The Methodology Center

Examining moderated effects of additional adolescent substance use treatment: Structural nested mean model estimation using inverse-weighted regression with residuals

Daniel Almirall
Faculty Research Fellow, Institute for Social Research, University of Michigan
Daniel F. McCaffrey
Senior Statistician & PNC Policy Chair, RAND Corporation
Beth Ann Griffin
Statistician, RAND Corporation
Rajeev Ramchand
Behavioral Scientist, RAND Corporation
Robert A. Yuen
Graduate Student, Department of Statistics, University of Michigan
Susan A. Murphy
Professor of Statistics, Psychiatry, & Research Scientist at the Institute for Social Research
University of Michigan

Technical Report Series
#12-121

Address correspondence to Daniel Almirall, dalmiral@umich.edu.

KEY WORDS: Structural nested mean model, Causal effect modification, Estimating equations, 2-stage estimator, Time-varying covariates, Time-varying treatment, Inverse-probability-of-treatment weighting

Funding for this work was provided by the following grants: R01-DA-015697 (Griffin, McCaffrey, & Ramchand), R01-MH-080015 (Murphy), and P50-DA-010075 (Murphy & Almirall). The authors would like to thank Cha-Chi Fan and Mary Ellen Slaughter for guidance with the data.

Abstract

This article considers the problem of examining causal effect moderation using observational, longitudinal data in which treatment, candidate moderators, and putative confounders are time-varying. Robins' (1994) 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 candidate time-varying 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 candidate time-varying moderators of interest are the sole time-varying confounders. The proposed IPTW+RR approach identifies the moderated time-varying causal effects in the SNMM in the presence of an additional, auxiliary set of known and measured putative time-varying confounders, which are not candidate time-varying moderators of interest. A small simulation experiment is used to compare IPTW+RR vs 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. The methodology is illustrated in a case study examining the moderated time-varying effects of additional adolescent substance use treatment on future substance use, as a function of time-varying frequency of substance use.

1. Introduction

Across a wide spectrum of the behavioral, medical, and social sciences, there is considerable interest in examining research questions having to do with the impact of time-varying treatments (or exposures) using longitudinal data. The methodology we discuss in this manuscript focuses on examining a particular set of scientific questions concerning *time-varying causal effect moderation* (Almirall, McCaffrey, Ramchand, & Murphy, 2011; Almirall, Ten Have, & Murphy, 2009), known as *time-varying causal effect modification* in the epidemiology literature (Petersen, Deeks, Martin, & van der Laan, 2007; Robins, Hernán, & Rotntzky, 2007; Petersen & van der Laan, 2007).

Moderator variables specify for whom (or under what conditions) treatment is more or less effective (Baron & Kenny, 1986; Kraemer, Wilson, & Fairburn, 2002). Often, moderator analyses are carried out in the context of point-treatment studies in which treatments are not time-varying or are not conceptualized as being time-varying (e.g., in secondary analyses of data arising from standard randomized trials). In the study of time-varying treatments, on the other hand, time-varying moderators are variables that specify for whom (or under what conditions) both initial treatment and the next step in treatment (e.g., treatment switch, augmentation, or dis/continuation) is more or less effective. 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. (Indeed, time-varying moderators may simultaneously be mediators of the impact of prior treatment; we note, however, that in this article our aim is not to develop methods for examining the mechanisms by which treatments exhibit their effects.)

To illustrate what we mean by time-varying causal effect moderation, consider a simplified version of our motivating example, in which 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) *yes*(1)/*no*(0) later treatment) on post-treatment substance use frequency (Y). One set of questions involves comparing 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 (Robins, 1997a; Robins, 1999; Robins, Hernán, & Brumback, 2000; Hernán, Brumback, & Robins, 2000; Cole et al., 2003). In this manuscript, we are interested in asking more detailed questions, 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 use prior to (S_0) and during (S_1) initial treatment 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,

(S_0, S_1) is a candidate time-varying moderator of the impact of time-varying 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, but without using person-specific information about concurrent or intermediate response to ongoing treatment.

An important challenge in the estimation of time-varying causal effects is that adjusting naïvely for other time-varying covariates may result in bias if the covariates are themselves impacted by prior treatment (Robins, 1987; Robins, 1989; Robins, 1997b). 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 that 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 that may be directly related to both subsequent treatment and outcome in order to reduce or eliminate time-varying confounding bias. However, in either case (i.e., whether adjusting for a time-varying covariate because it is a candidate moderator, or if 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, consider a naïve extension of the standard treatment-moderator interactions approach (Baron & Kenny, 1986) for studying effect moderation using observational study data, in which a regression model such as the following one is used:

$$\begin{aligned} E(Y|X_0,S_0,A_1,X_1,S_1,A_2) &= \beta_0+\eta_1X_0+\eta_2S_0+\beta_{1,1}A_1+\beta_{1,2}A_1S_0 \\ &\quad +\eta_3X_1+\eta_4S_1+\beta_{2,1}A_2+\beta_{2,2}A_2S_0+\beta_{2,3}A_2S_1 \end{aligned} \quad (1)$$

In this traditional regression analysis approach, the analyst adjusts for (S_0,S_1) because, as a candidate time-varying moderator, it is of particular scientific interest, whereas the analyst adjusts for (X_0, X_1) because it is a putative 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 that, 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. 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) 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 putative time-varying confounders, the X_ts are correlated with the A_ts. 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_ts and S_ts.

The third problem is especially problematic in most observational study settings, such as ours, where the list of observed putative time-varying confounders, the X_t, is significantly larger than the list of candidate time-varying moderators of interest. 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 problems above 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; Murphy, 2005). Further, these 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). The second problem mentioned, known as collider bias (Pearl, 1998), is particularly subtle; intuitive discussions of it are given in Almirall et al. (2011) and Cole et al. (2010).

The structural nested mean model (SNMM; Robins, 1994) provides a principled alternative to model (1), which specifies the moderated time-varying causal effects of interest in a conditional mean model for a continuous response given time-varying treatments and putative moderators. The structure of the SNMM provides a clue for how to condition on (S_0, S_1) appropriately to avoid the first two problems mentioned above. With regard to the third problem, the SNMM can be used 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_ts in the conditional mean model itself. Rather, the putative time-varying confounders are nuisances that are dealt with in the *estimation* of the causal parameters (via weighting, see below) of the SNMM, but not as part of the conditional mean model itself *defining* the causal effects of interest.

This technical report 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 previous work (Almirall et al., 2009; Almirall, Coffman, Yancy, & Murphy, 2010; Almirall et al., 2011; Henderson, Ansell, & Alshibani, 2011), the RR approach (when used without IPTW) identifies 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, putative time-varying confounders X_t. Following van der Laan, Murphy, and Robins (2002), van der Laan and Robins (2003, Section 6.5), and Robins (2004, pp. 78-80), 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 of scientific interest 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 scientist is happy to use X_t to adjust for confounding, but not necessarily otherwise interested in it scientifically.

In Section 2, we define the moderated time-varying causal effects more formally using the potential outcomes framework for causal inference; and we show how the moderated time-varying causal effects can be identified as part of a conditional mean model for the response given time-varying treatment and candidate moderators using Robins' (1994) 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 of the asymptotic versus bootstrap standard error estimates for the estimated parameters of the SNMM. In Section 6, we illustrate the methods in a case study examining the moderated effects of additional adolescent substance abuse treatment. Section 7 offers a discussion of the new methodology, including limitations and directions for future work.

2. A Model for Time-Varying Causal Effect Moderation

2.1 Potential outcomes notation

We define the causal parameters of interest and state the assumptions necessary for valid causal inference using the potential outcomes framework for causal inference (Rubin, 1974; Holland, 1986; Robins, 1987). Suppose there are K time intervals under study. Treatment at each time interval t is denoted by a_t (t=1,…,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,…, K. 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 A_K be the countable collection of all possible treatment vectors (e.g., for K=2, A_2 ={(0,0), (0,1), (1,0), (1,1)}; whereas for K = 3, A_3 is the set of $2^3 = 8$ triplets of 0 or 1). For each fixed value of the treatment vector, $\bar{a}_K$, we conceptualize potential, candidate time-varying moderators $\{S_1(a_1),\ldots, S_{K-1}(\bar{a}_{K-1})\}$ and a potential final response $Y(\bar{a}_K)$. Thus, $S_t(\bar{a}_t)$ is the vector of candidate time-varying moderators at the beginning of the tth interval had the client 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 client followed the treatment pattern $\bar{a}_K$. Baseline moderators (pre- $\bar{a}_K$) are denoted by the vector S_0. For shorthand, let

$\overline{S}_t(\bar{a}_t) = \{S_0, S_1(a_1), ..., S_t(\bar{a}_t)\}$, the history of candidate moderators up to the start of the tth time interval. For completeness we have indexed the candidate time-varying moderators $S_t(\bar{a}_t)$ by treatment $\bar{a}_t$ to acknowledge the potential for the moderators to be impacted by treatment; however, in this article we will not focus on the time-varying causal effects of $\bar{a}_t$ on $S_t(\bar{a}_t)$.

In our motivating example in Section 6, $K = 3$, a_1 denotes substance use treatment during months 1-3, a_2 denotes substance use treatment during months 4-6, a_3 denotes substance use treatment during 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 during months 4-6, and $Y(\bar{a}_3)$, our outcome of interest, is an end-of-study measure of frequency during months 10-12 (see Tables 3 and 6). For expositional simplicity, henceforth unless otherwise noted, we focus on the case where K=2 for defining the causal effects of interest and for giving intuition about the methodology. We do this by omitting discussion of $S_2(a_1, a_2)$ and a_3 in the context of our motivating example. The definitions extend easily to our case with K=2 and to general K time points. Thus, we have the following with which to define the causal parameters of interest: in temporal order, $\{S_0, a_1, S_1(a_1), a_2, Y(a_1, a_2)\} = \{S_0, a_1, S_1(a_1), a_2, Y(\bar{a}_2)\}$, or $\{\overline{S}_1(a_1), \bar{a}_2, Y(\bar{a}_2)\}$, in shorthand.

2.2 Moderated time-varying causal effects

The response $Y(a_1, a_2)$ is taken to be continuous. We are only concerned with modeling the mean of the response $Y(\bar{a}_2)$ as a function of $\bar{a}_2$ and $\overline{S}_1(a_1)$. Thus, for example, we do not explicitly consider treatment or covariate effects on the variance of the response.

The first causal effect function at t=1 is defined as

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

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, $\mu_1(1, s_0)$ 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. $\mu_1(1, s_0)$ is a comparison of substance use frequency at the end of the study had all clients 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. When a_1=0, the causal effect function is zero regardless of the value of s_0; that is, $\mu_1(0, s_0)$=0, as should be the case when comparing outcomes under the same treatment sequences.

The second causal effect function at t=2 is defined as

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

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 the study had all clients 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. As above, note that when $a_2=0$, the causal effect function is zero regardless of the values of s_0, a_1, and s_1; that is, $\mu_2(s_0,a_1,s_1,0)$, as should be the case when comparing outcomes under the same treatment sequences.

μ_1 and μ_2 are causal effect functions because at each time point, they represent contrasts (e.g., 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.

Finally, 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. It is possible to define μ_1 with future treatment set to a level other than zero (such as to the active level $a_2=1$). However, given that our interest is in examining the effects of additional substance use treatment as we move through time, setting future $a_2=0$ is a sensible choice dictated by our scientific interests. Further, in our motivating example (see Section 6), observed treatment sequences are predominantly monotonic whereby once an adolescent discontinues treatment, s/he rarely returns to treatment; therefore, these causal effects are also sensible given the treatment patterns observed in our data set.

2.3 The structural nested mean model

We present one version of the SNMM (Robins, 1994), which is a particular additive, telescoping decomposition of the conditional mean of $Y(a_1, a_2)$ given $\overline{S}_1(a_1)$. This decomposition includes the causal functions μ_1 and μ_2 as part of the decomposition. Specifically, for $K=2$, the SNMM is expressed as

$$E(Y(a_1,a_2) \mid \overline{S}_1(a_1) = \overline{s}_1) = \beta_0 + \varepsilon_1(s_0) + \mu_1(s_0,a_1) + \varepsilon_2(\overline{s}_1,a_1) + \mu_2(\overline{s}_1,\overline{a}_2), \quad (4)$$

where the intercept $\beta_0=E(Y(0,0))$ is the mean outcome for the population under no treatment; and the functions $\varepsilon_1(s_0)$ and $\varepsilon_2(\overline{s}_1,a_1)$ are defined as follows:

$$\varepsilon_1(s_0) = E(Y(0,0) \mid S_0 = s_0) - E(Y(0,0)) \quad (5)$$

$$\varepsilon_2(\overline{s}_1(a_1)) = E(Y(a_1,0) \mid \overline{S}_1(a_1) = \overline{s}_1) - E(Y(a_1,0) \mid S_0 = s_0). \quad (6)$$

Note that $\varepsilon_1(s_0)$ and $\varepsilon_2(\overline{s}_1, a_1)$ are defined just so the right-hand side of (4) equals the conditional mean on the left-hand side when it is also expressed as a function of the causal functions μ_1 and μ_2. Following Robins (1994), we label the functions ε_1 and ε_2 as "nuisance functions" to distinguish them from our primary 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, which forms the basis for how we model these quantities using the RR approach in Section 4 below. Namely, the nuisance functions are mean-zero functions conditional on the past; that is,

$$E(\varepsilon_1(S_0))=0, \text{ and} \tag{7}$$

$$E(\varepsilon_2(\overline{S}_1(a_1)) \mid S_0) = 0, \tag{8}$$

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. This conditional mean-zero property is what makes the SNMM a non-standard regression model. This property also forms the basis for how we properly model the conditional mean $E(Y(a_1,a_2)\mid \overline{S}_1(a_1))$ using the RR approach in Section 3 below. That is, understanding how to model the nuisance functions properly helps resolve the first two problems with the standard regression model (1) discussed in the Introduction.

3. Linear Parametric Models for the SNMM

3.1 Linear models for the causal functions

Up to this point, we have defined the components of the SNMM (including the moderated time-varying causal effects of interest), but we have not discussed how we will model the components of the SNMM. In this manuscript, we consider parametric linear models for the μ_ts of the form

$$\mu_t(\overline{s}_{t-1}, \overline{a}_t; \beta_t) = a_t(H_{t-1}\beta_t), \tag{9}$$

where β_t is an unknown q_t-dimensional column-vector of parameters, and $H_{t\text{-}1}$ is a corresponding row-vector that is a function of $(\overline{S}_{t-1}(\overline{a}_{t-1}), \overline{a}_{t-1}) = (\overline{s}_{t-1}, \overline{a}_{t-1})$. $H_{t\text{-}1}$ stands for *H*istory up to time t=1. This functional form for the causal functions is an extension of the standard treatment-moderator interaction (i.e., covariate-by-treatment product terms; Baron & Kenny, 1986) framework to the time-varying setting. For example, for t=2, let $H_1 = (1, s_1, a_1)$ and $\beta_2=(\beta_{2,1},\beta_{2,2},\beta_{2,3})^T$ (where v^T means transpose of v) so that

$$\mu_2(\overline{s}_1, \overline{a}_2; \beta_2) = a_2(\beta_{2,1} + \beta_{2,2}s_1 + \beta_{2,3}a_1) = \beta_{2,1}a_2 + \beta_{2,2}s_1a_2 + \beta_{2,3}a_1a_2. \tag{10}$$

In this model, the effects of additional substance abuse treatment depend on previous treatment a_1 (according to $\beta_{2,1}$) and also vary linearly in $S_1(a_1)$ (with slope equal to $\beta_{2,2}$). According to this model, $H_0: \beta_{2,2}=\beta_{2,3}=0$ is the null hypothesis that the effect of additional treatment is not moderated by $(S_0, a_1, S_1(a_1))$; that is, that the effect of additional treatment is constant given $(S_0, a_1, S_1(a_1))$.

3.2 Linear models for the nuisance functions

We also consider parametric linear models for the ε_ts. For univariate S_{t-1}, we consider models such as the following:

$$\varepsilon_t(\bar{s}_{t-1}, \bar{a}_{t-1}; \eta_t, \gamma_t) = \eta_t \delta_t(\bar{s}_{t-1}, \bar{a}_{t-1}; \gamma_t) \tag{11}$$

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; McCullagh & Nelder, 1989), 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 either the "inverse logit" transform or the "inverse probit" transform; on the other hand, for continuous $S_{t-1}(\bar{a}_{t-1})$, $g_t()$ would be the identity function.

Note that, consistent with properties (7) and (8), $E(\varepsilon_t(\bar{S}_{t-1}(\bar{a}_{t-1}); \eta_t, \gamma_t) \mid \bar{S}_{t-2}(\bar{a}_{t-2})) = 0$ since 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 naming the proposed estimator in Section 4 "regression-with-residuals." The notation appears overly complicated, but these ideas becomes more clear in the simple example given in equation (12) below, in equation (14) below where we show an example model for the full SNMM which involves models for the nuisance functions, and in Section 4 below where we describe how to implement the proposed estimator of the SNMM.

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). In this case, a sample model for ε_2 that is consistent with (11) is

$$\varepsilon_2(s_0, a_1, s_1; \eta_2, \gamma_2) = \eta_2 \left(s_1 - m_1(s_0, a_1; \gamma_2)\right), \text{ where} \tag{12}$$

$$m_1(s_0, a_1; \gamma_2) = \gamma_{2,0} + \gamma_{2,1} s_0 + \gamma_{2,2} a_1 + \gamma_{2,3} s_0 a_1 \tag{13}$$

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

The parametric form in (11) 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 $\varepsilon_{tk} = \eta_{tk} \delta_{tk}$, one for each S_{tk} as in (11), and then summing these models together to create an overall parametric model for tth time-point nuisance function: $\varepsilon_t = \sum_k^{r_t} \varepsilon_{tk}$ (see the appendix in Almirall et al., 2009). Note that in the multivariate case, δ_t is a 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 Putting it all together

Combining the linear parametric models for both the causal (μ_t) and nuisance (ε_t) functions, we arrive at a linear parametric SNMM, denoted m_γ. For instance, assuming the candidate time-varying moderator $S_t(\overline{a}_t)$ is univariate continuous, and using the example models above, 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 $E(S_0)$, implies the following example linear SNMM:

$$m_\gamma(\overline{s}_1, \overline{a}_2; \beta, \eta, \gamma) = \beta_0 + \delta_1\eta_1 + \beta_{1,0}a_1 + \beta_{1,1}s_0a_1 + \delta_2\eta_2 + \beta_{2,1}a_2 + \beta_{2,2}s_1a_2 + \beta_{2,3}a_1a_2, \qquad (14)$$

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, as above, $\delta_2 = s_1 - m_2$.

It is noteworthy that this linear SNMM is very similar to the traditional regression analysis approach, equation (1), except it differs in at least two important ways: first, in equation (14), the "main effects" of the candidate time-varying moderators are conditional-mean centered. That is, the S_ts in equation (1) are replaced by δ_ts in equation (14). The intuition here is that by "residualizing" the S_ts–in particular, residualizing S_1–we avoid the potential problems described in the Introduction related to naïvely conditioning on candidate moderators impacted by prior treatment. Second, equation (14) focuses solely on relating the outcome Y with time-varying treatments and candidate moderators; that is, it does not adjust for putative 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, focusing on the proposed estimator, describe how to estimate the parameters of the SNMM all the while adjusting for the putative time-varying confounders X_t using a weighted least squares regression approach.

More generally, letting $D_\gamma(1, \delta_1, a_1H_0, \delta_2, a_2H_1)$ 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 $m_\gamma = D_\gamma\theta$. 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_ts, which make up the models for the nuisance functions, the ε_ts.

Apart from the special case of fully saturated models which, by definition, can not be mis-specified (see Almirall et al., 2011 for an example of a saturated SNMM), note that these parametric models constitute modeling assumptions the scientist must make. This is the first of four assumptions made in this methodology. The other three assumptions–consistency, no unmeasured time-varying confounders, and positivity–are specific to estimating the SNMM; they are described in Section 4.1 below.

4. Estimation

4.1 Observed data and assumptions

In this subsection, we describe the observed data–and its connection to the potential outcomes–which are used to estimate the causal parameters of interest. 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 below). 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 to S_t. As before, $\bar{A}_t = (A_1, \ldots, A_t)$ for t=1,…,K, and similarly $\bar{S}_t = (S_1, \ldots, S_t)$ for t=1,…,$K-1$.

The link between the potential outcomes and the observed data O is established by invoking the consistency assumption (Robins, 1994) for both the S_ts and Y: for all clients in the study, the consistency assumption for the end of the study outcome states that $Y = Y(\bar{A}_K)$, where the $Y(\bar{A}_K)$ denotes the potential outcome indexed by values of a_K equal to A_K. This assumption says that the observed outcome Y for a client 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.

Intuitively, a confounder of μ_1 is a pre- A_1 covariate that is correlated with A_1 and Y; a confounder of μ_2 is a pre- A_2 covariate that is correlated with A_2 and Y. V_0 may include confounders of μ_1 and μ_2; (V_0, A_1, V_1) may include confounders of μ_2. Note that confounders of μ_2 may be mediators of earlier treatment. Confounders create pre-treatment imbalances in the types of clients observed on (A_t=1) or off (A_t=0) treatment in ways that are correlated with outcome; therefore, they complicate the identification of μ_t based on comparison of outcomes between adolescents on A_t=1 versus off A_t=0 treatment.

In order to identify the μ_ts using the observed data, we make the no unmeasured or unknown direct confounders assumption (Robins, 1994): for every t (t=1,2,…,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 (Murphy, 2005), this assumption is satisfied by design. In observational studies, it is not possible to know whether this assumption is satisfied; it is not testable given the observed data, unless it is replaced by (sometimes more stringent) additional assumptions. In the context of observational studies, this assumption informally states (for every t) that aside from the history of putative 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,

$$0 < Pr(A_t = 1 \mid \bar{V}_{t-1}, \bar{A}_{t-1}) < 1. \tag{15}$$

This technical assumption ensures we do not have true weights (which are inversely proportional to $Pr(A_t = 1 \mid \bar{V}_{t-1}, \bar{A}_{t-1})$ or its complement) with infinite values. Informally, this assumption states that every client could potentially be assigned to any of the treatments (at each time t) and that there is some overlap among the groups receiving the different treatments, such that there are no values of $(\bar{V}_{t-1}, \bar{A}_{t-1})$ that can occur only among units receiving treatment (or not receiving treatment) at time t.

4.2 A set of estimating equations

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

$$0 = \mathbb{P}_n \psi_\theta(O; \theta, \gamma, \alpha, \pi) = \mathbb{P}_n W(\alpha, \pi)(Y - D_\gamma \theta) D_\gamma^T, \tag{16}$$

where n is the number of clients 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 IPTWs $W(\alpha, \pi)$ are defined as

$$W(\bar{V}_{K-1}, \bar{A}_K; \alpha, \pi) = \prod_{t=1}^{K} W_t(\bar{V}_{t-1}, \bar{A}_t; \alpha_t, \pi_t), \tag{17}$$

where

$$W_t(\bar{V}_{t-1}, \bar{A}_t; \alpha_t, \pi_t) = A_t \frac{p_t^{num}(\pi)}{p_t^{den}(\alpha)} + (1 - A_t) \frac{(1 - p_t^{num}(\pi))}{(1 - p_t^{den}(\alpha))}, \tag{18}$$

where 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 (16) 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 putative time-varying confounders X_t are not a part of the linear SNMM (D_γ), but are adjusted for via the weights (W). van der Laan et al. (2002), van der Laan and Robins (2003, Section 6.5), and Robins (2004, pp. 78-80) provide the theory that shows that under the four assumptions listed above and known $W(\alpha,\pi)$ and γ, the proposed estimator identifies the parameters θ of the SNMM. In practice, however, implementing the estimator above requires more work 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.

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 clients who are treated (A_t=1) versus those who are untreated (A_t=0). The weights accomplish this by up-weighting clients who are unlikely to receive the treatment they were given $(\bar{A}_{t-1}, \bar{V}_{t-1})$, and by down-weighting clients who are likely to have received the treatment they were 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). The numerator p_t^{num} is not required for eliminating or reducing bias due to time-varying confounding. The numerator is used to project the $1/p^{den}$-weighted sample back to the space of conditional distributions given $\bar{S}_{t-1}$. Intuitively, the reason for doing this is because, at each time point *t*, we are interested in and we explicitly model the effect of A_t on Y given $\bar{S}_{t-1}$ (the moderated effect of A_t given levels of S_{t-1}). Therefore, weighting the sample back to "within observed levels of S_{t-1}" in this fashion makes sense. Another intuitive way to think about this projection in the context of SMARTs, which can be 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. Statistically, the advantage of using the numerator probabilities is that it potentially increases the statistical efficiency in the estimates $\hat{\theta}$ by making the weights W_t less variable since $0 < p_t^{num}(\pi) < 1$. This is why Robins and colleagues (2000) call these "stabilized weights." The weights used in this methodology are also discussed in Petersen et al. (2007) and used by Rosthøj, Keiding, and Schmiegelow (2009) to estimate history-adjusted marginal structural models.

Another way to think about the projection induced by the numerator model, intuitively, is that by conditioning on S_{t-1} in the numerator model, imbalances in A_t=1 versus A_t=0 due to S_{t-1} are preserved. This may appear counterproductive to the aim of removing or eliminating confounding; that is, the ideal is that the weights remove confounding due to $V_{t-1} = (S_{t-1}, X_{t-1})$, not just X_{t-1}. However, since by definition the SNMM conditions on $\bar{S}_K$ (because they are candidate moderators of the impact of A_t) and therefore adjusts for putative confounding by S_{t-1} explicitly as part of the linear model, then undoing the balancing on S_{t-1} (i.e., that achieved by $1/p^{den}$) in this fashion incurs no penalty in terms of bias in the estimated θ (assuming, of course, the four assumptions listed above are met).

4.3 Implementation steps: IPTW+RR

The (α,π)–and therefore, the weights $W(\alpha,\pi)$–are unknown, as is γ. This section describes steps to implement estimator (16) 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.

We propose the following steps for implementing the above estimator:

Step 1. Estimate the weights. As discussed above, for each *t*, the numerator propensity score is a function of $(\overline{S}_{t-1},\overline{A}_{t-1})$. 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 $(\overline{V}_{t-1},\overline{A}_{t-1})$.

1a. **Estimate the numerator model (obtain $\hat{\pi}$).** For each *t*, estimate the numerator propensity score p_t^{num} using a logistic regression model for $Pr(A_t=1\mid\overline{S}_{t-1},\overline{A}_{t-1})$ 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 the denominator propensity score p_t^{den} using a logistic regression model for $Pr(A_t=1\mid\overline{V}_{t-1},\overline{A}_{t-1})$ with unknown parameters α_t. Calculate and save the $\hat{p}_t^{den}(\hat{\alpha}_t)$ s.

1c. **Estimate the SNMM using RR+IPTW(obtain $\hat{\theta}$)**

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

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 for $E(S_{t-1}\mid\overline{S}_{t-1},\overline{A}_{t-1})$ 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 $t\geq1$, 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 $\hat{\theta}$). Specify a model $D_{\hat{\gamma}}$ for the SNMM. Note, the models for the nuisance functions (e.g., main effects of the candidate time-varying moderators) in $D_{\hat{\gamma}}$ uses the residuals $\hat{\delta}_t$ s from Step 1. 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 $\hat{\theta}$ that are smaller than appropriate. Asymptotic standard errors (ASE) obtained using the delta method (e.g., Taylor series arguments, see Appendix C), which take into account sampling error in the estimation of $(\hat{\gamma}, \hat{\alpha}, \hat{\pi})$ are used instead. However, since not all investigators have the resources to computer program the ASEs, we also compare results with bootstrap standard error estimates for $\hat{\theta}$, which are easier to calculate using most over-the-counter statistical software packages. To obtain the bootstrap standard error, 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 of the 500 estimates.

5. Simulations

Two small simulation experiments were conducted to (1) illustrate and compare IPTW+RR versus the RR versus the traditional regression approach under simple conditions where the true SNMM is known, and to (2) compare small and large sample properties of asymptotic versus bootstrap estimators of the standard errors for the IPTW+RR approach.

The data generating model for $\{U, V_0, A_1, V_1, A_2, Y\} = \{U, (X_0, S_0), A_1, (X_1, S_1), A_2, Y\}$ is described in Appendix A. The data were generated to mimic (to the extent possible) the adolescent substance use data (see Section 6; except with K=2 instead of K=3). We did this by ensuring that the marginal distributions (e.g., proportions, means, standard deviations) of the generated data were similar to the adolescent substance use data.[1]

The generated data will be used to estimate the SNMM for $E(Y(a_1, a_2) \mid S_0, S_1(a_1))$. Key features of the data generating model include the following:

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 are

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

mediators of the effect of A_1 on Y.

4. We generate a baseline variable U that affects both $S_1$and Y.[2] This variable creates a spurious (non-causal) correlation between A_1 and Y when adjusting naïvely for S_1.
5. In the first simulation we vary whether or not there exists time-varying confounding by X_t.

5.1 IPTW+RR vs RR vs traditional regression

The first simulation allows us to illustrate, in the context of a simple example, the main feature of the proposed IPTW+RR estimator and how it differs from the RR and two versions of the traditional regression approach for estimating the K=2 SNMM for the conditional mean $E(Y(a_1,a_2)\,|\,S_0,S_1(a_1))$. Two versions of the data generating model were considered in this experiment: with and without time-varying confounding by X_t. Both the IPTW+RR and the RR estimators use the correct functional form for $E(Y(a_1,a_2)\,|\,S_0,S_1(a_1))$. The first traditional regression estimator (TRAD1) fits the same functional form as fitted with the IPTW+RR and RR estimators except that the δ_ts are replaced with S_t. The second traditional regression estimator (TRAD2) also adjusts for X_t. Under each condition (with or without time-varying confounding) 1000 data sets of size n=2870 (the size of the adolescent substance use data) were generated. Rather than showing simulation results for all parameters of the SNMM, for simplicity we report summaries of the relative bias for

- β_0 = the intercept, or marginal mean outcome under no treatment, and
- $\beta_{t,0}$ = the effect of treatment at time t vs no treatment at time t given no substance use frequency prior to treatment, for t=1,2

under each estimator: IPTW+RR vs RR vs TRAD1 vs TRAD2.

Table 1 shows the results of the first simulation experiment, which confirm our expectations. We make the following observations:

- When there is time-varying confounding, RR and TRAD1 were biased, whereas the IPTW+RR was unbiased. This is because RR and TRAD1 do not adjust for X_t in any way.
- Whether or not 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) naïvely. The bias we see in TRAD2 is due to both cutting off the 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 due to adjusting naïvely 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 since the goal of IPTW is to adjust for

[2] U is not on the causal pathway between A_1 and Y. It is neither a time-varying moderator, nor a confounder, nor a mediator: it is simply a measure that explains variance in S_1 and Y.

time-varying confounders.

- When there is no time-varying confounding, TRAD1 and TRAD2 continue to be biased for the parameters in (β_0, $\beta_{1,0}$). This bias occurs because of the problems with traditional regression noted in the Introduction. The bias is greater in TRAD2 than with TRAD1 since more of the effect of A_1 on Y is cut off when we adjust for both (S_1, X_1) (TRAD2) than if we adjust for only S_1 (TRAD1). Recall that both (S_1, X_1) are mediators of the effect of A_1 on Y.
- When there is no time-varying confounding, all four estimators are unbiased for $\beta_{2,0}$.
- Whether there was time-varying confounding or not, 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 Almirall et al. (2009) for details.

The overarching conclusions are as expected: In general, (a) TRAD1 or TRAD2 (weighted or not) is not a principled estimator of the parameters of all of the SNMM; and (b) when there is no time-varying confounding, IPTW is not necessary and RR is by itself sufficient.

5.2 Asymptotic vs bootstrap standard errors

The second simulation experiment focuses on comparing the large and small sample properties of the bootstrap vs asymptotic estimates of the standard errors of the IPTW+RR. For this set of experiments, we employed the data generating model used above in which there is time-varying confounding. In this experiment

Table 1. Results from a simulation experiment to illustrate and compare IPTW+RR versus the RR versus the traditional regression approach under simple conditions where the true SNMM is known

		Bias			
Generative Model	Effect	IPTW+RR	RR	TRAD1	TRAD2
Time-varying confounding	β_0	0.00	**0.24**	**3.10**	**4.15**
	$\beta_{1,0}$	0.00	**0.19**	**0.51**	**0.74**
	$\beta_{2,0}$	0.00	**0.14**	**0.14**	0.00
	Average bias	0.00	**0.19**	**1.25**	**1.63**
No time-varying confounding	β_0	0.00	0.00	**2.99**	**4.14**
	$\beta_{1,0}$	0.00	0.00	**0.36**	**0.74**
	$\beta_{2,0}$	0.01	0.01	0.01	0.00
	Average bias	0.00	0.00	**1.12**	**1.63**

Appendix A describes the data generating 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 naïvely for S_t only (the candidate time-varying moderator). TRAD2 is the traditional regression estimator that adjusts naïvely for both S_t and X_t (a time-varying confounder). 1000 data sets of size n=2870 were used in all simulation conditions. Bias is defined as the relative bias |(TRUE−EST)/TRUE|. Conditions in which there is bias are shown in bold.

we varied the sample size: n=100(small), 250(medium),500(large),2870(very large, the size of the adolescent substance use data set). In Table 2, we report the standard deviation of the IPTW+RR estimates (SD), mean bootstrap standard errors (BOOT), mean asymptotic standard errors (ASE), and coverage of the 95% confidence intervals for both the bootstrap (BOOT95) and ASEs over the 1000 simulated data sets. As in the simulation above, we report results for estimates of β_0, $\beta_{1,0}$, and $\beta_{2,0}$. At very large samples, such as with n=2870, the bootstrap and the ASE were nearly indistinguishable in our simulation experiments. However, in small samples (n=100), the 95% confidence interval calculated using the ASE had lower than nominal coverage (0.932 for β_0) and much lower than nominal coverage (0.919 for $\beta_{1,0}$). In general, we noticed that BOOT95 had closer to nominal coverage than did ASE95.

6. Case Study: Moderated Effects of Additional Adolescent Substance Use Treatment

6.1 Data and measures

Sample. The methodology is illustrated using data (n=2870 clients) pooled from a number of adolescent treatment studies funded by the Substance Abuse and Mental Health Services Administration's Center for Substance Abuse Treatment involving adolescents entering community-based substance abuse treatment programs. All data points were collected using the Global Appraisal of Individual Needs (GAIN; Dennis, Titus, White, Unsicker, & Hodgkins, 2002), a structured clinical interview of client characteristics and functioning administered at baseline/intake and at the end of 3, 6, 9, and 12 months for a total of 5 measurement occasions. At each

Table 2. 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. 1000 data sets were used in all simulation conditions. BOOT95 and ASE95 refer to the coverage probabilities (over the 1000 data sets) for the 95% confidence interval constructed using either the bootstrap SE or the ASE, respectively.

Table 3. Notation and temporal ordering of the variables used as treatment, moderators, and confounders in the adolescent substance use data analysis. Measurements are taken at baseline and at the end of every 3 month interval.

	Baseline	0-3 Months	3-6 Months	6-9 Months	9-12 Months
Measurement taken, t'	0	3	6	9	12
Time notation, t'	0	1	2	3	4
Treatment	-	A_1	A_2	A_3	-
Outcome	-	-	-	-	Y
Moderators	S_0	S_1	S_2	-	-
Confounders	X_0	X_1	X_2	-	-

In the data, the treatment, outcome, and moderator variables are A_1=anytxt3, A_2= anytxt6, A_3= anytxt9, Y= sfs8p12, S_0=($S_{0,1}$, $S_{0,2}$, $S_{0,3}$,) = (sfs8p0, b2a, ce0), S_1 = sfs8p3, and S_2 = sfs8p6, where anytxt t' is an indicator of treatment (binary 0/1), sfs8pt' is the substance frequency scale (continuous), b2a is age (continuous), and ce0 is an indicator of whether or not the adolescent spent time in a controlled environment in the 90 days prior to baseline (binary 0/1). S_0 is multivariate. Variable names for X_t are given in Table 4.

measurement occasion, GAIN questions ask about constructs over the past 90 days (past 3 months). Table 3 describes the notation and temporal ordering of the variables used as treatment (A_t), moderators (S_t), and confounders (X_t) with respect to the GAIN data collection design. Next, we describe the actual measures used.

Treatment. For the illustrative analysis, time-varying treatment A_t=1 if a client reports receiving any substance use treatment in the past 90 days (i.e., the client reported receiving inpatient treatment, outpatient treatment, or both); and A_t=0 otherwise. In the data set, this variable is called anytxtt' where t'=3,6,9 denotes t=1,2,3 (see Table 3).

Moderator Variables. The primary time-varying moderator of interest is the Substance Frequency Scale (SFS) collected at baseline ($S_{0,1}$), and at the end of months 3 (S_1) and 6 (S_2). The SFS is a continuous measure (possible range is (0,1)) based on 8 items that assesses the average proportion of alcohol and other drug using days in the past 90 taking into account heavy use and problem days. 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 the data set, this variable is called sfs8pt' where t'=0,3,6 denotes t=0,1,2.

In addition to $S_{0,1}$, we also consider the following two variables as candidate baseline moderator variables as part of S_0: $S_{0,2}$ is age (continuous), so that we may explore the developmental heterogeneity in the effects of time-varying treatment, and $S_{0,3}$ is a binary indicator taken at baseline of whether or not the adolescent reports being in a controlled environment in the past 90 days. In the data set, ($S_{0,2}$, $S_{0,3}$) are called b2a and ce0, respectively. Note that S_0=($S_{0,1}$,$S_{0,2}$, $S_{0,3}$)=(sfs8p0, b2a, ce0) is multivariate, whereas S_1 and S_2 each are univariate.

Outcome. Y is the SFS collected at the end of month 12. In the data set, this variable is called sfs8p12. Note that the methodology, as presented thus far, considers only an end of study primary outcome Y. In the Discussion section, we discuss the opportunity and the challenges of extending the methodology to allow for longitudinal Y; that is, where SFS is both a time-varying outcome and a time-varying moderator.

Confounder Variables. Table 4 describes the large list of auxiliary, candidate confounder variables. X_t includes 38 time-varying confounder variables (for each t=0,1,2). X_0 also includes gender, race (4 categories=3 dummy variables), and two other measures about behavior problems and internal mental distress that were collected only at baseline (non-time-varying confounder variables). Therefore, V_0=(X_0,S_0) includes J_1=(38+6)+2=46 candidate pre- A_1 confounders. V_1=(X_1,S_1) and V_2=(X_2,S_2) each include 38+1=39 variables. Therefore, there are J_2=46+1+39=86 candidate pre- A_2 confounders (the size of (V_0, A_1, V_1)) and J_3=86+1+39=128 candidate pre- A_3 confounders (the size of (V_0,A_1,V_1,A_2,V_2)).

6.2 Missing data

Prior to analysis, we used multiple imputation to replace missing values. A sequential regression multivariate imputation algorithm was used, as implemented in the IVEware package for SAS (Raghunathan, Lepkowski, Hoewyk, & Solenberger, 2001; Raghunathan, Solenberger, & Hoewyk, 2002). The imputation model included the outcome measure of interest, the longitudinal putative moderators, the longitudinal putative confounders, longitudinal treatment indicators, and treatment-by-putative moderator interaction terms (so that the data analysis model is subsumed within the imputation model). Ten data sets were generated. All parameter estimates and standard errors (SE) reported below were calculated using Rubin's (Rubin, 1987; Schafer, 1997) rules for combining the results of identical analyses performed on each of the 10 imputed data sets.

6.3 Descriptive information

Table 5 shows the frequency and proportion of the 2^3=8 possible treatment patterns. The majority of clients (89%) were observed to follow a monotonic treatment pattern, whereby participants either never participate in treatment (A_1, A_2, A_3)=(0,0,0), fully participate in treatment for the full duration of 9 months (A_1,A_2,A_3)=(1,1,1), or begin to participate in treatment, but then discontinue at some point (A_1,A_2,A_3)={(1,0,0),(1,1,0)}. Table 6 shows the descriptive data for the candidate moderators and the primary outcome.

Table 4. Description of time-varying confounder variables X_t used in the illustrative data analysis

Variable name	Description
Baseline (non-time-varying) confounders	
female	Gender: 1=female, 0=male
race4g2	Race: 1=White, 0=other
race4g3	Race: 1=Black, 0=other
race4g4	Race: 1=Hispanic, 0=other
bcs	Behavior Complexity Scale - High means more problems
imds	Internal Mental Distress scale - High means more distress
Time-varying confounders	
arrtott'	Number of arrests in past 90 days
ceit'	Controlled Environment Index - High means more time
cjsit'	Number of days (in past 90) involved in justice system
cwst'	Current Withdrawal Scale - High means more symptoms
dcst'	Drug Crime Scale - High means more problems
emaspt'	Employment Activity Scale - High means more employment
eps7pt'	Emotional Problem Scale - High means more problems
everopt'	Ever used an opiate? 0=no, 1=yes
erst'	Environmental Risk Scale - High means more problems
hps3pt'	Health Problem Scale - High means more problems
ias5pt'	Illegal Activities Scale - High means more activity
lrit'	Living Risk Index - High means more risky environment
maxcet'	Number days in a controlled environment in past 90
phti4t'	Physical Health Treatment Index - High means more days in txt
r4at'	Tobacco use count past 90 days
s2s2t'	Number days (in past 90) that AOD kept from responsibilities
s5at'	Number times admitted to detoxification program for AOD
s7ft'	Currently treated regularly for AOD problems? 0=no, 1=yes
satit'	Substance Abuse Treatment Index - High means more txt for AOD
schoolt'	In school in the past 90 days? 0=no,1=yes
sdsmt'	Substance Dependance Scale Monthly - High means more problems
sdsyt'	Substance Dependance Scale Yearly - High means more problems
spsmt'	Substance Problem Scale Monthly - High means more problems
spsyt'	Substance Problem Scale Yearly - High means more problems
sri7t'	Social Risk Index - High means more time with people at risk
tas5pt'	Training Activity Scale - High means more recent days working
v90t'	Ever victimized in past 90 days: 0=no, 1=yes
wkyhrt'	Weekly heroin/opiod use: 0=no, 1=yes
workt'	Employed in past 90 days? 0=no, 1=yes
dlyusenewt'	Using drugs daily? 0=no, 1=yes
wkyfmpnewt'	Weekly family problems in past 90 days: 0=no, 1=yes
e10t'	Personal sources of stress index - High means more stress
mreri13pt'	Recovery environmental risk index - High means more recovery
hmlsrunnewt'	Homelessness/runaway in past 90? 0=no, 1=yes
p3newt'	Index for limited abilities due to health in past 90 days?
s6newt'	Ever attended a self-help group for AOD use? 0=no, 1=yes
mhti3t'	Mental Health Treatment Index - High means more services received
recovt'	Incarcerated < 14 and no drugs and alcohol? 0=no, 1=yes

$t'=0,3,6$ denotes $t=1,2,3$ (see Table 1)

Table 5. Treatment trajectories

Treatment (A_1, A_2, A_3)	Frequency	Proportion
(0,0,0)	310	11%
(0,1,0)	56	2%
(1,0,0)	1184	41%
(1,1,0)	555	19%
(0,0,1)	56	2%
(0,1,1)	56	2%
(1,0,1)	153	5%
(1,1,1)	499	17%

Table 6. Descriptive data for the candidate moderators and the primary outcome

Moderators		Mean	SD	Range
$S_{0,1}$	sfs8p0	0.18	0.18	(0, 0.89)
$S_{0,2}$	b2a	15.98	1.4	(12, 25)
$S_{0,3}$	ce†	0.49	-	(0, 1)
S_1	sfs8p3	0.07	0.11	(0, 0.67)
S_2	sfs8p6	0.08	0.13	(0, 0.73)
Outcome		Mean	SD	Range
Y	sfs8p12	0.09	0.13	(0, 0.78)

†$S_{0,3}$=ce is a binary 0/1 random variable; hence, in the mean column, we report $\widehat{Pr}(S_{0,3}=1)$.

6.4 Step 1. Estimating the weights

Recall that in this methodology, IPTW is a *statistical tool* used to adjust for time-varying confounding (this is the primary role of the denominator models $p_t^{den}(\alpha_t)$ used in the weights). In addition, the numerator models $p_t^{num}(\pi_t)$ used in the weights are also statistical tools used to improve statistical efficiency (some of which may be lost due to inverse-weighting). Therefore, the results of the logistic regressions for the numerator and denominator models provided in this section–while interesting and while they provide some information on the associations between covariates and treatment uptake–are not to be interpreted from a causal clinical/public health point of view.

Step 1a. Estimate the numerator models $p_t^{num}(\pi_t)$. Table 7 shows the results of the logistic regression numerator models $p_t^{num}(\pi_t)$ for $Pr(A_t=1 \mid \bar{S}_{t-1}, \bar{A}_{t-1})$. The results suggest that higher frequency of use at baseline (−3 to 0 months prior to initial treatment opportunity) is associated with higher probability of treatment uptake throughout; whereas, higher 0-3 and 3-6 month frequency of use is associated with lower probability of subsequent treatment uptake.

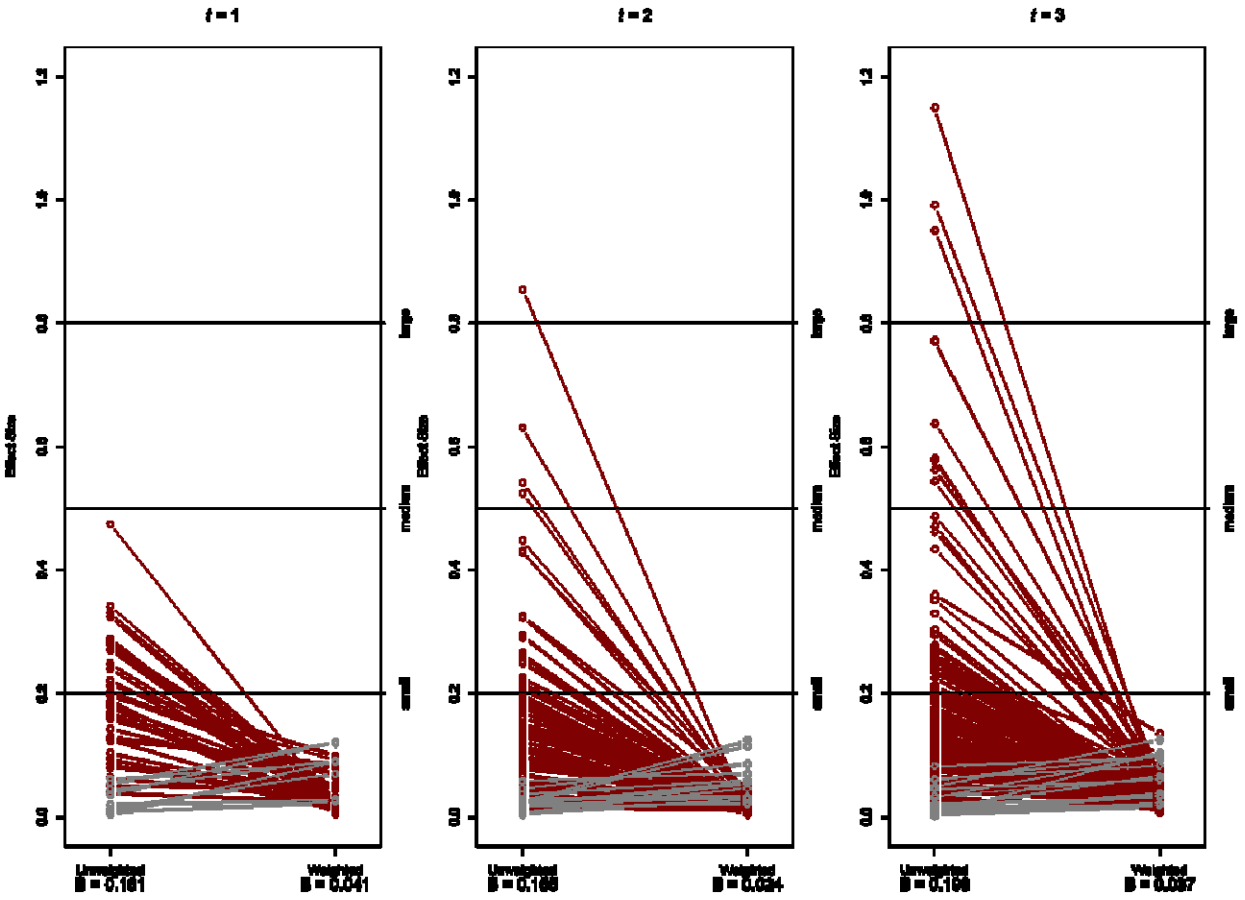
Table 7. Results of the logistic regression models for $p_t^{num}(\pi_t)$

t	Covariate	$\hat{\pi}_t$	SE	z	$Pr(>\|z\|)$
1	(Intercept)	2.437	0.597	4.084	<0.01
	sfs8p0	1.853	0.325	5.701	<0.01
	ce	-0.188	0.101	-1.850	0.06
	b2a	-0.065	0.037	-1.741	0.08
2	(Intercept)	-1.709	0.492	-3.469	<0.01
	sfs8p0	1.470	0.230	6.394	<0.01
	ce	0.367	0.079	4.659	<0.01
	b2a	0.012	0.030	0.409	0.68
	anytxt3	0.925	0.122	7.602	<0.01
	sfs8p3	-1.412	0.376	-3.754	<0.01
3	(Intercept)	-2.189	0.606	-3.610	<0.01
	sfs8p0	1.036	0.269	3.856	<0.01
	ce	0.497	0.099	5.039	<0.01
	b2a	0.001	0.039	0.029	0.98
	anytxt3	-0.257	0.206	-1.248	0.21
	sfs8p3	0.751	0.477	1.573	0.12
	anytxt6	1.843	0.103	17.891	<0.01
	sfs8p6	-0.996	0.471	-2.117	0.03

Consistent with the treatment trajectory patterns described above, treatment in the previous 3 months is associated with treatment uptake in the next 3-month interval.

Step 1b. Estimate the denominator models $p_t^{den}(\alpha_t)$. 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) clients based on observed time-varying confounders up to $t-1$. Following McCaffrey, Ridgeway, and Morral (2004), we employed a strategy for selecting the denominator logistic regression models for $p_t^{den}(\alpha_t)$ that leads to improved balance. Balance is measured based on summaries of the effect sizes $ES_{t,j}^*$ (e.g., mean and max over covariates j=1,…,J_t; Cohen, 1988) between treated vs untreated clients in the W_t^*-weighted sample, where $W_t^* = A_t / p_t^{den} + (1-A_t)/(1-p_t^{den})$.[3] Smaller

Figure 1. Unweighted vs W_t^*-weighted balance between treated and untreated groups



[3]These are not the final weights W (Step 1c) used in the weighted regression used to estimate the SNMM; however, they can be thought of as the part of the final weights responsible for adjusting for time-varying confounders up to time $t-1$.

$ES^*_{t,j}$s indicate better balance. In Appendix B, we define $ES^*_{t,j}$ and we describe our model selection approach in more detail, which aims to trade off improvements in balance with the losses in statistical efficiency which may result from models that are too complex (e.g., including all J_t confounders). Our approach resulted in selecting $c_t - 1 = 19$, 45, and 76 out of $J_t = 46$, 86, and 128 possible confounders to be included in the final models for $p_t^{den}(\alpha)$ for $t = 1$, 2, 3, respectively.

Figure 1 summarizes pictorially the balance before vs after weighting. For all t, $\max_j ES^*_{t,j} =< 0.16$ (averaged over the imputed data sets) and the average effect size (averaged over covariates and imputed data sets) was reduced from $B \geq 0.155$ (unweighted) to $B \leq 0.041$ (W_t^*-weighted). A more detailed summary and discussion is given in Appendix B and Table 16. Tables 8-10 show the results of the selected logistic regression models for $p_t^{den}(\alpha_t)$.

Table 8. Results of the logistic regression model for $p_1^{den}(\alpha_1)$

Covariate	$\widehat{\alpha}_3$	SE	z	$Pr(>\|z\|)$
(Intercept)	1.353	0.269	5.03	0.000
sati0	-1.089	0.229	-4.76	0.000
spsy0	0.050	0.016	3.15	0.002
sri70	0.016	0.012	1.33	0.185
sfs8p0	0.056	0.439	0.13	0.899
e10	-0.147	0.083	-1.77	0.077
ias5p0	0.536	0.366	1.46	0.144
race4g3	-0.913	0.189	-4.84	0.000
wkyfmpnew0	0.197	0.142	1.39	0.166
arrtot0	0.107	0.072	1.49	0.135
lri70	-0.004	0.018	-0.24	0.813
r4a0	0.001	0.002	0.66	0.511
race4g2	-0.584	0.155	-3.76	0.000
race4g4	-0.434	0.183	-2.37	0.018
hmlsrunnew0	-0.362	0.127	-2.85	0.004
emasp0	0.322	0.171	1.89	0.059
recov0	-0.125	0.139	-0.90	0.366
cws0	-0.012	0.018	-0.65	0.514
mreri13p0	0.182	0.810	0.23	0.822

Table 9. Results of the logistic regression model for $p_2^{den}(\alpha_2)$

Covariate	$\widehat{\alpha}_2$	SE	z	$Pr(>\|z\|)$	Covariate	$\widehat{\alpha}_2$	SE	z	$Pr(>\|z\|)$
(Intercept)	-2.723	0.300	-9.07	0.000	sati0	-1.183	0.277	-4.28	0.000
s7f3	1.583	0.105	15.04	0.000	s5a3	0.004	0.006	0.75	0.454
sati3	0.522	0.242	2.16	0.031	cws0	-0.015	0.013	-1.16	0.246
cei3	0.533	0.416	1.28	0.200	s2s20	0.000	0.003	0.06	0.954
maxce3	-0.001	0.006	-0.11	0.909	r4a3	0.004	0.002	2.37	0.018
spsy3	0.019	0.031	0.62	0.536	recov3	0.009	0.114	0.08	0.939
s6new3	0.154	0.127	1.22	0.224	dcs3	0.102	0.076	1.35	0.178
sdsy3	0.037	0.057	0.66	0.513	arrtot3	0.054	0.060	0.89	0.374
e13	0.195	0.087	2.24	0.025					
anytxt3	0.206	0.152	1.35	0.177					
eps7p3	0.642	0.323	1.99	0.047					
spsy0	-0.014	0.015	-0.92	0.357					
sfs8p0	0.054	0.565	0.10	0.923					
dcs0	0.008	0.053	0.16	0.876					
mreri13p0	-0.052	0.670	-0.08	0.939					
cjsi3	0.400	0.135	2.96	0.003					
eps7p0	0.039	0.296	0.13	0.895					
ers210	0.000	0.005	-0.08	0.940					
everop3	0.066	0.128	0.51	0.607					
hps3p0	0.348	0.317	1.10	0.273					
everop0	0.066	0.127	0.52	0.601					
s6new0	0.075	0.116	0.64	0.522					
arrtot0	0.061	0.046	1.32	0.186					
hps3p3	0.052	0.372	0.14	0.889					
r4a0	0.000	0.002	-0.15	0.879					
cjsi0	-0.073	0.139	-0.52	0.600					
phti43	2.417	1.566	1.54	0.123					
tas5p0	0.076	0.166	0.46	0.645					
p3new0	0.016	0.046	0.35	0.724					
wkyhr0	0.033	0.275	0.12	0.903					
mhti33	-1.054	0.912	-1.16	0.248					
emasp3	0.229	0.656	0.35	0.727					
work3	-0.260	0.454	-0.57	0.567					
cei0	-0.154	0.200	-0.77	0.442					
sri73	-0.016	0.010	-1.56	0.118					
hmlsrunnew0	0.037	0.122	0.31	0.758					
sfs8p3	-0.338	0.519	-0.65	0.516					

Table 10. Results of the logistic regression model for $p_3^{den}(\alpha_3)$

Covariate	$\hat{\alpha}_3$	SE	z	$Pr(>\|z\|)$	Covariate	$\hat{\alpha}_3$	SE	z	$Pr(>\|z\|)$
(Intercept)	-2.926	0.475	-6.16	0.000	wkyhr0	0.285	0.323	0.88	0.378
s7f6	1.353	0.139	9.75	0.000	dlyusenew6	-0.159	0.280	-0.57	0.571
s6new6	0.172	0.142	1.21	0.225	anytxt3	-0.153	0.219	-0.70	0.486
maxce3	0.014	0.007	2.01	0.044	e10	0.000	0.096	-0.001	0.999
sati3	0.114	0.284	0.40	0.687	phti43	-1.318	1.522	-0.87	0.386
cei3	-0.725	0.496	-1.46	0.144	hmlsrunnew0	0.032	0.138	0.23	0.817
s7f3	0.176	0.133	1.33	0.184	hps3p3	0.711	0.432	1.65	0.100
e13	-0.091	0.106	-0.86	0.391	ers210	-0.003	0.009	-0.34	0.733
spsy6	0.041	0.023	1.79	0.073	mhti33	-0.295	1.079	-0.27	0.785
cjsi6	0.504	0.173	2.91	0.004	s5a6	-0.006	0.008	-0.76	0.447
everop6	0.339	0.156	2.17	0.030	arrtot6	0.100	0.072	1.39	0.166
eps7p6	0.356	0.396	0.90	0.369	p3new6	0.116	0.064	1.82	0.069
spsy3	-0.028	0.040	-0.69	0.489	ias5p3	0.308	0.580	0.53	0.596
sdsy3	0.011	0.073	0.15	0.885	r4a3	0.005	0.002	2.47	0.013
s6new3	-0.392	0.161	-2.43	0.015	p3new0	-0.002	0.062	-0.03	0.973
s2s20	-0.004	0.004	-1.11	0.269	tas5p6	-0.001	0.196	-0.004	0.997
mhti36	0.798	0.850	0.94	0.348	tas5p0	0.234	0.210	1.11	0.267
emasp6	-0.015	0.177	-0.08	0.935	s7f0	0.116	0.165	0.70	0.482
s5a3	0.013	0.008	1.67	0.094	sfs8p0	0.812	0.674	1.21	0.228
emasp3	-0.643	0.873	-0.74	0.461	ers216	0.010	0.008	1.20	0.231
ias5p0	0.397	0.372	1.07	0.286	ias5p6	0.167	0.433	0.39	0.699
work3	0.294	0.594	0.50	0.620	sri70	-0.002	0.018	-0.12	0.902
cei0	0.291	0.217	1.34	0.180	mhti30	1.339	0.941	1.42	0.155
everop3	0.014	0.154	0.09	0.928	cws6	0.002	0.024	0.08	0.933
eps7p3	-0.333	0.385	-0.87	0.386	dcs0	-0.063	0.068	-0.93	0.355
cjsi3	-0.132	0.167	-0.79	0.430	race4g2	0.232	0.200	1.16	0.246
recov3	-0.175	0.141	-1.24	0.214	dcs3	0.005	0.121	0.04	0.969
bcs	-0.015	0.009	-1.63	0.103	mreri13p3	-1.697	0.924	-1.84	0.066
sri73	-0.002	0.019	-0.10	0.924	recov6	0.150	0.154	0.98	0.330
emasp0	-0.343	0.180	-1.90	0.058	p3new3	-0.012	0.064	-0.18	0.856
hps3p6	0.584	0.436	1.34	0.180	s2s23	0.005	0.004	1.14	0.253
r4a0	0.002	0.002	0.79	0.429	sfs8p6	0.094	0.964	0.10	0.922
arrtot3	0.115	0.065	1.76	0.078	mreri13p0	0.222	0.791	0.28	0.779
ers213	-0.011	0.009	-1.24	0.215	wkyhr6	0.081	0.460	0.18	0.860
phti46	-0.951	1.549	-0.61	0.539	maxce6	0.001	0.009	0.07	0.943
s6new0	-0.121	0.137	-0.88	0.380	cei6	0.500	0.667	0.75	0.454
arrtot0	0.048	0.051	0.95	0.341	e16	0.306	0.091	3.36	0.001
lri76	-0.034	0.022	-1.53	0.126	anytxt6	0.866	0.143	6.06	0.000

Step 1c. Calculate final weights. The estimated final weights $\hat{W} = \prod_{t=1}^{3} \hat{W}_t$ ranged from 0.01 to 31.74, with a mean of 1.06 (as expected, the mean of the stabilized weights should estimate unity), and median of 0.78. Following McCaffrey et al. (2004), the effective sample size $ESS = (\sum \hat{W})^2 / \sum (\hat{W})^2$ is a conservative estimate of the sample size after weighting by $\hat{W}$; that is, intuitively $n - ESS$ is an upper bound estimate of the amount of data spent on reducing or eliminating time-varying confounding using $\hat{W}$. In our data analysis, $ESS = 1165.5$.

Interestingly, if at each time point we use only the denominator propensity score in the estimation of the weights, the final weights range from 1.23 to 308.8, with a mean of 8.45, median of 3.3, and ESS=448.1. This type of comparison allows analysts to gauge the potential for improved statistical efficiency as a result of using the stabilized weights.

6.5 Exploratory data analysis

Prior to estimation, we first did 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) as $S^*_{0,2} = \text{plus16} = 1$ if $S_{0,2} = \text{b2a} \geq 16$ or $S^*_{0,2} = \text{plus16} = 0$ if $S_{0,2} = \text{b2a} < 16$. The top,

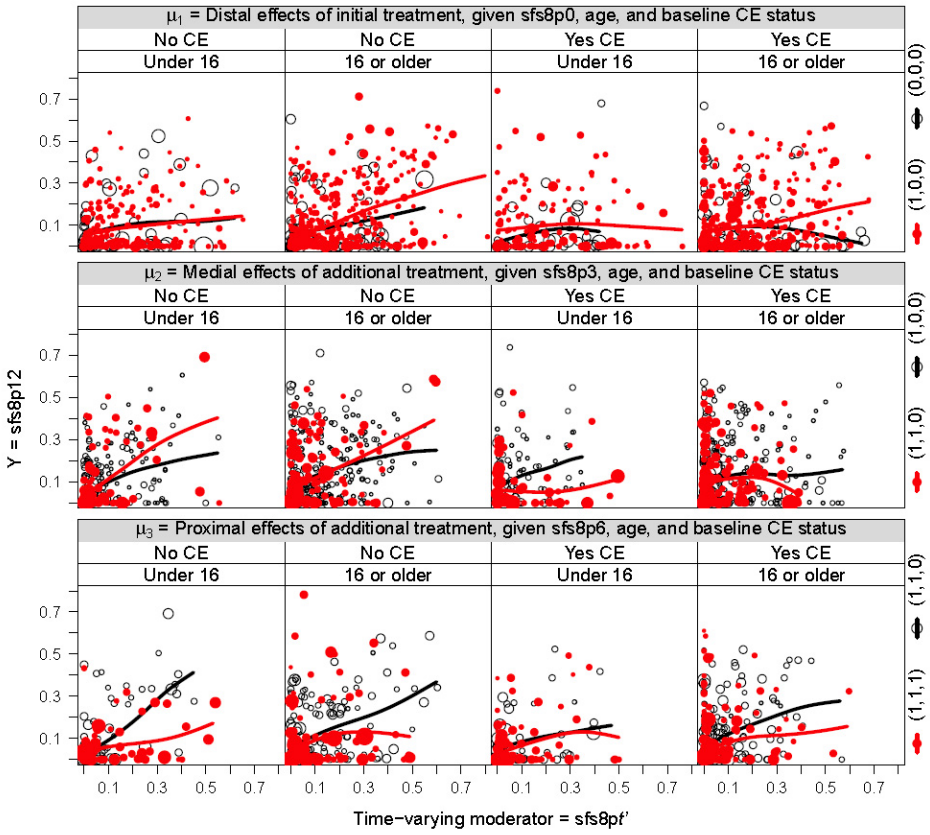


Figure 2. Exploratory data analysis of the distal, medial, and proximal effects of treatment on 12-month substance use frequency, conditional on baseline, 3-month, and 6-month substance use frequency, respectively, age (<16 vs. ≥16) and whether or not the adolescent was in a controlled environment (CE) prior to beginning any treatment. Treatment-group-specific smoothing spline curves are weighted by W_t^*; curves are plotted over the range of sfs8pt′ 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 W_t^* .

middle, and bottom four panels in Figure 2 compare data for adolescents treated with (1,0,0) vs (0,0,0), (1,1,0) vs (1,0,0), and (1,1,1) vs (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=sfs8p12; the x-axis for the top, middle, and bottom four panels are the time-varying candidate moderators S_0= sfs8p0, S_1=sfs8p3, and S_2=sfs8p6, respectively. Each panel presents a scatter plot of Y vs S_t, with smoothing curves for each treatment trajectory for each of the four combinations of whether or not the adolescent had a history of controlled environment prior to intake × whether or not the adolescent is ≥16. The points and the fitted smoothing splines were weighted by W.[4] For μ_1, the top four panels of Figure 2 suggest that the distal effect of treatment is iatrogenic among adolescents ≥16 and 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 have a 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 vs untreated clients 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. We also did similar EDAs (not shown) to inform the choice of linear model for m_j used in ε_j (for j=1,2).

6.6 Estimating the SNMM

Based on the EDA, we fitted the following SNMM to the adolescent substance use data using the IPTW+RR estimator:

$$\begin{aligned} m_Y = \beta_0 &+ \overbrace{\eta_{1,1}\delta_{1,1} + \eta_{1,2}\delta_{1,2} + \eta_{1,3}\delta_{1,3}}^{\varepsilon_1} + \overbrace{a_1\left(\beta_{1,0} + \beta_{1,1}s_{0,1}s^*_{0,2}\right)}^{\mu_1} + \overbrace{\eta_2\delta_2}^{\varepsilon_2} \\ &+ \overbrace{a_2a_1\left(\beta_{2,0} + \beta_{2,1}s_1s_{0,3} + \beta_{2,2}s_1(1-s_{0,3})\right) + a_2(1-a_1)\left(\beta_{2,3} + \beta_{2,4}s_1\right)}^{\mu_2} + \overbrace{\eta_3\delta_3}^{\varepsilon_3} \\ &+ \overbrace{a_3a_2a_1\left(\beta_{3,0} + \beta_{3,1}s_2\right) + a_3(1-a_2a_1)\left(\beta_{3,2} + \beta_{3,3}s_2\right)}^{\mu_3}, \end{aligned} \tag{19}$$

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 $m_2(\bar{s}_1, \bar{a}_2; \gamma_3) = \gamma_{3,0} + \gamma_{3,1}s_{0,1} + \gamma_{3,2}s_{0,2} + \gamma_{3,3}s_{0,3} + \gamma_{3,4}a_1 + \gamma_{3,5}s_1 + \gamma_{3,6}a_2 + \gamma_{3,7}a_2s_{0,1} + \gamma_{3,8}a_2s_1$. Table 11 describes the

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

Table 11. The meaning of the β terms for the SNMM that was fitted to the adolescent substance use treatment data

Term	Description of the effect
	Mean distal effect of initial 0-3 month treatment alone: $Y(1,0,0)$ vs $Y(0,0,0)$,
$\beta_{1,0}$	Among clients who report not using 90 days prior to intake or younger than 16 years
$\beta_{1,0}+\beta_{1,1}\, s_{0,1}$	Among clients who are older than 16 years and with baseline substance frequency use of $s_{0,1} > 0$
	Mean medial effect of additional 4-6 month treatment, $Y(1,1,0)$ vs $Y(1,0,0)$,
$\beta_{2,0}$	Among clients who report not using in months 0-3
$\beta_{2,0}+\beta_{2,1}\, s_1$	Among clients who have been in a controlled environment in the 90 days prior to intake and with months 0-3 substance use frequency of $s_1 > 0$
$\beta_{2,0}+\beta_{2,2}\, s_1$	Among clients who were not in a controlled environment in the 90 days prior to intake and with Months 0-3 substance use frequency of $s_1 > 0$
	Mean medial effect of 4-6 month treatment alone, $Y(0,1,0)$ vs $Y(0,0,0)$,
$\beta_{2,3}$	Among clients who report not using in months 0-3
$\beta_{2,3}+\beta_{2,4}\, s_1$	Among clients with months 0-3 substance use frequency of $s_1 > 0$
	Mean proximal effect of additional 7-9 month treatment, $Y(1,1,1)$ vs $Y(1,1,0)$,
$\beta_{3,0}$	Among clients who report not using in months 4-6
$\beta_{3,0}+\beta_{3,1}\, s_1$	Among clients with months 4-6 substance use frequency of $s_2 > 0$
	Mean proximal effect of 7-9 month treatment & no or inconsistent past treatment, $Y(a_1, a_2, 1)$ vs $Y(a_1, a_2, 0)$ with $a_1 a_2 \neq 1$,
$\beta_{3,2}$	Among clients who report not using in months 4-6
$\beta_{3,2}+\beta_{3,3}\, s_1$	Among clients with months 4-6 substance use frequency of $s_2 > 0$

meaning of the βs within the treatment effect terms, the μ_ts. As described in subsection 3.2, the η_ts are nuisance parameters, which model the association (causal and non-causal effects) of the time-varying moderators with the outcome (under no treatment); these are like the "main effects" of the time-varying moderators.

Estimates for $\gamma_{1,j} \forall j$ are given in Table 6. Table 12 presents the results for the weighted regression estimates of γ_2 and γ_3 for m_2 and m_3, respectively. We note that similar to how we use propensity score models in the weights as a tool to eliminate or reduce time-varying confounding bias, we use the δ_ts based on the m_ts as a tool to get around the problems with the standard regression method. Therefore, our aim is not to

Table 12. Results of the weighted regression models for m_t (for t=2,3)

t	Term	γ_t	$\hat{\gamma}_t$	SE	z	$Pr(>\|z\|)$
2	(Intercept)	$\gamma_{2,0}$	-0.036	0.084	-0.43	0.67
	sfs8p0	$\gamma_{2,1}$	0.204	0.069	2.97	<0.01
	b2a	$\gamma_{2,2}$	0.005	0.005	1.02	0.31
	ce	$\gamma_{2,3}$	-0.014	0.005	-3.10	<0.01
	anytxt3	$\gamma_{2,4}$	0.133	0.088	1.52	0.13
	anytxt3:sfs8p	$\gamma_{2,5}$	-0.070	0.070	-0.99	0.32
	anytxt3:b2a	$\gamma_{2,6}$	-0.008	0.005	-1.46	0.14
3	(Intercept)	$\gamma_{3,0}$	0.066	0.032	2.05	0.04
	sfs8p0	$\gamma_{3,1}$	0.106	0.017	6.16	<0.01
	anytxt3	$\gamma_{3,4}$	-0.003	0.007	-0.39	0.69
	sfs8p3	$\gamma_{3,5}$	0.457	0.050	9.05	<0.01
	anytxt6	$\gamma_{3,6}$	-0.006	0.009	-0.73	0.46
	b2a	$\gamma_{3,2}$	-0.002	0.002	-0.84	0.40
	ce	$\gamma_{3,3}$	-0.003	0.007	-0.48	0.63
	anytxt6:sfs8p	$\gamma_{3,7}$	-0.031	0.075	-0.41	0.68
	anytxt6:ce	$\gamma_{3,8}$	-0.006	0.011	-0.54	0.59

interpret the estimated γ_2 and γ_3 in m_2 and m_3, respectively, as causal. In Section 7, we discuss additional assumptions that are necessary, and other challenges, to interpreting γ_2 and γ_3 as causal.

Table 13 presents the IPTW+RR estimates for the SNMM in equation 19. For comparison, we report estimated asymptotic standard errors and bootstrap standard errors for the estimated SNMM. As expected according to our simulation study, given our large sample, the asymptotic and bootstrap estimates of the standard errors are nearly identical.

To facilitate interpretation of the fitted SNMM, Table 14 reports estimates, effect sizes, and confidence intervals and p-values (using bootstrap standard errors) for the linear contrasts described in words in Table 11. 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. Second, we find that among adolescents who are severe at the end of month 3, those who report receiving $\bar{A}_3 = (0,1,0)$ vs those never reporting receiving treatment $\bar{A}_3 = (0,0,0)$ also do poorly at the end of the year (a moderate to large effect size, ES=0.862, P-val=0.10). Again, this is consistent with the conjecture above the insufficiency of just

Table 13. Results of the IPTW+RR estimates of a SNMM, examining the time-varying moderated effects of additional adolescent substance use treatment

			Bootstrap SE		Asymptotic SE	
Term	θ	$\hat{\theta}$	SE	$Pr(>\|z\|)$	SE	$Pr(>\|z\|)$
(Intercept)	β_0	0.094	0.010	<0.01	0.010	<0.01
$\delta_{1,1}$	$\eta_{1,1}$	0.124	0.027	<0.01	0.028	<0.01
$\delta_{1,2}$	$\eta_{1,2}$	0.004	0.003	0.12	0.003	0.11
$\delta_{1,3}$	$\eta_{1,3}$	#	0.008	0.99	0.008	0.99
anytxt3	$\beta_{1,0}$	-0.001	0.011	0.94	0.011	0.94
anytxt3:sfs8p0:plus16	$\beta_{1,1}$	0.065	0.035	0.06	0.038	0.09
δ_2	η_2	0.268	0.046	<0.01	0.047	<0.01
anytxt6:anytxt3	$\beta_{2,0}$	-0.009	0.010	0.35	0.011	0.41
anytxt6:anytxt3:sfs8p0:ce	$\beta_{2,1}$	-0.080	0.083	0.33	0.097	0.41
anytxt6:anytxt3:sfs8p0:(1-ce)	$\beta_{2,2}$	0.066	0.106	0.54	0.117	0.57
anytxt6:(1-anytxt3)	$\beta_{2,3}$	-0.033	0.024	0.17	0.026	0.21
anytxt6:(1-anytxt3):sfs8p3	$\beta_{2,4}$	0.294	0.261	0.26	0.315	0.35
δ_3	η_3	0.483	0.047	<0.01	0.052	<0.01
hlineanytxt9:anytxt6:anytxt3	$\beta_{3,0}$	-0.006	0.011	0.59	0.012	0.63
anytxt9:anytxt6:anytxt3:sfs8p6	$\beta_{3,1}$	-0.338	0.080	<0.01	0.085	<0.01
anytxt9:(1-anytxt6:anytxt3)	$\beta_{3,2}$	0.021	0.017	0.21	0.018	0.26
anytxt9:(1-anytxt6:anytxt3):sfs8p6	$\beta_{3,3}$	-0.359	0.091	<0.01	0.102	<0.01

# Indicates that |estimate| is smaller than 1×10^{-3}.

some (or one bout of) treatment. Finally, 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. 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).

Finally, Table 15 shows how the estimates of θ compare across the three estimators IPTW+RR, 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 the estimates of IPTW+RR are more negative (further away from zero). We conjecture 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 get 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 elminate. 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 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 naïvely conditioning on month 3 use) that is reduced or eliminated by the IPTW+RR and RR estimators (see Almirall et al., 2011). 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) estimatesobtained under RR and TRAD. We do not have a reasonable explanation for this difference in estimates.

Table 14. Linear contrasts of interest for the SNMM fitted with IPTW+RR

Contrast #	EST	ES†	95%CI‡	$Pr(>\|z\|)$
$\beta_{1,0}$	-0.001	-0.006	(-0.174, 0.161)	0.94
$\beta_{1,0} + \beta_{1,1}$ 0.5	0.032	0.240	(-0.038, 0.518)	0.09
$\beta_{2,0}$	-0.009	-0.068	(-0.211, 0.074)	0.35
$\beta_{2,0} + \beta_{2,1}$ 0.5	-0.049	-0.371	(-0.951, 0.210)	0.21
$\beta_{2,0} + \beta_{2,2}$ 0.5	0.024	0.179	(-0.555, 0.913)	0.63
$\beta_{2,3}$	-0.033	-0.248	(-0.559, 0.062)	0.117
$\beta_{2,3} + \beta_{2,4}$ 0.5	0.114	0.862	(-0.193, 1.916)	0.109
$\beta_{3,0}$	-0.006	-0.042	(-0.216, 0.132)	0.63
$\beta_{3,0} + \beta_{3,1}$ 0.5	-0.174	-1.32	(-1.952, -0.679)	<0.01
$\beta_{3,2}$	0.021	0.158	(-0.058, 0.374)	0.15
$\beta_{3,2} + \beta_{3,3}$ 0.5	-0.159	-1.20	(-1.629, -0.766)	<0.01

See Table 11 for an explanation, in words, of these linear contrasts.

† Effect sizes are calculated as EST/SD(Y) = EST/0.17.

‡ 95% confidence intervals for the effect sizes are calculated using (EST ± 1.96 × BOOTSE)/0.17, where BOOTSE is the bootstrap standard error for the linear contrast. Results were near identical when the ASEs were used. P-values were calculated based on a Wald statistics using BOOTSE.

Table 15. Comparing estimates of the β parameters using IPTW+RR vs RR alone vs traditional regression estimator†

		Estimates of θ		
Term	θ	IPTW+RR	RR	TRAD
(Intercept)	β_0	0.094	0.091	0.13
$\delta_{1,1}$	$\eta_{1,1}$	0.124	0.121	0.029
$\delta_{1,2}$	$\eta_{1,2}$	0.004	#	0.002
$\delta_{1,3}$	$\eta_{1,3}$	#	-0.003	0.003
anytxt3	$\beta_{1,0}$	-0.001	#	-0.002
anytxt3:sfs8p0:plus16	$\beta_{1,1}$	0.065	0.074	0.080
δ_2	η_2	0.268	0.333	0.115
anytxt6:anytxt3	$\beta_{2,0}$	-0.009	-0.005	#
anytxt6:anytxt3:sfs8p0:ce	$\beta_{2,1}$	-0.080	-0.127	-0.080
anytxt6:anytxt3:sfs8p0:(1-ce)	$\beta_{2,2}$	0.066	-0.054	-0.019
anytxt6:(1-anytxt3)	$\beta_{2,3}$	-0.033	-0.001	0.004
anytxt6:(1-anytxt3):sfs8p3	$\beta_{2,4}$	0.294	-0.090	-0.048
δ_3	η_3	0.483	0.439	
anytxt9:anytxt6:anytxt3	$\beta_{3,0}$	-0.006	-0.012	
anytxt9:anytxt6:anytxt3:sfs8p6	$\beta_{3,1}$	-0.338	-0.194	
anytxt9:(1-anytxt6:anytxt3)	$\beta_{3,2}$	0.021	0.002	
anytxt9:(1-anytxt6:anytxt3):sfs8p6	$\beta_{3,3}$	-0.359	-0.293	

# Indicates that |estimate| is smaller than 1×10^{-3}.
† The traditional regression estimator fits the same linear parametric model as fitted using the IPTW+RR and RR estimators except that the δ_ts are replaced with S_t; that is, in the traditional regression estimator, the "main effects" of the time-varying moderators are not residualized as they are with the ITPW+RR and RR estimators.

7. Discussion

This manuscript presents an application of the SNMM (Robins, 1994) for examining time-varying causal effect moderation. In interventions research developing time-varying treatments, 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 (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 (Murphy, 2005), 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 (Murphy, van der Laan, Robins, & CPPRG, 2001; Murphy & Almirall, 2009).

The methods presented in this article can be used as preliminary analyses for, or be supplemented by, more sophisticated analyses or methodological development aimed explicitly at developing optimal dynamic

treatment regimes (Murphy, 2003; Robins, 2004; Hernán, Lanoy, Costagliola, & Robins, 2006; Rosthøj et al., 2009; Moodie, Richardson, & Stephens, 2007; Orellana, Rotnitzky, & Robins, 2010; Henderson et al., 2010). In particular, Henderson et al. (2010) use regret-regression (Murphy, 2003), which is the analogue to the RR estimator (Almirall et al., 2009) we build on in this manuscript, for estimating optimal dynamic treatment regimes.

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, the proposed IPTW+RR estimator can be used to obtain high quality starting values for the G-estimator (Robins, 1994; Almirall et al., 2009) of the SNMM. The G-estimator has an advantage over the IPTW+RR of not requiring correct specification for the nuisance functions. However, as suggested by Almirall and colleagues (2009), this may come at a cost in terms of statistical efficiency. The IPTW+RR may help improve this efficiency by providing high-quality guesses for estimates of the nuisance functions in the G-estimator. Second, in the context of the SNMM, the IPTW+RR shows that it is possible to hybridize parametric estimators and IPTW estimators (often thought of as "semi-parametric" in the time-varying covariates). 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, behavioral scientists may first understand how to fit the SNMM using the RR+IPTW estimator prior to moving to more sophisticated estimators such as the G-estimator. 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 which allows scientists to deal with the nuisance of time-varying confounding bias, all the while reserving the linear model for examining scientific questions of interest (in this case, moderated time-varying causal effects). In particular, this methodology helps clarify the difference between the SNMM and Robins' marginal structural model (Robins et al., 2000), which is more commonly used in the epidemiological, behavioral, and medical sciences. Indeed, by averaging over the $(S_t(\overline{a}_{t-1}))$ in the estimated SNMM, it is possible to obtain estimates of the marginal structural model.

Simulation experiments were presented in Section 5 to illustrate the methodology under various scenarios. More careful simulation experiments could be conducted to quantify the biases incurred under more 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 (1988) $d \approx 0.1$ 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. In addition, 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. In other scenarios, such as when there is no time-varying confounding and the time-varying moderator S_t is not a mediator of the effect of treatment (as it is in the data generative models shown in Section 5), all three estimators can yield identical point estimates! We also do not show conditions under which the assumptions of the IPTW+RR are violated–namely, conditions under which not all confounders were modeled as part of the propensity scores making up the weights; in such cases, the IPTW+RR can also be biased.

References

Almirall, D., Coffman, C., Yancy, W., & Murphy, S. (2010). Maximum likelihood estimation of the structural nested mean model using SAS PROC NLP. In D. Faries, A. Leon, J. Haro, & B. Obenchain (Eds.), *Analysis of observational health-care data using SAS*. SAS Press.

Almirall, D., McCaffrey, D., Ramchand, R., & Murphy, S. (2011). Subgroups analysis when treatment and moderators are time-varying*., Prevention Science,* Advance online publication.

Almirall, D., Ten Have, T., & Murphy, S. (2009). Structural nested mean models for assessing time-varying effect moderation. *Biometrics, 66*(1), 131–139.

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology, 51*, 173–1182.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Earlbaum Associates.

Cole, S., Hernan, M., Robins, J., Anastos, K., Chmiel, J., Detels, R., … Munoz, A. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology, 158*(7), 687–694.

Cole, S., Platt, R., Schisterman, E., Chu, H., Westreich, D., Richardson, D., & Poole, C. (2010). Illustrating bias due to conditioning on a collider, *International Journal of Epidemiology, 39*

Dennis, M. L., Titus, J. C., White, M. K., Unsicker, J. I., & Hodgkins, D. (2002). Global Appraisal of Individual Needs (GAIN): Administration guide for the GAIN and related measures. Bloomington, IL: Chestnut Health Systems. Available online at http://www.chestnut.org/li/gain.

Henderson, R., Ansell, P., & Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes, *Biometrics, 66*(4), 1192–201.

Henderson, R., Ansell, P., & Alshibani, D. (2011). Optimal dynamic treatment methods. *REVSTAT Statistical Journal, 9*(1), 19–36.

Hernán, M., Brumback, B., & Robins, J. (2000). Marginal structural models to estimate the causal effect of zidovudi ne on the survival of HIV-positive men, *Epidemiology, 11*, 561-70.

Hernán, M., Lanoy, E., Costagliola, D., & Robins, J. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic and Clinical Pharmacology and Toxicology, 98*, 237-242.

Holland, P. (1986). Statistics and causal inference, *Journal of the American Statistical Association, 81*, 945–970.

Kraemer, H. C., Wilson, G., & Fairburn, C. (2002). Mediators and M*oderators of Treatment Effects in Randomized Clinical Trials. Archives of General Psychiatry, 59*, 877–883.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403–425.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.) Boca Raton, FL: Chapman and Hall/CRC.

Moodie, E., Richardson, T., & Stephens, D. (2007). Demystifying optimal dynamic treatment regimes, *Biometrics, 63,* 447-455.

Murphy, S. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine, 24*(10), 1455–1481.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B, 65*(2), 331–366.

Murphy, S. A., van der Laan, M. J., Robins, J. M., & CPPRG (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association, 96*, 1410–1423.

Murphy, S., & Almirall, D. (2009). Dynamic treatment regimens. In M. W. Kattan (Ed.), *Encyclopedia of medical decision making* (pp. 419–422). Thousand Oaks, CA: Sage Publications.

Orellana, L., Rotnitzky, A., & Robins, J. (2010). Dynamic Regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. *International Journal of Biostatistics, 6*(2).

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research, 27*, 226–284.

Petersen, M., Deeks, S., Martin, J., & van der Laan, M. (2007). History-adjusted marginal structural models to estimate time-varying effect modification. *American Journal of Epidemiology, 166*(9), 985–93.

Petersen, M., & van der Laan, M. (2007). Response to invited commentary on 'History-adjusted marginal structural models to estimate time-varying effect modification.' *American Journal of Epidemiology, 166*(9), 1003–1004.

Raghunathan, T., Lepkowski, J., Hoewyk, J., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, Survey Methodology, 27, 85–95.

Raghunathan, T., Solenberger, P., & Hoewyk, J. (2002). *IVEware: Imputation and variance estimation software.* Survey Research Center, Institute for Social Research, Ann Arbor, MI.

Robins, J. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease, 40*(Supplement 2), 139s–161s.

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

Robins, J. (1997a). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality.* Lecture Notes in Statistics, New York: Springer-Verlag.

Robins, J. (1997b). Estimating causal effects of time-varying endogenous treatments by Gstimation of structural nested models. In M. Berkane (Ed.), *Latent variable modeling and applications to causality*. Lecture Notes in Statistics, New York: Springer, pp. 69–117.

Robins, J. (2004). Optimal structural nested models for optimal sequential decisions. In D. Lin, & P. Heagerty (Eds.), *Proceedings of the Second Seattle Symposium on Biostatistics* (pp. 189–326). New York, NY: Springer.

Robins, J., Hernán, M., & Rotnitzky, A. (2007). Invited commentary on 'History-adjusted marginal structural models to estimate time-varying effect modification.' *American Journal of Epidemiology, 166*(9), 994–1002.

Robins, J. M. (1989). The control of confounding by intermediate variables. *Statistics in Medicine, 8*, 679–701.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese, 121*, 151–179.

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

Rosthøj, S., Keiding, N., & Schmiegelow, K. (2009). 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, 31*(5). 470–88.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley.

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

Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall / CRC Press.

van der Laan, M. J., Murphy, S. A., & Robins, J. M. (2002). Analyzing dynamic regimes using structural nested mean models. Unpublished Manuscript.

van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Series in Statistics, New York, NY: Springer-Verlag.

Appendix A. Data Generating Model for the Simulation Experiments

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

$U \sim N(0,\sigma)$,	$\sigma=0.1$
$S_0 \sim N(\gamma_0,\sigma)$,	$\gamma_0=0.4$
$S_1(a_1) \mid S_0, U \sim N(F_1\gamma_1 + \gamma_U U, \sigma)$,	$F_1=(1, a_1, s_0, a_1 s_0)$,
	$\gamma_1=(0.5,-0.5,0.1,-0.1)^T$,
	$\gamma_U=0.1$
$\mu_1(S_0,a_1) = \beta_{11}a_1 + \beta_{12}a_1 S_0$,	$\beta_{11=}-0.1, \beta_{12=}-0.1$
$\mu_2(\bar{S}_1(a_1),\bar{a}_2) = \beta_{21}a_2 + \beta_{22}a_2 S_1(a_1)$,	$\beta_{21=}-0.1, \beta_{22=}-0.1$
$\varepsilon_1(U,S_0) = \eta_1(S_0-\gamma_0)+\gamma_U U$,	$\eta_1=0.15$
$\varepsilon_2(U,S_0,a_1,S_1(a_1)) = \eta_2(S_1 - F_1\gamma_1 - \gamma_U U)$,	$\eta_2=0.30$
$Y(a_1,a_2) \mid U, \bar{S}_1(a_1) \sim N(\beta_0+\varepsilon_1+\mu_1+\varepsilon_2+\mu_2,\sigma)$,	$\beta_0=0.8$
$X_1(a_1) \mid Y(\bar{a}_2) \sim N\left(\rho_1+\rho_2 a_1+\rho_3\sum_{a_1,a_2\in(0,1)} Y(a_1,a_2), \sigma/2\right)$,	$\rho_1=-0.65$
	$\rho_2=-0.25$
	$\rho_3=0.50$
$X_0 \mid Y(\bar{a}_2) \sim N\left(\rho_4\sum_{a_1,a_2\in(0,1)} Y(a_1,a_2), \sigma\right)$,	$\rho_4=0.25$
$A_1 \mid S_0, X_0 \sim \text{Bernoulli}\left(p=\Lambda(G_1\alpha_1)\right)$,	$G_1=(1,S_0,X_0)$,
	$\alpha_1=(-1.3,1.5,2.5)$
$S_1 = A_1 S_1(1) + (1-A_1)S_1(0)$	
$A_2 \mid S_1, X_1 \sim \text{Bernoulli}\left(p=\Lambda(G_2\alpha_2)\right)$,	$G_2=(1,S_1,X_1)$,
	$\alpha_2=(-1.3,2.0,2.5)$
$Y = Y(A_1,A_2) = \sum_{a_1,a_2\in(0,1)} I(A_1=a_1, A_2=a_2) Y(a_1,a_2)$	

The parameter values given above were used 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).

A.1 Deriving the SNMM for the conditional mean of $Y(a_1,a_2)$ given $(S_0, S_1(a_1))$

In the above, Y is generated according to the following known SNMM for $Y(a_1,a_2)$ conditional on $U(S_0, S_1(a_1))$

$$m_Y(U,S_0,S_1(a_1)) = E(Y(a_1,a_2) \mid U,S_0,S_1(a_1)) =$$
$$\beta_0 + \varepsilon_1(U,S_0) + \mu_1(S_0,a_1) + \varepsilon_2(U,S_0,a_1,S_1(a_1)) + \mu_2(\bar{S}_1(a_1),\bar{a}_2). \quad \text{(A1)}$$

By definition $\varepsilon_1(U,S_0) = E(Y(0,0) \mid U,S_0) - E(Y(0,0)$ and it is generated using the linear form $\varepsilon_1(U,S_0) = \eta_1(S_0 - E(S_0)) + \gamma_U(U - E(U))$ where $E(S_0)=\gamma_0$ and $E(U)=0$; whereas, by definition

$\varepsilon_2(U, S_0, a_1, S_1(a_1)) = E(Y(a_1,0) \mid U, S_0, S_1(a_1)) - E(Y(a_1,0) \mid U, S_0)$ and it is generated using $\varepsilon_2(U, S_0, a_1, S_1(a_1)) = S_1(a_1) - E(S_1(a_1) \mid U, S_0)$ where $\gamma_1 F_1 + \gamma_U U$ 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_ts are independent of *U*), nor an unmeasured confounder (i.e., *U* is not used to generate A_t), nor a mediator (i.e., it is a baseline measure).

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 $E(m_Y(U, S_0, S_1(a_1)) \mid S_0, S_1(a_1))$. Since only the ε_ts are a function of *U*, all we need is

$$
\begin{aligned}
&E(\varepsilon_1(U, S_0) + \varepsilon_2(U, S_0, a_1, S_1(a_1)) \mid S_0, S_1(a_1)) \\
&\quad = E(\eta_1(S_0 - E(S_0)) + \gamma_U U \mid S_0, S_1(a_1)) + E(\eta_2(S_1(a_1) - E\left(S_1(a_1) \mid U, S_0\right)) \mid S_0, S_1(a_1)) \\
&\quad = \eta_1(S_0 - E(S_0)) + \gamma_U E(U \mid S_0, S_1(a_1)) + \eta_2(S_1(a_1) - E(E\left(S_1(a_1) \mid U, S_0\right) \mid S_0)), \\
&\quad = \eta_1(S_0 - \gamma_0) + \gamma_U E(U \mid S_0, S_1(a_1)) + \eta_2(S_1(a_1) - (\gamma_1 F_1 + \gamma_U E\left(U \mid S_0\right))), \qquad \text{(A.2)}
\end{aligned}
$$

which relies on knowing the conditional means $E(U|S_0)$ and $E(\mathrm{U}|\mathrm{S}_0, S_1(a_1))$. Since *U* is independent of S_0 and $E(U)=0$, then $E(U|S_0)=0$. Using standard normal theory, it can be shown that

$E(U \mid S_0, S_1(a_1)) = \frac{\gamma_U}{1+\gamma_U^2}(S_1(a_1) - \gamma_1 F_1)$. It follows that

$$E(\varepsilon_1(U, S_0) + \varepsilon_2(U, S_0, a_1, S_1(a_1)) \mid S_0, S_1(a_1)) = \eta_1(S_0 - \gamma_0) + (\eta_2 + \frac{\gamma_U^2}{1+\gamma_U^2})(S_1(a_1) - \gamma_1 F_1).$$

Therefore, the following design matrix can be used 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. Choosing the Denominator Propensity Score Models.

As discussed in Section 4, the denominator propensity score $p_t^{den}(\alpha_t)$ is used in the weights to balance treated and untreated groups based on time-varying confounders. For each *t*, we used the following steps to choose the models for $p_t^{den}(\alpha_t)$:

1b(i). **Examine unweighted balance.** For each *t*, examine the balance in the distribution of the covariates in $\bar{V}_{t-1}$ between treated (A_t=1) and untreated (A_t=0) groups. We do this by estimating the absolute standardized difference in means (i.e., effect size), denoted $ES_{j,t-1}$, for each covariate $V_{j,t-1}$ in $\bar{V}_{t-1}$. Following Cohen (1988), for continuous $V_{j,t-1}$, $ES_{j,t-1}$ is defined as

$$ES_{j,t-1} = |\tilde{V}^1_{j,t-1} - \tilde{V}^0_{j,t-1}| / \sigma_{j,t-1},$$

where $\tilde{V}^1_{j,t-1}$ is the average of $V_{j,t-1}$ for clients observed under A_t=1, $\tilde{V}^0_{j,t-1}$ is the average of $V_{j,t-1}$ for clients observed under A_t=0, and $\sigma_{j,t-1}$ is the standard deviation of $V_{j,t-1}$. For binary $V_{j,t-1}$, we set $\sigma_{j,t-1} = 1$ (the effect size is a difference in proportions). Rank-ordering the $ES_{j,t-1}$s allows examination of the most and least influential putative confounders at time *t*. The balance score, BAL_t, obtained by averaging the $ES_{j,t-1}$s provides a sense for the magnitude of the observed time-varying confounding at time *t*.

1b(ii). **Estimate the denominator model.** For each *t*, estimate the denominator propensity score p_t^{den} using a logistic regression model for $Pr(A_t = 1 | \bar{V}_{t-1}, \bar{A}_{t-1})$ with unknown parameters α_t. Calculate and save the $\hat{p}_t^{den}(\hat{\alpha}_t)$ s.

1b(iii). **Calculate working weights** $\boldsymbol{W_t^* = \frac{A_t}{\hat{p}^{den}(\hat{\alpha})} + \frac{1 - A_t}{1 - \hat{p}^{den}(\hat{\alpha})}}$. These are the denominator-only weights. These are not the weights used in the final weighted regression of the SNMM. These can be seen as the part of the final weights responsible for adjusting for time-varying confounders.

1b(iv). **Examine weighted balance and other diagnostic statistics.** For each *t*, examine the balance in the distribution of the covariates as in Step 1b(i), except replace $\tilde{V}^{A_t}_{j,t-1}$ with the W_t^*-weighted average. That is, for each $V_{j,t-1}$, calculate the W_t^*-weighted effect size, denoted $ES^*_{t,j}$. Calculate the weighted balance score *WBAL* by averaging the weighted effect sizes. If the weights have been successful at reducing imbalances, the weighted balance *WBAL* should be smaller than *BAL* (first-order success). Ideally, the largest imbalance is a small one (second-order success); that is, $\max_j ES_{j,t-1} \leq 0.20$

(Cohen, 1988). A third-order measure of success is the working effective sample size ESS_t^* of the W_t^*-weighted sample. Following Mccaffrey et al. (2004), we define $ESS_t^* = (\sum W_t^*)^2 / \sum (W_t^*)^2$. Typically, $ESS^* < n$; $n - ESS^*$ is an estimate of the sample size that is spent adjusting for the putative confounders at time *t* using W_t^*. ESS^* also measures the variability in the weights: very large weights $\Leftrightarrow$ small ESS^*, and vice-versa.

1b(v). **Select/fine-tune form of denominator model.** In this step, we repeat Steps 1b(ii)-1b(iv), each time fitting different models for $p_t^{den}(\alpha)$ and obtaining new working weights W_t^*. We seek to choose the denominator propensity score model which trades off the best balance with the largest ESS. Doing this is more art than science! We used the following 4-step approach which appeared to work well (see Figure 3):

1) **All confounders.** First, we fit a denominator model with all confounders. Results: As shown in Figure 3, this step had the most significant effect on balance, as expected. For all imputed data sets, the balance improved from $BAL > 0.15$ to $WBAL < 0.04$ based on this step alone. Notably, for $t=2$, the balance ($WBAL < 0.015$), distribution of the weights, and effective sample size ($ESS > 1750$) diagnostics are superb based on this step alone.

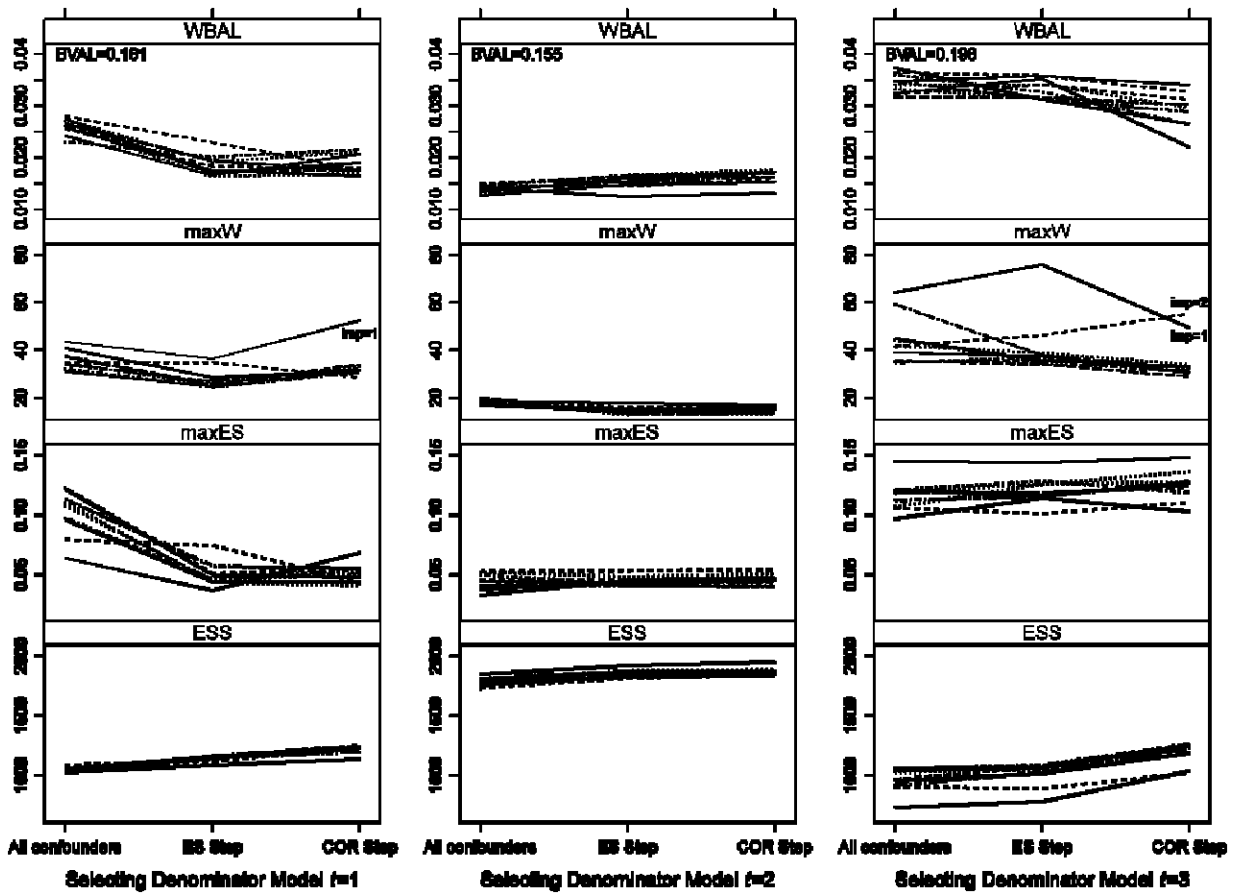
2) **ES Step.** Second, we chose to limit the model by removing candidate confounders with $ES_t < 0.10$. Results: As shown in Figure 3, for $t=1$ this step resulted in significant improvements in *WBAL* by reducing the size of $\max ES_1$, whereas for $t=2,3$ *WBAL* remained about the same. Improvements were made in *ESS* for $t=1,2$, whereas *ESS* remained about the same for $t=3$. Bucking the trend in the other 8 imputed data sets, $\max W_3^*$ became larger in imputed data sets 1 and 2.

3) **COR Step.** Third, we chose to limit the models further by removing redundant candidate confounders from the model. To do this, we obtained correlations between all pairs of candidate confounders still in the model after the ES Step. For each confounder-pair with |correlation|>0.5, we removed from the model the confounder in the pair with the smallest *ES*. Results: As shown in Figure 3, *WBAL* remained about the same for $t=1$, but significant improvements were made on *WBAL* for $t=3$. *ESS* continued to improve for all *t*, especially for $t=3$, though. It is intuitive that this step would result in the greatest improvement for $t=3$ (especially for *ESS*) given the large number of covariates still present in this model after the ES step. This step considerably improved the $\max W_3^*$ for imputed data set 1; however, bucking the trend in all other imputed data sets,

this step resulted in a slightly worse $\max W_3^*$ for imputed data set 2.

4) **Union Step.** The first three steps above result in different confounders being included in each of the imputed data sets. However, for inferential purposes, we sought to have one "unique" model for $p_t^{den}(\alpha)$ (that we can fit to and summarize over the different imputed data sets). In this fourth step, we chose to include in the final model for $p^{den}(\hat{\alpha})$, the union of all covariates selected for each of the imputed data sets. Results: Table 16 shows balance measures and other diagnostics for the final selected model at each time point. The statistics shown in Table 16 are for the same model fitted to each of the ten imputed data sets. $c_t - 1 = 19$, 45, and 76 confounders out of $J_t = 46$, 86, and 126 were included in the final models for $p_t^{den}(\alpha)(t = 1,2,3)$. In every case, $\max W_t^*$ (averaged over the imputed data sets) was below 0.20 which is one of our benchmarks for success. Figure 1 summarizes pictorially the balance before vs after weighting and the

Figure 3. Balance, distribution of the working weights , and effective sample size diagnostics for different models for p_t^{den} at each step of refinement (see Appendix B).



distribution of the W_t^* weights (averaged over the ten imputed data sets). (Note: we also investigated selecting the final model based on the intersection of all covariates selected for each imputed data set, but the union approach resulted in far better balance; these results not shown.)

Table 16. A summary of the selected logistic regression models chosen for $p_t^{den}(\alpha_t)$, the weights W_t^* calculated using $p_t^{den}(\alpha_t)$, and balance in the W_t^*-weighted sample

	Probabilities, $\hat{p}_t^{den}$			Denominator weights, $\hat{W}^*$				Balance		
t	c−1	(min, max)	med	(min, max)	med	$\overline{\max}$	ESS	J	B	M
1	19	(0.20,0.98)	0.86	(1.02,32.83)	1.18	47.63	1214.8	46	0.041	0.13
2	45	(0.03,0.95)	0.40	(1.03,15.43)	1.51	15.92	1867.7	86	0.024	0.13
3	76	(0.01,0.97)	0.15	(1.01,37.93)	1.18	49.02	1057.0	126	0.037	0.16

C_t-1 is the number of variables (number of parameters c_t minus 1 for the intercept) selected for the logistic regression model for $p_t^{den}(\alpha_t)$. Appendix B describes our approach to selecting the logistic regression models.

The denominator weights are defined as $W_t^* = A_t / p_t^{den} + (1-A_t)/(1-p_t^{den})$.

(min, max) and med are used to denote the range and median for the probabilities and the weights. Each statistic is reported as the average over the imputed data sets.

$\overline{\max}$ is the maximum weight over all units and across all imputed data sets.

$ESS_t = (\sum W_t^*)^2 / \sum (W_t^*)^2$ is the effective sample size of the W_t^*-weighted sample. For comparison, the total sample size is $n=2870$. The quantity reported is the average over the imputed data sets.

J_t is the total number of covariates (putative confounders) available up to time t used to calculate balance measures.

B_t is the average (over the imputed data sets) of the weighted balance scores $WBAL_t$. $WBAL_t$ is the average of $ES_{t,j}^*$ ($j=1,\ldots,J_t$), the W^*-weighted effect sizes (there is one for each covariate j, see text for definition).

M_t is the average (over the imputed data sets) of $\max_j ES_{t,j}^*$.

Appendix C. Asymptotic Standard Errors of the IPTW+RR Estimator

Asymptotic standard errors (ASE) 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 $\hat{\pi}_t$ 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 $\hat{\alpha}_t$ 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 $V_t = (S_t, X_0)$). Based on 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 $J_{\alpha_t} = E \dfrac{\partial}{\partial \alpha_t} \psi_{\alpha_t}$ ($b_t \times b_t$ matrix) and $J_{\pi_t} = E \dfrac{\partial}{\partial \pi_t} \psi_{\pi_t}$ ($c_t \times c_t$ 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). Based on 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 $J_{\gamma_t} = E \dfrac{\partial}{\partial \gamma_t} \psi_{\gamma_t}\left(\bar{\alpha}_t, \bar{\pi}_t\right)$ ($k_t \times k_t$ matrix), $J_{\gamma_t \alpha_j} = E \dfrac{\partial}{\partial \alpha_j} \psi_{\gamma_t}\left(\bar{\alpha}_t, \bar{\pi}_t\right)$ ($k_t \times b_j$ matrix), and $J_{\gamma_t \pi_j} = E \dfrac{\partial}{\partial \pi_j} \psi_{\gamma_t}\left(\bar{\alpha}_t, \bar{\pi}_t\right)$ ($k_t \times c_j$ matrix) where $\left(\bar{\alpha}_t, \bar{\pi}_t\right) = \left(\alpha_1, \ldots, \alpha_t, \pi_1, \ldots, \pi_t\right)$. Finally, the estimates $\hat{\theta}$ 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 $\tilde{\psi}_{\gamma_t} = \psi_{\gamma_t} - \sum_{j=1}^{t} J_{\gamma_t \alpha_j} J_{\alpha_j}^{-1} \psi_{\alpha_j} - \sum_{j=1}^{t} J_{\gamma_t \pi_j} J_{\pi_j}^{-1} \psi_{\pi_j}$. Based on 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 $J_\theta = E \dfrac{\partial}{\partial \theta} \psi_\theta$ ($d \times d$ matrix), $J_{\theta \gamma_t} = E \dfrac{\partial}{\partial \gamma_t} \psi_\theta$ ($d \times k_t$ matrix), $J_{\theta \alpha_t} = E \dfrac{\partial}{\partial \alpha_t} \psi_\theta$ ($d \times b_t$ matrix), and $J_{\theta \pi_t} = E \dfrac{\partial}{\partial \pi_t} \psi_\theta$ ($d \times c_t$ matrix). For simplicity, let $\tilde{\psi}_\theta = \psi_\theta - \sum_{t=1}^{K} J_{\theta \gamma_t} J_{\gamma_t}^{-1} \tilde{\psi}_{\gamma_t} - \sum_{t=1}^{K} J_{\theta \alpha_t} J_{\alpha_t}^{-1} \psi_{\alpha_t} - \sum_{t=1}^{K} J_{\theta \pi_t} J_{\pi_t}^{-1} \psi_{\pi_t}$. Since $E(\tilde{\psi}_\theta) = 0$, 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 matrices in Σ_θ (defined on the basis of expectations "E") using their corresponding empirical means averages based on the data (e.g., $\hat{J}_\theta = \mathbb{P}_n \dfrac{\partial}{\partial \theta} \psi_\theta$).