Biometrics 66, 131–139
March 2010

DOI: 10.1111/j.1541-0420.2009.01238.x

Structural Nested Mean Models for Assessing Time-Varying Effect Moderation

Daniel Almirall,[1,2,*] Thomas Ten Have,[3] and Susan A. Murphy[4]

[1]Center for Health Services Research in Primary Care, VA Medical Center, Durham, North Carolina 27705, U.S.A.
[2]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham,
North Carolina 27705, U.S.A.
[3]Department of Biostatistics and Epidemiology, University of Pennsylvania,
Philadelphia, Pennsylvania 19104, U.S.A.
[4]Department of Statistics and Institute for Social Research, University of Michigan,
Ann Arbor, Michigan 48109, U.S.A.
*email: daniel.almirall@duke.edu

Summary. This article considers the problem of assessing causal effect moderation in longitudinal settings in which treatment (or exposure) is time varying and so are the covariates said to moderate its effect. *Intermediate causal effects* that describe time-varying causal effects of treatment conditional on past covariate history are introduced and considered as part of Robins' structural nested mean model. Two estimators of the intermediate causal effects, and their standard errors, are presented and discussed: The first is a proposed two-stage regression estimator. The second is Robins' G-estimator. The results of a small simulation study that begins to shed light on the small versus large sample performance of the estimators, and on the bias–variance trade-off between the two estimators are presented. The methodology is illustrated using longitudinal data from a depression study.

Key words: Bias–variance trade-off; Causal inference; Effect modification; Estimating equations; G-estimation; Time-varying treatment; Time-varying covariates; Two-stage estimation.

1. Introduction

In this article, we are interested in assessing the causal effect of treatment as a function of variables that may lessen or increase this effect—that is, we are interested in *causal effect moderation*—in the time-varying setting. A distinctive feature of assessing effect moderation in the *time-varying setting* is that both treatment and the set of putative moderators vary over time. This feature of the data provides both an *opportunity* for improved empirical research and also provides a *methodological challenge*. An *opportunity* presents itself in the form of more varied and interesting questions that scientists may ask from time-varying data. For instance, consider our motivating example, the Prevention of Suicide in Primary Care Elderly—Collaborative Trial (PROSPECT) study (Bruce et al., 2004), in which some patients switch out of depression treatment with their mental health specialist. Using time-varying information about suicidal thoughts and depression, we can ask "How does switching out of treatment early versus later affect future depression severity scores as a function of history of suicidal ideation and levels of depression?"

A *methodological challenge* arises because moderators of the effect of future treatment may themselves be outcomes of earlier instances of treatment (Robins, 1987, 1989b, 1994, 1997); or, in the context of PROSPECT, suicidal ideation measured at the second visit (S_2) is a moderator of the effect of switching out of treatment after the second visit (A_2) on depression severity (Y), and switching off of treatment after the baseline visit (A_1) affects suicidal ideation at the second visit (S_2). In this setting, a naïve extension of the treatment–moderator interaction framework (Kraemer et al., 2002), in which, for instance, a regression model such as the following one is used, creates at least two problems for causal inference:

$$\begin{aligned} E(Y \mid S_1, A_1, S_2, A_2) = {} & \beta_0 + \eta_1 S_1 + \beta_1 A_1 + \eta_2 S_2 + \beta_2 A_2 \\ & + \beta_3 A_1 S_1 + \beta_4 A_1 S_2 + \beta_5 A_2 S_2 \\ & + \beta_6 A_1 A_2 + \eta_3 S_1 S_2. \end{aligned} \tag{1}$$

First, conditioning on S_2 cuts off any portion of the effect of A_1 on Y that occurs via S_2, including $A_1 S_1$ interaction effects. Secondly, there are likely common, unknown, causes of both S_2 and Y; thus, conditioning on S_2 (an outcome of treatment A_1) in equation (1) may introduce biases in the coefficients of the A_1 terms. The end result is that A_1 and its interactions (e.g., β_1 and β_3) may appear to be (un)correlated with Y solely because A_1 impacts S_2 and both S_2 and Y are affected by a common unknown cause. These problems can occur regardless of whether A_1 and/or A_2 are randomized (Robins, 1987, 1989b, 1994, 1997).

A framework for studying time-varying effect moderation that also addresses both of these challenges involves the notion of a *conditional intermediate causal effect* at each time point. These causal effects are a part of Robins' structural nested mean model (SNMM; Robins, 1994). They isolate the average effects of treatment at each time interval as a function of moderators available prior to that time interval.

This article contributes to the literature on modeling and estimating causal effects in the time-varying setting by (1) clarifying and illustrating the use of Robins' SNMM to assess time-varying effect moderation, (2) proposing a two-stage parametric regression estimator for the parameters of an SNMM, and (3) comparing the proposed parametric estimator to two versions of the semiparametric G-estimator (Robins, 1994) in terms of a bias–variance trade-off.

In Section 2, the causal effects of interest are defined in the context of Robins' SNMM. The two estimators of the intermediate causal effects are presented and discussed in Section 3. The results of a small simulation study that sheds light on the bias–variance trade-off between the two estimators is presented in Section 4. The methodology is illustrated in Section 5 using data from the PROSPECT study. A discussion of the methods is presented in Section 6.

2. Effect Moderation with Time-Varying Treatment and Time-Varying Moderators

To define the structural parameters and to state the structural assumptions necessary for valid causal inference we use the potential outcomes framework for causation (Rubin, 1974; Holland, 1986; Robins, 1987, 1989a, 1994, 1997, 1999). Suppose there are K time intervals under study. Treatment is denoted by a_t, at each time interval t, where $t = 1, \ldots, K$. Denote the treatment pattern over K intervals by $\bar{a}_K = (a_1, \ldots, a_K)$; where $a_t = 0$ represents standard, or baseline, treatment. For simplicity, we consider binary time-varying treatments a_t, though the ideas presented below apply with little modification to continuous or categorical a_t. Let $\mathcal{A}_K$ be the countable collection of all possible treatment vectors. For each fixed value of the treatment vector, $\bar{a}_K$, we conceptualize potential time-varying moderators $\{S_2(a_1), \ldots, S_K(\bar{a}_{K-1})\}$ and a potential final response $Y(\bar{a}_K)$. Thus, $S_t(\bar{a}_{t-1})$ is the value of the time-varying moderator at the beginning of the tth interval, had the subject followed the treatment pattern $\bar{a}_{t-1}$ through the end of the $t-1$ interval. Baseline moderators are denoted by the vector S_1. Let $\bar{S}_t(\bar{a}_{t-1}) = \{S_1, S_2(a_1), \ldots, S_t(\bar{a}_{t-1})\}$, the history of moderators up to the start of the tth time interval.

2.1 *The Conditional Intermediate Causal Effects*

For simplicity, let $K = 2$; thus, we have $\{S_1, a_1, S_2(a_1), a_2, Y(a_1, a_2)\}$. The response $Y(a_1, a_2)$ is taken to be continuous with unbounded support. We are only concerned with modeling the mean of the response $Y(\bar{a}_K)$ as a function of $\bar{a}_K$ and $S_K(\bar{a}_{K-1})$. Using potential outcomes we can express the marginal average causal effect of $\bar{a}_2$ on $Y(a_1, a_2)$ as $E\{Y(a_1, a_2) - Y(0, 0)\}$, where $a_t = 0$ is the baseline level of treatment. We can write this difference as a decomposition of conditional means: $E\{Y(a_1, a_2) - Y(0, 0)\} = E\{Y(a_1, a_2) - Y(a_1, 0)\} + E\{Y(a_1, 0) - Y(0, 0)\} = E[E\{Y(a_1, a_2) - Y(a_1, 0) \mid \bar{S}_2(a_1)\}] + E[E\{Y(a_1, 0) - Y(0, 0) \mid S_1\}]$, with the outer expectations over $\bar{S}_2(a_1)$ and S_1, respectively. The inner expectations on the right-hand side are *conditional intermediate causal effects* of treatment. Let $\mu_2(\bar{S}_2(a_1), \bar{a}_2)$ denote $E\{Y(a_1, a_2) - Y(a_1, 0) \mid \bar{S}_2(a_1)\}$, the effect of treatment (a_1, a_2) relative to the treatment $(a_1, 0)$ within levels of $\bar{S}_2(a_1)$; and let $\mu_1(S_1, a_1)$ denote $E\{Y(a_1, 0) - Y(0, 0) \mid S_1\}$, the effect of treatment $(a_1, 0)$ relative to $(0, 0)$ within levels of S_1. S_1 is defined as a *moderator* of the impact of $(a_1, 0)$ relative to $(0, 0)$ if μ_1 is nonconstant in S_1; similarly for μ_2. The effects μ_1 and μ_2 are *intermediate* causal effects because they isolate the causal effect of treatment at time 1 and time 2, respectively; note the constraints $\mu_2(\bar{S}_2(a_1), a_1, 0) = 0$ and $\mu_1(S_1, 0) = 0$. The "isolation" is achieved by setting future instances of treatment at their inactive levels—in our case, the zero level. Hence, μ_1 corresponds to a contrast of the potential outcomes in a_1 with a_2 set to its inactive level. 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.

2.2 *Robins' Structural Nested Mean Model*

We use the SNMM, developed by Robins (1994, 1997), to combine the μ_t's (also known as Robins' *blip* functions) additively in a model for the conditional mean of $Y(a_1, a_2)$ given $\bar{S}_2(a_1)$. The SNMM is expressed as a telescoping sum:

$$E\{Y(a_1, a_2) \mid \bar{S}_2(a_1)\} = \beta_0 + \epsilon_1(S_1) + \mu_1(S_1, a_1) + \epsilon_2(\bar{S}_2(a_1), a_1) + \mu_2(\bar{S}_2(a_2), \bar{a}_2), \quad (2)$$

where $\beta_0 = E\{Y(0, 0)\}$, the mean response to baseline treatment averaged over levels of $\bar{S}_2(a_1)$; $\epsilon_2(\bar{S}_2(a_1), a_1) = E\{Y(a_1, 0) \mid \bar{S}_2(a_1)\} - E\{Y(a_1, 0) \mid S_1\}$; and $\epsilon_1(S_1) = E\{Y(0, 0) \mid S_1\} - E\{Y(0, 0)\}$. In Web Appendix A, we define the SNMM for K time points. The SNMM depicts how the intermediate effect functions relate to the conditional mean of $Y(a_1, a_2)$ given the past; it places no restriction on the distribution of $Y(a_1, a_2)$. ϵ_1 and ϵ_2 are defined such that the right-hand side of equation (2) is indeed equal to $E\{Y(a_1, a_2) \mid \bar{S}_2(a_1)\}$. They satisfy the constraints $E\{\epsilon_2(\bar{S}_2(a_1), a_1) \mid S_1\} = 0$, and $E\{\epsilon_1(S_1)\} = 0$, and they are considered nuisance functions because they contain no information regarding the conditional intermediate causal effects of $\bar{a}_2$ on the mean of $Y(a_1, a_2)$. The nuisance functions carry causal and noncausal information about the relationship between the time-varying moderators and the response.

3. Estimation Strategies for the SNMM

3.1 *Observed Data and Assumptions Underlying Estimation*

Denote the *observed* treatment history by the random vector $\bar{A}_K := (A_1, A_2, \ldots, A_K)$; denote the *observed* time-varying covariate history by the random vector $\bar{S}_K := (S_1, S_2, \ldots, S_K)$; and denote the *observed* outcome by the random variable Y. The two estimation strategies described below rely on the assumptions of *consistency* and *sequential ignorability* (Robins, 1994, 1997) in order *to make causal inferences.*

The *consistency assumption* states that $Y = Y(\bar{A}_K) = \sum_{\bar{a}_K \in \mathcal{A}} I(\bar{a}_K = \bar{A}_K) Y(\bar{a}_K)$, where $I(\bar{a}_K = \bar{A}_K)$ denotes the indicator function that $\bar{a}_K$ is equal to $\bar{A}_K$. The consistency assumption is the link between objects defined as potential outcomes, and objects that are actually observed. Assuming consistency for the time-varying moderators $\{\bar{S}_K(\bar{a}_{K-1}) : \bar{a}_K \in A_K\}$ as well, then a similar relationship holds between the counterfactual objects in the SNMM and a corresponding set of observed data. The actual data observed (not to be confused with the complete set of potential outcomes defined above) for one individual in our study are $D = (S_1, A_1, S_2, A_2, \ldots, S_K, A_K, Y)$, where for each $t > 1$, S_t takes on some value in the set $\{S_t(\bar{a}_{t-1}) : \bar{a}_{t-1} \in A_K\}$, A_t takes on some value in the collection A_K, and Y takes on some value in the set $\{Y(\bar{a}_K) : \bar{a}_K \in A_K\}$.

Another key assumption used to identify the causal parameters of the SNMM using observed data is the *sequential ignorability assumption*: For each $t = 1, 2, \ldots, K$, A_t is independent of $\{Y(\bar{a}_K) : \bar{a}_K \in \mathcal{A}_K\}$ given $(S_1, A_1, S_2, A_2, \ldots, S_t)$. This assumption is sometimes known as the *no unmeasured confounders assumption* (Robins, 1997). Intuitively, the assumption implies that aside from $(\bar{S}_t, \bar{A}_{t-1})$, the history of observed treatment and moderator variables up to the beginning of the tth time interval, no other variables known or unknown, measured or unmeasured, directly affect both A_t and $\{Y(\bar{a}_K) : \bar{a}_K \in \mathcal{A}_K\}$.

The causal meaning of models for $\mu_t(\bar{S}_t, \bar{A}_t) = E(Y \mid \bar{S}_t, \bar{A}_t) - E(Y \mid \bar{S}_t, \bar{A}_{t-1}, A_t = 0)$ relies on the above assumptions (see Web Appendix B), whereas their estimation requires other modeling (or statistical) assumptions, such as the choice of parametric models. One possible parameterization of the intermediate causal effects is linear in the parameters. For example, in our presentation of both estimators below, we use

$$\mu_t(\bar{S}_t, \bar{A}_t; \beta_t) = A_t(H_t\beta_t), \tag{3}$$

where A_t is the univariate treatment variable at time *t*, β_t represents a q_t-dimensional column vector of unknown parameters at time *t*; and H_t, a q_t-dimensional row vector, is a known function of $(\bar{S}_t, \bar{A}_{t-1})$. Using this general form ensures that the following constraint is always satisfied: $\mu_t(\bar{S}_t, \bar{A}_{t-1}, 0; \beta_t) = 0$. Typically, the first element in the row vector H_t is one.

3.2 *Parametric Two-Stage Estimator: $\hat{\beta}$*

We propose a parametric two-stage estimator that employs the following *general approach* for estimating the parameters of an SNMM: In the first stage, for every *t*, we model the conditional distribution of S_t given $(\bar{S}_{t-1}, \bar{A}_{t-1})$, denoted by $f_t(S_t \mid \bar{S}_{t-1}, \bar{A}_{t-1})$, based on unknown parameters γ_t. We then pose a model $\epsilon_t(\bar{S}_t, \bar{A}_{t-1}; \eta_t, \gamma_t) = g_t(\bar{S}_{t-1}, \bar{A}_{t-1}; \eta_t)\, \delta_t(\bar{S}_t, \bar{A}_{t-1}; \gamma_t)$ for the nuisance functions $\epsilon_t(\bar{S}_t, A_{t-1}) = E(Y \mid \bar{S}_{t-1}, \bar{A}_{t-1}, S_t, A_t = 0) - E(Y \mid \bar{S}_{t-1}, \bar{A}_{t-1}, A_t = 0)$. Under this general model, g_t is any function of $(\bar{S}_{t-1}, \bar{A}_{t-1})$ based on unknown parameters η_t; and δ_t is any function of $(\bar{S}_t, \bar{A}_{t-1})$ based on one or more features of the distribution f_t (and parameters γ_t), such that $E(\delta_t \mid \bar{S}_{t-1}, \bar{A}_{t-1}) = 0$. The constraint ensures that $E(\epsilon_t \mid \bar{S}_{t-1}, \bar{A}_{t-1}) = 0$, as required. In the second stage, these models for the nuisance functions are put together with models for the intermediate causal effects in an SNMM for the conditional mean of Y given $(\bar{S}_K, \bar{A}_K)$. Estimates for β are then based on solutions to the following estimating equations:

$$\mathbb{P}_n\left[\left\{Y - \beta_0 - \sum_{t=1}^{K} A_t(H_t\beta_t) - \sum_{t=1}^{K} \epsilon_t(\bar{S}_t, \bar{A}_{t-1}; \eta_t, \gamma_t)\right\} \times \begin{pmatrix} A_1 H_1^T \\ \cdots \\ A_K H_K^T \end{pmatrix}\right] = 0 \tag{4}$$

where for any function $V()$ of the observed data D, $\mathbb{P}_n V(D)$ denotes $1/n \sum_{i=1}^{n} V(D_i)$. In the following, we present a *particular linear implementation* of this general approach.

3.2.1 *A linear regression implementation of the two-stage estimator.* For simplicity, assume that S_t at each time point is univariate (i.e., one time-varying moderator per time t is used); an extension of the method to multivariate S_t is presented in Web Appendix C. The proposed linear implementation sets $\delta_t = S_t - m_t(\bar{S}_{t-1}, \bar{A}_{t-1}; \gamma_t)$ and uses $g_t = G_t\eta_t$, where G_t is a row vector summary of the past $(\bar{S}_{t-1}, \bar{A}_{t-1})$, η_t is an unknown w_t-dimensional vector of parameters, and m_t is a model for the conditional mean of S_t given the past based on an unknown l_t-dimensional vector of parameters γ_t. Let F_t be a row vector of the data $(\bar{S}_{t-1}, \bar{A}_{t-1})$. We employ generalized linear models (GLMs; McCullagh and Nelder, 1989) for the m_t: Thus, when S_t is continuous, we use $m_t(\bar{S}_{t-1}, \bar{A}_{t-1}; \gamma_t) = F_t\gamma_t$; when S_t is binary, we use $m_t(\bar{S}_{t-1}, \bar{A}_{t-1}; \gamma_t) = Pr(S_t = 1 \mid \bar{S}_{t-1}, \bar{A}_{t-1}) = \text{expit}(F_t\gamma_t)$. (It is also possible to model the m_t simultaneously, in a longitudinal model.) A simple model for ϵ_t will have $G_t = (1)$, so that $\epsilon_t = \eta_{t0}\delta_t(\bar{S}_t, \bar{A}_{t-1}; \gamma_t)$, for example. Note that under this implementation, we can multiply every element of G_t by the residual δ_t, denoted $G_t^*(\gamma_t)$, and re-write the parametric model for the nuisance functions as $\epsilon_t(\gamma_t, \eta_t) = G_t^*(\gamma_t)\eta_t$. If γ_t were known, this would imply a linear (in the β's and η's) parametric model for the SNMM. For example, for $K = 2$, $E(Y \mid \bar{S}_2, \bar{A}_2) = \beta_0 + A_1H_1\beta_1 + A_2H_2\beta_2 + G_1^*\eta_1 + G_2^*\eta_2$. This idea forms the basis for the linear implementation of the two-stage approach, given here for general K time points:

(1) Stage 1 Regression. GLM analyses are used in the first stage to obtain $\hat{\gamma}_t$ based on regressions of S_t on (S_{t-1}, A_{t-1}). These are carried out for each time point $t = 1, 2, \ldots, K$.

(2) Use the predicted means $\hat{m}_t(\hat{\gamma}_t)$ to construct the predicted residuals $\hat{\delta}_t = S_t - \hat{m}_t$.

(3) Combine the model vectors for the conditional intermediate effects (and a column for the intercept) and denote this quantity by *X*; that is, $X = (1, A_1H_1, \ldots, A_KH_K)$. Note that $X\beta = \beta_0 + \sum_{t=1}^{K} A_tH_t\beta_t$ represents the functional of interest of the SNMM.

(4) Multiply each element in G_t by the predicted residual $\hat{\delta}_t$ and denote this quantity by $\hat{G}_t^*$; that is, $\hat{G}_t^* = \hat{\delta}_tG_t$. Note that if $\eta = (\eta_1^T, \ldots, \eta_K^T)^T$ were known, then $G^*\eta = \sum_{t=1}^{K} \hat{G}_t^*\eta_t$ would represent an estimate of the sum of the nuisance functionals of the SNMM.

(5) Augment the row vector X to include the $\hat{G}_t^*$'s; that is, $X_{aug} = (X, \hat{G}_1^*, \ldots, \hat{G}_K^*)$. Define the $(1 + \sum_{t=1}^{K} q_t + \sum_{t=1}^{K} w_t)$-dimensional column vector of parameters $\theta = (\beta^T, \eta^T)^T$.

(6) Stage 2 Regression. Use standard linear regression of Y on X_{aug} to obtain $\hat{\theta} = \hat{\theta}(\hat{\gamma})$.

3.3 *Robins' Semiparametric Efficient G-Estimator: $\tilde{\beta}$*

The following estimator, derived in Robins (1994), does not rely solely on correct models for the nuisance functions to achieve consistency for β. Instead, its consistency relies on either correct models for the nuisance functions *or* correct models for the observed treatment mechanism (conditional on the time-varying moderators). In $K = 2$ (see Web Appendix D for the general form), the estimate is based on these estimating functions:

$$\begin{aligned} \mathbb{P}_n \Bigg[& \sigma_2^{-1}(\bar{S}_2, A_1)(Y - A_2(H_2\beta_2) - b_2(\bar{S}_2, A_1; \xi_2)) \\ & \times \{A_2 - p_2(\bar{S}_2, A_1; \alpha_2)\} \left\{ \begin{matrix} 0_{q_1} \\ H_2^T \end{matrix} \right\} \\ & + \sigma_1^{-1}(S_1)\{Y - A_2(H_2\beta_2) - A_1(H_1\beta_1) - b_1(S_1; \xi_2)\} \\ & \times \{A_1 - p_1(S_1; \alpha_1)\} \left\{ \begin{matrix} H_1^T \\ \Delta(S_1; \kappa) \end{matrix} \right\} \Bigg], \end{aligned} \tag{5}$$

where $p_t(\bar{S}_t, \bar{A}_{t-1}; \alpha_t)$ is a model for $Pr(A_t = 1 \,|\, \bar{S}_t, \bar{A}_{t-1})$; $b_2(\bar{S}_2, A_1; \xi_2)$ is a model for $E\{Y - A_2(H_2\beta_2) \,|\, \bar{S}_2, A_1\}$, and $b_1(S_1; \xi_1)$ is a model for $E\{Y - A_2(H_2\,\beta_2) - A_1(H_1\beta_1) \,|\, S_1\}$; $\Delta(S_1;\ \kappa)$ is a model for $E\{A_2 H_2^T \,|\, S_1,\ A_1 = 1\} - E\{A_2\ H_2^T \,|\, S_1,\ A_1 = 0\}$; and 0_{q_1} is a q_1-dimensional row vector of zeros. The set of equations (5) is $(q_1 + q_2)$-dimensional because H_t is q_t-dimensional for $t = 1,\ 2$. We denote this system of equations by $\mathbb{P}_n \psi_\beta(D; \alpha, \xi, \kappa)$, where $\alpha = (\alpha_1^T,\ \alpha_2^T)^T$ (of dimension $r_1 + r_2$), $\xi = (\xi_1^T,\ \xi_2^T)^T$, and κ are all unknown parameters. The conditional variances $\sigma_1^2(S_1)$ and $\sigma_2^2(\bar{S}_2, A_1)$ are defined as $\sigma_1^2(S_1) = Var\{Y - A_2(H_2\beta_2) - A_1(H_1\beta_1) \,|\, S_1\} = Var\{Y - A_2(H_2\beta_2) - A_1(H_1\beta_1) \,|\, S_1,\ A_1\}$ and $\sigma_2^2(\bar{S}_2, A_1) = Var\{Y - A_2(H_2\beta_2) \,|\, \bar{S}_2, A_1\} = Var\{Y - A_2(H_2\beta_2) \,|\, \bar{S}_2, \bar{A}_2\}$ where the second equality in each follows by assumption (without this partially homogenous variance assumption, the estimating equations are intractable). In our implementation, we further assume that these variances are constant in $(\bar{S}_2, A_1)$ and S_1, respectively. To use $\mathbb{P}_n \psi_\beta(D; \alpha, \xi, \kappa)$ for estimation, we substitute estimates of the parameters α, ξ, and κ in $p_t(\alpha_t)$, $b_t(\xi_t)$, and $\Delta(\kappa)$—denoted $\hat{p}_t(\hat{\alpha}_t), \hat{b}_t(\hat{\xi}_t)$, and $\hat{\Delta}(\hat{\kappa})$—and solve for β in the estimating equations $0 = \mathbb{P}_n \psi_\beta(D; \hat{\alpha}, \hat{\xi}, \hat{\kappa})$. The resulting estimator $\tilde{\beta} := \tilde{\beta}(\hat{\alpha}, \hat{\xi}, \hat{\kappa})$ is known as *Robins' locally efficient semiparametric G-estimator* for β.

3.3.1 *Implementing Robins' G-estimator.* We consider two implementations of Robins' G-estimator: In both implementations, we estimate $\hat{\alpha}$ using logistic regression models at each time point t based on Z_t, a row vector of the data $(\bar{S}_t, \bar{A}_{t-1})$. The predicted probabilities from the logistic regression are used to get $\hat{p}_t$. In both implementations, we obtain $\hat{\Delta}$ by first using ordinary multivariate regression models for $\lambda(S_1,\ A_1;\ \kappa) = E(A_2 H_2^T \,|\, S_1,\ A_1)$ to get $\hat{\kappa}$, and then predicting Δ using $\hat{\Delta}(S_1; \hat{\kappa}) = \hat{\lambda}(S_1, 1; \hat{\kappa}) - \hat{\lambda}(S_1, 0; \hat{\kappa})$. The two implementations differ in the way we obtain estimates for b_t and how we find the solution to $\mathbb{P}_n \psi_\beta = 0$.

Implementation A. For a fixed t, note that b_t is a model for $E\{Y - \sum_{j=t}^K A_j(H_j\beta_j) \,|\, \bar{S}_t, \bar{A}_{t-1}\} = \beta_0 + \sum_{j=1}^{t-1} A_j(H_j\beta_j) + \sum_{j=1}^t \epsilon_t(\bar{S}_t, \bar{A}_{t-1})$, the sum of the intercept, the nuisance functions through time t, and the intermediate causal effects through time $t - 1$. Hence, the two-stage estimator presented above can be used to obtain the "guesses" $\hat{b}_t(\hat{\xi}_t)$ needed to solve the equations. We do this under G-estimator implementation A, by using the relevant portions of the (two-stage) estimated conditional mean at each time t to create the estimates for $\hat{b}_t(\hat{\xi}_t)$. Thus, at time 1, for instance, we use $\hat{b}_1(S_1; \hat{\xi}_1) = \hat{\beta}_0 + \hat{\epsilon}_1(S_1; \hat{\eta}_1, \hat{\gamma}_1)$, where $\xi_1 := (\beta_0,\ \eta_1^T,\ \gamma_1^T)^T$ are estimated using the two-stage estimator. The numerical search for the solution to $\mathbb{P}_n \psi_\beta(D; \hat{\alpha}, \hat{\xi}, \hat{\kappa}) = 0$ is itself an iterative process that requires starting values. We use the estimates obtained from the two-stage estimator, $\hat{\beta}$, as the starting values for this search. A `FORTRAN` subroutine was used to find the zeros of $\mathbb{P}_n \psi_\beta$. The estimates for β under G-estimator implementation A are labeled $\tilde{\beta}_A$.

Implementation B. For the second implementation, we employ two successive evaluations of a one-step Newton–Rhapson (NR) estimator to obtain the solutions to $\mathbb{P}_n \psi_\beta(D; \hat{\alpha}, \xi, \hat{\kappa}) = 0$, as follows: First set starting values $\beta^{(0)} = 0$ and starting guesses $b_t = 0$, and find $\tilde{\beta}^{(1)}$ using one iteration of the NR method. Second, obtain predicted values $\hat{b}_t(\hat{\xi}_t)$ based on the results of linear regressions of $Y - \sum_{j=t}^K A_j(H_j\tilde{\beta}_j^{(1)})$ on $(\bar{S}_t, \bar{A}_{t-1})$. Third, using starting values $\tilde{\beta}^{(1)}$ and starting guesses $\hat{b}_t(\hat{\xi}_t)$, find $\tilde{\beta}^{(2)}$ using one iteration of the NR method. We label the resulting estimates $\tilde{\beta}_B$ for G-estimator implementation B.

3.4 *Estimated Standard Errors (SEs) for $\hat{\beta}$ and $\tilde{\beta}$*

Estimated asymptotic SEs for $\hat{\beta}$ and $\tilde{\beta}$ are computed using the delta method, based on one-step Taylor series expansions (see Web Appendix D). $\hat{SE}(\hat{\beta})$ takes into account the variability in the estimation of γ. $\hat{SE}(\tilde{\beta})$ takes into account the variability in the estimation of α and ξ.

3.5 *Comparing the Properties of the Two-Stage Estimator and the G-Estimator*

The G-estimating equations $\mathbb{P}_n \psi_\beta$ provide unbiased estimating functions for β given correct models for the intermediate causal effects and the p_t's, regardless of our choice of models for the b_t's and Δ (or the outcome variances, the σ_t^2's). Indeed, even if $\Delta = 0$ and $b_t = 0$ for all t, we still have $E(\psi_\beta) = 0$. Conversely, given correct models for the intermediate causal effects and the b_t's, unbiasedness is still achieved with the G-estimator regardless of our choice of models for the p_t's and Δ. This is known as the *double-robustness property* of the G-estimator (Robins and Rotnitzky, 2001). Now, provided true models for both b_t and p_t (for all t; and true model for Δ), the resulting estimates are also asymptotically efficient. By efficient, we mean that the asymptotic variance of the resulting G-estimates of β achieve the semiparametric efficiency variance bound (Bickel et al., 1993) for the class of models in which μ_t are parameterized linearly and the remaining aspects of the model, including the ϵ_ts, are left unspecified.

The two-stage estimator relies on correct models for *both* the intermediate causal effects and the nuisance functions to provide unbiased estimates for β. At the correct model fit, we conjecture that the two-stage estimator enjoys better efficiency than the G-estimator. This gain in precision, however, may be offset by a lack of robustness to misspecifications in the ϵ_t's. Exactly how to balance the trade-off between bias and variance is an open question. The simulation experiments in the next section shed light on this question.

4. Simulation Experiments

All simulations are based on $N = 1000$ simulated data sets and $K = 3$. We generated continuous time-varying covariates $\{S_1,\ S_2,\ S_3\}$ and continuous outcome Y such that their implied marginal distributions and bivariate correlations are similar to those found in the PROSPECT data (see Section 5 below), where S_t is suicidal ideation at time t, and Y is end-of-study depression scores. Specifically, $[S_1] \sim N(m_1 = 0.5, \text{sd} = 0.82), [S_2|S_1, A_1] \sim N(m_2 = 0.5 + 0.10S_1 - 0.5A_1 + 0.35S_1A_1, \text{sd} = 0.65)$, and $[S_3|\bar{S}_2, \bar{A}_2] \sim N(m_3 = 0.5 + 0.17S_2 + 0.1S_1 - 0.5A_2 + 0.5S_2A_2, \text{sd} = 0.65)$, where binary treatment $A_t \in (0, 1)$ at each time point is generated as a binomial random variable with $Pr(A_t = 1 \,|\, \bar{S}_t, \bar{A}_{t-1}) = p_t = \text{expit}(0.5 - 1.5S_t)$. The nuisance functions were chosen as $\epsilon_1 = 0.1 \ (S_1 - m_1)$, $\epsilon_2 = \{0.2 + 0.18S_1 + 0.4A_1 + 0.35A_1S_1 + \sin(4.5 \ S_1)\} \ (S_2 - m_2)$, and $\epsilon_3 = \{0.3 + 0.18S_2 + 0.4A_2 + 0.35A_2S_2 + \sin(2.5 \ S_2)\} \ (S_3 - m_3)$. The intermediate causal effect functions in the SNMM were set to: $\mu_t = A_t(H_t\beta_t) = A_t(1, S_t) \ (\beta_{t,0}, \beta_{t,1})^T = A_t(\beta_{t,0} + \beta_{t,1}S_t)$ for $t = 1, 2, 3$. The true value for all six causal parameters was set to $\beta_{t,j} = 0.45$, where $j = 0, 1$. The outcome Y was generated as a normal random variable with a conditional mean structure according to the above SNMM, and intercept and residual standard deviation (SD) set to 1.0.

Experiments were carried out to compare the performance of the two-stage regression estimator and the two G-estimator implementations (A and B; see Section 3). For both G-estimator implementations, true logistic regression models were always fit to obtain the $\hat{p}_t$ predictions; and multivariate linear regression models that included all main effects and all second-order interaction terms were always used for $E(A_2H_2^T \,|\, S_1, \ A_1)$, $E(A_3H_3^T \,|\, S_1, \ A_1)$, and $E(A_3H_3^T \,|\, \bar{S}_2, \bar{A}_2)$ to obtain $\hat{\Delta}$. True models were always fit for the b_t under G-estimator B. True models were always specified for the intermediate causal effects, the μ_t's, for all three estimators. Models for the ϵ_t under the two-stage estimator varied, as described below.

First, we compared the small versus large sample properties of the three estimators when the *true* two-stage estimator is fit to the data. Sample sizes were varied and performance measures for point estimates and SEs were compared across the estimators. The $n = 300$ sample size is approximately the size of the data set used in the analysis in Section 5; $n = 50$ and $n = 1000$ were chosen to study relatively smaller and larger data sets.

In a second set of experiments, we sought to shed light on the bias–variance trade-off between the two-stage estimator and the G-estimator. In this set of experiments, the nuisance functions in the two-stage estimator were misspecified and the relative performance of the estimators (in terms of mean squared error [MSE]) was assessed. Only data sets of size $n = 300$ were considered. Misspecification of the nuisance functions is measured using the scaled root-mean squared difference $= \text{SRMSD}(\nu) = \sqrt{\sum_t^K E\{\epsilon_t^{\text{TRUE}} - \epsilon_t(\nu)\}^2/\text{Var}(Y)}$, where for a fixed value of ν, $\epsilon_t(\nu)$ denotes the misspecified nuisance function at time t. SRMSD has the interpretation of an *effect size*, so that SRMSD values of 0.2 and 0.5, for example, correspond to small and moderate levels of misspecification, respectively (see Cohen, 1988). We varied values of SRMSD using ν, by replacing S_t's in the ϵ_t's with S_t times U, where U is a draw from the normal distribution $N(1, sd = \nu)$. When $\nu = 0$, the true two-stage estimator is fit to the data.

4.1 *Simulation Results and Discussion*

Table 1 shows the results of the first experiment. We report results only for β_{t1}, $t = 1, 2, 3$; results for β_{t0} are similar unless otherwise noted. As expected according to large sample theory, all three estimators are unbiased when $n = 1000$; and empirical SDs and mean SEs (MEAN SE) show good agreement for all three estimators when $n = 1000$. All 95% CI coverage probabilities at $n = 1000$ show coverages between the expected 93.6% and 96.4% range for $n = 1000$ replicates. Performance in terms of mean bias is only slightly worse at $n = 300$ relative to $n = 1000$ for the three estimators. As expected, the variance (both SD and AVG SE) increases significantly with the smaller sample size. The 95% CIs show under-coverage at smaller sample sizes, especially at $n = 50$.

Table 1

Small versus large sample performance of the two-stage estimator, G-estimator implementation A, and G-estimator implementation B. Results are shown only for β_{t1}, the time-varying interaction terms at each time point t. True parameter values are $\beta_{t1} = 0.45$ for t = 1, 2, 3.

		True Two-stage estimator					G-estimator implementation A					G-estimator implementation B				
		AVG		AVG	95%		AVG		AVG	95%	REL	AVG		AVG	95%	REL
n	β_{2j}	EST	SD	SE	COV	MSE	EST	SD	SE	COV	MSE	EST	SD	SE	COV	MSE
50	β_{11}	0.453	0.580	0.532	91.0	0.337	0.459	0.807	0.735	88.5	1.9	0.463	0.846	0.804	89.8	2.1
	β_{21}	0.412	0.739	0.637	90.0	0.547	0.379	0.859	0.669	83.2	1.4	0.402	0.863	0.759	89.6	1.4
	β_{31}	0.430	0.634	0.493	85.8	0.402	0.302	0.829	0.552	80.9	1.8	0.483	0.688	0.491	78.7	1.2
300	β_{11}	0.452	0.197	0.186	93.5	0.039	0.468	0.310	0.297	93.2	2.5	0.470	0.310	0.295	93.3	2.5
	β_{21}	0.456	0.229	0.222	93.8	0.053	0.454	0.279	0.273	94.6	1.5	0.457	0.285	0.278	94.4	1.6
	β_{31}	0.439	0.184	0.182	93.4	0.034	0.444	0.207	0.208	92.8	1.3	0.426	0.283	0.269	93.3	2.4
1000	β_{11}	0.452	0.102	0.100	94.6	0.010	0.447	0.171	0.165	93.9	2.8	0.447	0.170	0.164	94.0	2.8
	β_{21}	0.440	0.117	0.120	95.7	0.014	0.439	0.150	0.147	94.1	1.6	0.438	0.155	0.150	94.1	1.8
	β_{31}	0.454	0.099	0.099	94.8	0.010	0.456	0.114	0.113	94.0	1.4	0.455	0.154	0.146	93.7	2.4

REL MSE denotes MSE relative to the true two-stage estimator.

Table 2

Results of a simulation experiment to understand the bias–variance trade-off between the two-stage estimator and the two G-estimator implementations. Data sets of size $n = 300$ used for these simulations. Results are shown only for β_{t1}, $t = 1, 2, 3$, the time-varying interaction terms. $\nu > 0$ corresponds to misspecified two-stage regression estimators.

		Two-stage regression estimator						G-estimator implementation A				
ν	β_{tj}	AVG EST	SD	AVG SE	95% COV	REL MSE$^{\#}$	REL MSE†	AVG EST	SD	AVG SE	95% COV	REL MSE†
0	β_{11}	0.452	0.197	0.186	93.5	0.4	0.4	0.468	0.310	0.297	93.2	1.0
	β_{21}	0.456	0.229	0.222	93.8	0.7	0.7	0.454	0.273	0.266	94.6	1.0
	β_{31}	0.439	0.184	0.182	93.4	0.8	0.4	0.444	0.208	0.208	92.8	0.5
1	β_{11}	0.345	0.219	0.208	89.7	0.6	0.6	0.468	0.314	0.292	98.3	1.0
	β_{21}	0.348	0.275	0.267	92.5	0.9	1.0	0.454	0.311	0.308	98.6	1.2
	β_{31}	0.350	0.233	0.219	90.5	0.9	0.9	0.444	0.263	0.286	99.0	0.9
3	β_{11}	0.247	0.230	0.215	81.9	0.9	1.0	0.468	0.316	0.292	91.8	1.0
	β_{21}	0.105	0.300	0.279	73.4	2.1	2.6	0.459	0.317	0.317	94.7	1.2
	β_{31}	0.157	0.235	0.228	72.7	2.0	1.7	0.431	0.268	0.316	97.0	0.9
5	β_{11}	0.230	0.230	0.217	81.7	1.0	1.1	0.468	0.316	0.292	92.0	1.0
	β_{21}	0.049	0.297	0.282	68.6	2.4	3.0	0.456	0.319	0.318	94.0	1.3
	β_{31}	0.124	0.231	0.230	67.7	2.3	2.0	0.430	0.265	0.318	96.5	0.9

$\nu = 0, 1, 3, 5$ corresponded to SRMSD $= 0.0, 0.47, 0.53, 0.55$, respectively.

$^{\#}$denotes MSE relative to the G-estimator implementation A.

†denotes MSE relative to the G-estimator implementation B.

For the three estimators, the coverage gets worse for the parameters at later time points.

REL MSE denotes relative MSE of the G-estimator relative to the two-stage estimator. As conjectured, the two-stage estimator (evaluated at the true model) outperforms the G-estimator in terms of MSE across all the scenarios.

Table 2 shows the results of the second experiment. We report results only for β_{t1}. Values of ν were varied from 0 to 5 corresponding to SRMSD values between 0.0 and 0.55—that is, from no misspecification to just beyond "moderate" amounts of misspecification in the two-stage estimator. We report REL MSE's of the two-stage estimator relative to the two G-estimators, and REL MSE of G-estimator A relative to G-estimator B. Because we always fit the correct models for b_t (and p_t) for G-estimator implementation B, a panel is not shown for G-estimator implementation B. As expected, bias increases for the two-stage estimator, whereas the G-estimator remains unbiased, for higher values of ν. As expected, as well, REL MSE of the two-stage estimator relative to the G-estimators increases for higher values of ν. The results show this is largely due to increased bias in the two-stage estimator. The results also demonstrate that the G-estimators begin to dominate the two-stage estimator (in terms of MSE) at $\nu = 3$, corresponding to roughly a "moderate" amount of misspecification (SRMSD ≈ 0.5), although there is some minimal variation by parameter in the trajectories. For both β_{31} (and β_{30}, not shown), G-estimator A dominates G-estimator B when $\nu = 0$. However, G-estimator B is equivalent to or slightly better than G-estimator A under misspecified models for the b_t in G-estimator A. This is due to increases in variance in G-estimator A for misspecified b_t models, as expected according to large sample theory and as demonstrated in the simulations.

5. An Illustration Using the PROSPECT Data

The PROSPECT study evaluated the effect of a primary care intervention, as compared to usual care, on suicidal ideation and depression in older patients (Bruce et al., 2004). Intervention patients had access to a mental health specialist who offered education, care management, and support. Not all patients met with their health specialist throughout the entire course of the study; and those that stopped meeting with their health specialist at some time point, never met with them again thereafter (i.e., a monotonic treatment pattern). The continuous scale for suicidal ideation (SSI) and continuous Hamilton depression severity scale (HAMD) were collected at baseline and at clinic visits at 4 months, 8 months, and 12 months post-baseline. Restricting our sample to patients in the intervention arm ($n = 277$), we illustrate the SNMM methodology to assess the impact of switching off treatment (defined as meeting with the health specialist) early versus later on end-of-study (12-month) depression severity, as a function of SSI and HAMD between baseline and 8 months. From a practical point of view, understanding these effects can serve as a useful guide to clinical decision making by suggesting whether or not a patient should continue to meet with their health specialist depending on what is known about their history of depression and suicidal ideation.

Let $S_t = (\mathrm{SSI}_t, \mathrm{HAMD}_t)$, where $t = 1, 2, 3$ denotes clinic visits at baseline, 4 months, and 8 months, respectively. Let A_t denote the binary (1/0) indicator measuring whether a patient met with his/her assigned health specialist between clinic visits (i.e., $A_1 = 1$ means the patient met with his/her health specialist between baseline and 4 months; A_2 and A_3 defined similarly for the 4-to-8 and 8-to-12 month intervals, respectively). Let Y denote 12-month HAMD; that is, depression severity at the end of study.

Table 3

An illustrative analysis of the effect of switching off treatment early versus later on end-of-study depression scores given time-varying suicidal ideation and depression severity

	Two-stage regression estimator			G-estimator implementation A			G-estimator implementation B		
Effect	$\hat{\beta}$	SE	P-val	$\tilde{\beta}_A$	SE	P-val	$\tilde{\beta}_B$	SE	P-val
$\beta_{10}A_1$	−0.201	0.373	0.59	−0.211	0.369	0.57	−0.106	0.349	0.76
$\beta_{11}A_1\text{SSI}_1$	0.106	0.149	0.48	0.091	0.338	0.79	0.142	0.288	0.62
$\beta_{12}A_1\text{HAMD}_1$	−0.241	0.347	0.49	−0.105	0.334	0.75	−0.159	0.324	0.62
$\beta_{20}A_2$	0.396	0.312	0.20	0.519	0.327	0.11	0.418	0.337	0.21
$\beta_{21}A_2\text{SSI}_2$	0.221	0.272	0.42	−0.028	0.366	0.94	−0.200	0.352	0.57
$\beta_{22}A_2\text{HAMD}_2$	0.109	0.158	0.49	0.106	0.237	0.65	0.274	0.258	0.29
$\beta_{30}A_3$	−0.268	0.258	0.30	−0.416	0.300	0.17	−0.403	0.311	0.20
$\beta_{31}A_3\text{SSI}_3$	−0.119	0.261	0.65	−0.604	0.499	0.23	−0.620	0.519	0.23
$\beta_{32}A_3\text{HAMD}_3$	−0.015	0.187	0.94	0.026	0.228	0.91	0.039	0.240	0.87

Causal models. Parsimonious linear models were chosen for the intermediate causal effects, whereby the effect of switching off treatment at time t is a function of suicidal ideation and depression severity at time t only: $\mu_t = A_t\,(1, \text{SSI}_t, \text{HAMD}_t)\,(\beta_{t0}, \beta_{t1}, \beta_{t2})^T = \beta_{t0}A_t + \beta_{t1}A_t\text{SSI}_t + \beta_{t2}A_t\text{HAMD}_t$. The variables SSI_t and HAMD_t in μ_t were centered at their grand mean so that β_{t0} has the interpretation as the average effect of switching off treatment at time t for patients with average suicidal ideation and depression severity. Due to monotonicity (i.e., $\mathcal{A}_3 = \{(0,0,0), (1,0,0), (1,1,0), (1,1,1)\}$), the intermediate causal effects cannot vary according to previous levels of treatment.

Nuisance models. Simple linear regressions were used for the m_t in the two-stage estimator, for the Δ_t in both G-estimator implementations, and for the b_t in G-estimator B. Logistic regressions were used for the p_t in both G-estimator implementations. In each of these working models, and in the G_t in the two-stage estimator, we included all variables in the past as main effects together with all second-order interaction terms.

Variable selection. To arrive at parsimonious working models for m_t, Δ_t, b_t, and p_t we employed a hierarchical forward-and-backward stepwise AIC variable selection procedure, as implemented in R (Venables and Ripley, 2002). The same variable selection method was used on the final two-stage regression fit to find the best fitting models for G_t. Variable selection results and working model fits are not shown here for reasons of space.

The results of the data analysis are shown in Table 3 for the two-stage estimator and the two implementations of Robins' G-estimator. The three estimators show good agreement in terms of direction and magnitude of the effects, and in the magnitude of estimated SEs. The results show no effect moderation by either time-varying suicidal ideation or previous levels of depression severity, nor evidence of an effect of treatment for patients with average levels of SSI and HAMD. Indeed, the results show no significant impact whatever on end-of-study depression severity of meeting with a mental health specialist early or later during the course of the study.

6. Discussion

This article presents and discusses the use of intermediate conditional causal effects to study time-varying causal effect moderation using Robins' SNMM (Robins, 1994). A parametric two-stage regression estimator was proposed and compared with the existing semiparametric G-estimator.

The SNMM and Robins' G-estimator have also been used previously to study the effects of randomization to an intervention in the presence of noncompliance (Goetghebeur and Lapp, 1997), including when the outcome is binary (Vansteelandt and Goetghebeur, 2003; Robins and Rotnitzky, 2005); and the G-estimator has been used recently for studying causal effect mediation (Joffe, Small, and Hsu, 2007; Ten Have et al., 2007). Petersen et al. (2007) have also proposed a method for assessing time-varying effect moderation, called historically adjusted marginal structural models (HA-MSMs). With HA-MSMs, MSMs (Robins, 1999) have been generalized to allow conditioning on time-varying covariates; this is accomplished by positing different MSMs, one per time point, and estimating them simultaneously. HA-MSM's differ from SNMMs in one important respect; namely, SNMMs are fully structural models for the conditional mean of Y given $(\bar{S}_K, \bar{A}_K)$, whereas with HA-MSMs there is no requirement, for instance, that the model posed for the causal effect of a_1 in the MSM at $t = 1$ be equivalent to the model for the causal effect of a_1 that is *implied* by the last MSM at $t = K$. Future work that compares HA-MSMs and SNMMs will be important.

The two-stage estimator requires more knowledge about portions of the conditional mean of Y given $(\bar{S}_K, \bar{A}_K)$ than does Robins' G-estimator. If this additional knowledge (concerning the nuisance functions ϵ_t) is incorrect, it is possible that $\hat{\beta}$ is biased for the true β. On the other hand, scientists may tolerate bias in $\hat{\beta}$ if its variance is smaller than an unbiased $\tilde{\beta}$. The simulation studies presented above begin to shed light on this bias–variance trade-off. The simulation experiments suggest that it may be useful to consider parametric estimators such as the two-stage estimator over the G-estimator up to *moderate* misspecifications in models for the nuisance functions using the parametric estimator. This interpretation relies on fitting true models for the μ_t; in practice, it is possible the scientist will misspecify the μ_t as well.

An important limitation is that the results of our simulations are contingent upon our method for exploring the space of misspecified two-stage regression fits. Though we have found similar results (not shown) when we have considered other one-dimensional paths through the truth, it is possible that other approaches to making the fitted model differ from the correct model may lead to different results. More work is needed in this area to understand the extent to which parametric estimators in noisy settings may dominate semiparametric estimators of the SNMM, including the different possible scenarios under which this may or may not be true.

While the two-stage estimator can serve a stand-alone estimator for the intermediate causal effects, we also sought to evaluate its use as a method for obtaining guesses for the b_t in the G-estimator. We did this in a simulation experiment by comparing two implementations of the G-estimator, where one method (G-estimator A) uses estimates for b_t derived from the two-stage estimator. In the simulations, true models were used for the b_t in G-estimator B. G-estimator B dominated G-estimator A only slightly when misspecified models were used for the b_t. When the true b_t were used for G-estimator A, the two estimators performed similarly except for the parameters at the last time point where G-estimator A dominated. In the data analyses, both G-estimators provided similar results.

The methodology was illustrated with observational data from the PROSPECT study. A potential limitation of our analysis is that the assumption of sequential ignorability may be violated. The illustrative analysis assumes that only previous suicidal ideation and history of depression both (a) affect future depression outcomes, and (b) determines whether or not a patient switches off treatment at the next time point. Yet it is possible that subjects that were worse off, in terms of having a history of emotional problems or other physical ailments, are more likely to receive treatment at subsequent visits to the clinic. If this is true, then the estimates of β (under both estimators) are likely biased due to confounding. In observational studies, such as ours, the assumption of sequential ignorability is likely to be fulfilled when researchers proactively measure all important variables thought to be associated with time-varying treatment (e.g., in our case, switching off treatment). By design, the assumption is fulfilled in sequential multiple assignment randomized trials, in which patients are randomized to different treatment options (e.g., stay on treatment, or switch off treatment) at critical decision points over time.

A more in-depth analysis of the PROSPECT data will seek to better understand observed confounding, by discovering what are the predictors of time-varying treatment $\bar{A}_3$. If these observed, time-varying covariates (possible confounders) are also time-varying moderators of interest, they could be entered into the SNMM (as a part of $\bar{S}_K$) as described in this article; doing so would help resolve observed confounding bias *in addition* to understanding their role as putative time-varying moderators. If, on the other hand, these covariates are auxiliary (i.e., they are not time-varying moderators of interest), it may be possible to adjust for them using the estimation methods proposed here in combination with inverse-probability-of-treatment weights (Robins, 1999). This is a promising future research direction, currently being explored.

Under sequential ignorability and the (testable) assumption that there is no causal effect moderation by $\bar{S}_t(\bar{a}_{t-1})$, which implies that $\mu_1(s_1, a_1) = E\{Y(a_1, 0) - Y(0, 0)\} = \omega_1(a_1)$ and $\mu_2(\bar{s}_2, \bar{a}_2) = E\{Y(a_1, a_2) - Y(a_1, 0)\} = \omega_2(a_1, a_2)$, both the G-estimator and the two-stage estimator may be used to identify ω_1 and ω_2. Thus, the methods proposed here may be used in special cases to identify *marginal causal effects*, as in those indexing MSMs (Robins, 1999).

Robins (1994) describes a generalization of the G-estimator presented here for longitudinal outcome Y_t. In this case, a separate SNMM is specified for each t, one for each conditional mean Y_t given $(\bar{S}_t, \bar{A}_t)$. It should be relatively straightforward to extend the two-stage estimator in a similar fashion to accommodate longitudinal outcomes.

The notion of time-varying causal effect moderation, and the use of the SNMM, should find ready application in the medical sciences, as well as in the social and behavioral sciences where structural equation modeling (SEM) approaches are more common. The methods described here can be generalized to accommodate latent constructs that are typical in SEM implementations. An important distinction is that with the SNMM we parameterize the causal functions of interest directly; whereas, typical SEM implementations parameterize the conditional mean as in equation (1) then derive causal meanings from the estimated parameters.

It remains to be seen how the two estimators discussed here will compare to a fully parametric maximum likelihood estimator (MLE), such as the one proposed by Robins (1997). As moments-based estimators, neither the two-stage estimator nor Robins' G-estimator require distributional assumptions (e.g., normality) on the full likelihood for $(\bar{S}_K, \bar{A}_K, Y)$. An MLE, on the other hand, requires positing distributional assumptions for the distributions of Y given $(\bar{S}_K, \bar{A}_K)$, and S_t given $(\bar{S}_{t-1}, \bar{A}_{t-1})$. In addition, because parametric models for the ϵ_t's in the SNMM rely on the distribution of S_t given $(\bar{S}_{t-1}, \bar{A}_{t-1})$, parameters in models for the distributions of S_t given $(\bar{S}_{t-1}, \bar{A}_{t-1})$ will appear in models for the distribution of Y given $(\bar{S}_K, \bar{A}_K)$. Finding the MLEs for the parameters of an SNMM, therefore, may require nonlinear optimization routines and, we conjecture, may be sensitive to starting values. Some of the ideas discussed in this article, such as the use of linear models for the ϵ_t's (as in our implementation of the two-stage estimator) may facilitate a full likelihood implementation.

7. Supplementary Materials

Web Appendices referenced in Sections 2.2, 3.1, 3.2.1, 3.3, and 3.4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

Funding was provided by NIMH grants R01-MH-61892 (TTH), R01-MH-080015 (SAM), and a NIDA grant P50-DA-010075 (SAM). The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veteran Affairs.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, Maryland: Johns Hopkins University Press.

Bruce, M. L., Ten Have, T. R., Reynolds III, C. F., Katz, I. I., Schulberg, H. C., Mulsant, B., Brown, G., McAvay, G., Pearson, J., and Alexopoulos, G. (2004). Reducing suicidal ideation and depressive symptoms in depressed older primary care patients. *Journal of the American Medical Association* **291,** 1081–1091.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Goetghebeur, E. and Lapp, K. (1997). The effect of treatment compliance in a placebo-controlled trial: Regression with unpaired data. *Applied Statistics* **46,** 351–364.

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

Joffe, M. M., Small, D., and Hsu, C. (2007). Defining and estimating intervention effects for groups who will develop an auxiliary outcome. *Statistical Science* **22,** 74–97.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., and Agras, W. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry* **59,** 877–883.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

Petersen, M. L., Deeks, S. G., Martin, J. N., and Van Der Laan, M. J. (2007). History-adjusted marginal structural models for estimating time-varying effect modification. *American Journal of Epidemiology* **166,** 985–993.

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

Robins, J. M. (1989a). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Services Research Methodology: Focus on AIDS*, 113–159. Washington, DC: US Public Health Service.

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

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

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

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

Robins, J. M. and Rotnitzky, A. (2001). Comment on the Bickel and Kwon article, “Inference for semiparametric models: Some questions and an answer.” *Statistica Sinica* **11,** 920–936.

Robins, J. M. and Rotnitzky, A. (2005). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika* **91,** 763–783.

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

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Maisto, S. A., Brown, G. K., and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics* **63,** 926–934.

Vansteelandt, S. and Goetghebeur, E. (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society, Series B* **65,** 817–835.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th edition. New York: Springer.

Received September 2007. *Revised December* 2008.
Accepted December 2008.