Jennifer A. Faerber, Marshall M. Joffe*, Dylan S. Small, Rongmei Zhang,
Gregory K. Brown and Thomas R. Ten Have

A Simple Model Allowing Modification of the Effect of a Randomized Intervention by Post-Randomization Variables

DOI 10.1515/jci-2015-0016

Abstract: We address several questions relating to the use of standard regression and Structural Nested Mean Model (SNMM) approach (e. g., Ten Have et al. 2007) to analyze post-randomization effect modifiers of the intent-to-treat effect of a randomized intervention on a subsequent outcome, which has not been well examined. We show through simulations that the SNMM performs better with respect to bias of estimates of the intervention and interaction effects than does the corresponding standard interaction approach when the baseline intervention is randomized and the post-randomization factors are subject to confounding, and even when there is no association between the intervention and effect modifier. However, causal inference under the SNMM makes untestable assumptions that the causal contrasts do not vary across observed levels of the intervention and post-randomization factor. In addition, the precision of the SNMM-based estimators depends on the effect of the randomized intervention on the post-randomization factor varying across baseline covariate combinations. These issues and methods are illustrated with the application of the standard and causal methods to a randomized cognitive therapy (CT) trial, for which there is a conceptual model of negative cognitive styles or distortions impacted by CT but then in turn modifying the effect of CT on subsequent suicide ideation and social problem solving outcomes.

Keywords: cognitive therapy, depression, interaction, stratification

A causal model for post-randomization effect modification of randomized interventions

In the context of randomized trials of interventions of some duration, there is interest in identifying post-randomization factors early in the intervention that may lead to tailoring later stages of the intervention to enhance its effectiveness on subsequent outcomes, which would be ultimately assessed with within-patient sequential randomization designs [1]. Preliminary evidence of such tailoring to an individual's behavioral factors in the early stages of an intervention may include assessments of modification by these post-randomization factors of the effect of the intervention on subsequent outcomes of clinical interest. Early evidence that the treatment is performing poorly might be used to suggest that the treatment course be modified.

***Corresponding author: Marshall M. Joffe**, Department of Biostatistics and Epidemiology, Perlman School of Medicine, University of Pennsylvania, Blockley Hall, 6th floor 423 Guardian Drive, Philadelphia, PA 19104, USA, E-mail: mjoffe@mail.med.upenn.edu

Jennifer A. Faerber, Roberts Center for Pediatric Research, 2716 South Street, Philadelphia, PA 19146, USA, E-mail: faerberj@email.chop.edu

Dylan S. Small, Department of Statistics, The Wharton School, University of Pennsylvania, 400 Huntsman Hall, Philadelphia, PA 19104, USA, E-mail: dsmall@wharton.upenn.edu

Rongmei Zhang, Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD, USA, E-mail: rongmei.zhang@fda.hhs.gov

Gregory K. Brown, Department of Psychiatry, Perlman School of Medicine, University of Pennsylvania; 3535 Market Street, Philadelphia, PA 19104, USA, E-mail: gregbrown@mail.med.upenn.edu

Thomas R. Ten Have, Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104, USA

To search for post-randomization factors that may lead to effective tailoring, we consider whether a variable is a “post-randomization effect modifier,” which we define in the following way. The standard terminology is that a mediator is a variable that is on the causal pathway between the treatment and the outcome whereas a moderator is a pre-randomization variable that modifies the size of the causal effect of a treatment. We say that a post-randomization variable is an effect modifier if the value that the post-randomization variable would take if the person was randomized to treatment is an effect modifier. The value that the post-randomization variable would take if the person was randomized to treatment can be thought of as a pre-randomization covariate since it is defined regardless of whether the person was randomized to treatment or control under the potential outcomes framework, even though it is only observed when the person is randomized to treatment [2, 3]. Emsley et al. (2010) [4] discuss post-randomization effect modifiers measured after treatment assignment. A post-randomization effect modifier may or may not be a mediator but it also might just be associated with a mediator (i. e., the post randomization effect modifier might not be on the causal pathway itself but be associated with a mediator that is on the causal pathway). Such a post-randomization effect modifier might be useful in tailoring a treatment in a setting where the mediator is not directly measurable but the post-randomization effect modifier, which is associated with the mediator, is measurable. Emsley (2013) [5] provide an example of such post-randomization effect modifiers when examining the effect of cognitive behavioral therapy on depression using antidepressant therapy as a mediator of interest. Here treatment process measures (treatment fidelity, therapeutic alliance) are potential post-randomization effect modifiers that might not be available for both treatment arms, but might be associated with antidepressant therapy and might explain the differential treatment effects observed with this mediator. These relationships can be better visualized using the path diagram below (see Figure 1). In this diagram, the arrows indicate a causal effect of one variable on another variable.

Figure 1: Path diagram.

For assessing whether a variable measured at baseline is an effect modifier, it is appropriate to examine a baseline covariate-intervention interaction in a standard linear regression or ANCOVA approach to examine effect modification [6, 7]. But when the effect modifier is measured after randomization, these standard ways of examining such effect modification may be biased, because they deal inappropriately with a post-treatment variable that may have been affected by the treatment [4, 8, 9]. Similar issues arise in landmark analyses in randomized cancer trials, where patients are stratified on early post-randomization response to treatment and then treatment effects on mortality are estimated and tested within the resulting strata [10].

To deal with these issues, we present a causal model based on a modified version of a Structural Nested Model that allows for modification of the effect of the main randomized treatment by a post-randomization factor (SNMM; [9, 11, 12]). We consider an estimation method based on two-stage least squares (2SLS) that may be fit using standard software.

This paper is organized as follows. We begin by providing motivation for our methods, using a trial of cognitive behavioral therapy (CT) on subsequent suicide ideation and social problem solving. We then present our models and estimation procedures. We use simulation to investigate the bias and precision of

our approaches and compare them with standard approaches. We apply the various methods to the randomized trial CT, and conclude with a discussion.

Cognitive therapy and tailoring factors

In studies of the effect of CT on depression, there is strong reason to think that there are distinct subgroups defined by post-treatment variables in which the effects of later treatments may vary, and so subsequent treatments might be tailored to those early post-treatment variables. Thus, in preliminary investigations, we may be interested in identification of psychological factors in the early window of the intervention that may modify the effects of later stages of the intervention on outcome. In the context of efficacy studies of cognitive-behavioral therapy for depression, Ilardi and Craighead [13] identified that the majority of symptom amelioration takes place within the first month of treatment and their research provided the groundwork for other researchers (e. g., [14]) to explore the cognitive and therapeutic mechanisms through which therapies achieved these effects. Instead of focusing on rapid early response to treatment, we examine here negative cognitive styles, such as early depression that might indicate that patients are likely to fail on their current treatment.

The presence of such early post-treatment modifiers of treatment effect may play a role in difficulty in finding effectiveness of some behavioral interventions [15, 16]. In the context of the CT intervention and its impact on suicidal thinking and problem solving in suicide attempters, such a factor may involve hopelessness, which is thought to consist of several constituent factors, including a stable, enduring component and a more variable component based on conditional or situational factors responsive to the initiation of the cognitive therapy [17]. This latter component has been found to predict depression severity above and beyond baseline levels of hopelessness and depression [17, 18], and involves negative thinking or problem orientation (i. e., hopelessness, self-criticism), behavioral problems (i. e., avoidance, passivity), strategies to instill hope (i. e., helping patients get a job), and suicide ideation and needs related to the recent suicidal crisis [19]. All of these factors may represent a common factor that identifies early in the intervention patients who will not respond to early treatment but may respond later in or after treatment. As representatives of these factors, we focus on an early post-randomization measure of depression, which is an early target of the CT intervention and whose levels predict the effect of the CT intervention on subsequent outcomes such as future suicide attempts and reduction of suicide ideation. Such an analysis of measures related to a common post-treatment factor may help tailor potential adaptive treatment strategies that optimize separate treatments for early responders and non-responders. In general, chronic or relapsing psychiatric disorders (e. g. major depression, substance abuse) are effectively treated using a dynamic treatment plan, which involves sequential treatment components [20]. In the absence of a universally effective approach to treating these disorders, clinicians have to individually adjust each patient's dose and/or type of treatment to manage his/her illness(es). According to the dynamic treatment approach, a clinician might decide on a course of treatment for a patient, decide when to change the treatment, and choose the next treatment if the prior treatment was not successful [1]. Prior to reassigning a patient to the next treatment, a clinician might use tailoring factors in the clinical decision making process. In this context, tailoring factors are patient characteristics (e. g. psychological characteristics, response to medication etc.) measured after the administration of a treatment (and therefore affected by the treatment) that might be used to inform what next treatment would be best. Depending on how the patient responds to the earlier treatments determines the range of treatment options available to the patient in later stages of the intervention. These patient responses measured in theoretically guided ways serve as tailoring factors for the subsequent treatment decisions.

In the context of this paper's focus on exploratory identification of tailoring factors for a cognitive therapy (CT) intervention on suicide ideation and social problem solving style in suicide attempters, a potential dynamic treatment plan involves two decisions:

- 1) an initial decision assigning patients to CT or usual care; and
- 2) stratifying patients on the basis of a tailoring factor (e. g. psychological state affected by CT that is predictive of future behavior in the intervention) and within each strata, deciding on whether to use a relapse prevention task [20]. Patients in the lowest stratum of the tailoring factor (i. e. patients better off, exhibiting the lowest levels of depression severity at this one month intermediate time) might be candidates to take a relapse-prevention task, which, if successfully completed, means that they could end their CT sessions. In a relapse-prevention task, patients are asked to focus on the events and cognitions associated with the previous suicide attempt and explain whether they could employ adaptive rather than self-harming behaviors. Cognitive therapy would end for patients who successfully passed this task. Patients in the highest stratum of the tailoring factor would likely not be administered the relapse prevention task, but their performance on the tailoring factor could suggest that they need additional CT and the remaining course of therapy modified based on the severity of their depression score in this intermediate window.

While the specific CT trial used for this paper does not offer actual data on such a dynamic treatment regime, it does offer data to assess if the effect of CT on depression is enhanced for certain percentiles of negative cognitive style factors (e. g., depression). Examination if such effect modification took place may be used to justify using this negative cognitive style factor as a tailoring variable, i. e., stratifying patients on these factors for subsequent CT treatment decisions. In general, the approach of this paper offers a way of analyzing data from a simple randomized trial such that it provides preliminary evidence of how tailoring factors can be used in a future sequential randomization treatment study.

A cognitive therapy study [21] provided the context for examining how the randomized CT effect was modified by a post-randomization factor. The subjects were 120 suicide attempters who received medical or psychiatric evaluation at the Hospital of the University of Pennsylvania within 48 hours of the attempt and were recruited from the hospital emergency department for the original study. In this study, the subjects were randomized to receive or not receive ten sessions of either cognitive therapy for suicidal behavior or no therapy, but everyone received usual care services in the community (UC). Depression score at one month was examined as a potential modifier of the effect of the CT treatment on suicide ideation at three months and also on six month impulsive/careless style social problem solving. Reducing levels of depression severity might not have been one of the primary goals of the CT treatment for suicide attempters, but depression severity was expected to be affected by cognitive therapy trial one month after randomization and predictive of subsequent improvement at three and six months in suicide ideation and social problem solving. In addition, one month depression severity might be associated with other (unmeasured) clinical markers of improvement in the early window of the trial. One month depression severity was measured by the Beck Depression Inventory-II (BDI-II; [22]), three month suicide ideation by the Scale for Suicide Ideation-Worst (SSIW; [23]), and six month impulsive/careless problem solving style (ICS) by the Social Problem Solving Inventory-Revised (SPSI-R; [24]). We will compare our formal causal model approach to estimating effect modification by one-month depression with two standard interaction regression approaches using data from the cognitive therapy study. We will also examine the performance of the causal model compared to the same two standard regression models in simulation studies in which we vary the strength of the contribution of baseline covariates and baseline covariate interaction terms in explaining the variability in the model.

Methods

Notation

This section presents notation for observed variables, then for the potential outcomes used to define causal effects [2, 3].

For the observed variables, Y defines the outcome random variable [three month suicide ideation (SSIW) or six month impulsive/careless style (ICS)]; R is the binary indicator of randomized treatment assignment (1 = CT; 0 = usual care only); M is the post-randomization variable (1 month depression severity (BDI-II), for which there will be separate analyses on suicide ideation and impulsive/careless style); and X is the vector of observed baseline covariates.

Average causal effects are defined in terms of potential outcomes. Thus, Y_{rdec} would be observed if a participant were assigned to level r of the CT intervention. Similarly, M_r denotes the level of M that would be observed were a participant assigned level r of the CT intervention. Effects are defined in terms of contrasts of potential outcomes; thus, the effect of the CT intervention on outcome for an individual is $Y_1 - Y_0$. The expected value of these contrasts is defined as the average causal effect and is no longer on an individual-level. Throughout, we use uppercase letters to refer to a random variable and lowercase letters to refer to a hypothetical level of that variable.

Models

We present the proposed interaction SNMM model and the analogous standard regression model. Of particular importance are the contrasts between the association effects under the standard regression approach and the causal effects under the SNMM.

SNMM model

The proposed interaction Structural Nested Mean Model is defined in terms of the potential outcome for fixed levels of random assignment (R) and post-randomization factors (M):

$$E(Y|R, M, X) - E(Y_0|R, M, X) = R\theta_R + RM\theta_{RM}. \quad (1)$$

which can be expressed as

$$E(Y|R=1, M_{r=1}, X) - E(Y_0|R=1, M_{r=1}, X) = R\theta_R + RM\theta_{RM}$$

The parameter θ_R is the effect of treatment for subjects whose mediating variable would be 0 if they were assigned to treatment (i. e., subjects with $\theta_R = E(Y_1 - Y_0|M_{r=1}=0)$). The parameter θ_{RM} describes how the effect of R varies with the value of M a subject would have if she were assigned to the treatment group, specifically θ_{RM} is how different the treatment effect is for subjects whose M would be 1 if they were assigned to the treatment than for subjects whose M would be 0 if they were assigned to the treatment, $\theta_{RM} = E(Y_1 - Y_0|M_{r=1}=1) - E(Y_1 - Y_0|M_{r=1}=0)$

There is said to be post-randomization effect modification by M if $E(Y_{r=1} - Y_{r=0}|M_{r=1}=m) \neq E(Y_{r=1} - Y_{r=0}|M_{r=1}=n)$, for some $m \neq n$ [25].

Standard regression model

We present two forms of standard linear regression models corresponding to (1), above. The first has a nearly identical form, including an interaction between R and M :

$$E(Y|\mathbf{X}, R, M) = \mathbf{X}\beta_S^T + R\beta_R + RM\beta_{RM}, \quad (2)$$

where β_S^T is a transposed vector of effects of the baseline covariates, $\mathbf{X}$. An alternative standard regression follows the usual convention by also including a main effect of M :

$$E(Y|\mathbf{X}, R, M) = \mathbf{X}\gamma_S^T + R\gamma_R + M\gamma_M + RM\gamma_{RM}. \quad (3)$$

We will contrast both of these models with the structural nested mean model (1). Both the SNMM and standard linear regression models adjust for the baseline level covariates in the model.

Model assumptions

We present in more detail assumptions for identifiability and unbiased estimation of parameters and conditions for more efficient estimation under the SNMM and standard regression methods in the sections below. We present additional assumptions and conditions for more efficient estimation but these assumptions are not necessarily required for identifiability and unbiased estimation. We assess such assumptions and conditions in the simulations and data analyses.

SNMM assumptions

For the SNMM model, the assumptions necessary for unbiased inference are: (1) randomization (i. e., ignorability) of baseline intervention assignment; (2) Stable Unit Treatment Value Assumption (SUTVA) (not testable) [3]; (3) independence of observations for standard error estimation (not testable); and (4) model assumptions including no-interaction assumptions among baseline covariates, the baseline randomized intervention, and the post-randomization modifier (both testable and not testable).

The randomization assumption implies stochastic independence between the randomization indicator for the baseline intervention, R , and the potential outcomes within strata defined by baseline covariates $\mathbf{X}$:

$$\Pr(Y_0, Y_1 | R, X) = \Pr(Y_0, Y_1 | X) \quad (4)$$

This assumption means that the distribution of observed and unobserved baseline covariates should be balanced between the randomized baseline intervention groups. Although this assumption is not testable, it is implied by physical randomization.

The Stable Unit Treatment Value Assumption (SUTVA; [3, 26]) consists of two sub-assumptions. First, for each participant, there is a single value of the potential outcome random variable corresponding to the random assignment level r , Y_r , regardless of the randomization assignment of any other participant. This assumption implies that Y_r is defined with scalar indices for a given participant, rather than vectors of indices representing baseline intervention assignments and post-randomization modifier levels of all patients. Additionally, SUTVA implies that for each treatment level r , a single value for the potential outcome Y_r and potential post-randomization variable M_r exists regardless of how the treatment level r was administered. We outline two approaches to estimating parameters in the SNMM. First is a two-stage least squares (2SLS) approach. To apply this approach, we regress the outcome on randomization and baseline covariates:

$$\begin{aligned} E(Y|R, X) &= \sum_m E(Y|R, M=m, X) \Pr(M=m|R, X) \\ &= \sum_m \{E(Y_0|R, M=m, X) + R\theta_R + RM\theta_{RM}\} \Pr(M=m|R, X) \\ &= \sum_m \{E(Y_0|R, M=m, X) \Pr(M=m|R, X) + R\theta_R \Pr(M=m|R, X) + RM\theta_{RM} \Pr(M=m|R, X)\} \\ &= E(Y_0|R, X) + R\theta_R + RE(M|R, X)\theta_{RM} \\ &= E(Y_0|X) + R\theta_R + RE(M|R, X)\theta_{RM} \end{aligned}$$

Here, the second equality follows by substituting the SNMM into the equation, and the final equality $E(Y_0|R, X) = E(Y_0|X)$ follows due to randomization. We can use a two stage procedure to estimate the parameters in the model; we first estimate $E(M|R, X)$ by regressing M on R and X , then estimate $\theta \equiv \{\theta_R, \theta_{RM}\}$ by regressing Y on X , R , and R times the estimate $\hat{E}(M|R, X)$. If both stages are performed

using least squares, this is a two-stage least squares procedure. Standard software for 2SLS accounts in the variance estimation for the fact that $\hat{E}(M|R, X)$ is estimated.

2SLS is an asymptotically efficient approach under the SNMM assumptions, homoscedasticity, and a linear outcome model [27]. The SNMM estimate is less efficient than a standard regression model because it is using only a part of the variation in M to estimate the effect modification of R by M . The SNMM extracts the part of the variation in M that is due to X and uses this confounder-free variation (under the SNMM assumptions) to estimate the effect modification by M whereas the standard regression approach uses the whole variation in M (but this variation may be confounded, leading to bias of the standard regression approach).

We concentrate on the 2SLS approach and use it in simulation and practice, because standard software could be used. G-estimation [28, 29] is a more general approach that can be used in a larger variety of settings. G-estimation is less reliant on parametric assumptions and extends to time-varying treatments. For G-estimation, we note that Y_0 is independent of R given X . We can then create a random variable $U(\theta) \equiv Y - R\theta_R - RM\theta_{RM}$. We can use estimating equations:

$$\sum_i (R - p) \{U(\theta) - \mu_X\} W = 0,$$

where W is a vector of non-collinear weights of dimension 2, and μ_X is some function of X . Typically, we choose $\mu_X = E\{U(\theta)|X\}$ or $\mu_X = 0$, and $W_X = \{1, E(M|R=1, X)\}^T$.

Estimation using these equations or the 2SLS approach requires that $E(M|R=1, X)$ vary with X ; otherwise we have collinearity. Estimation using these approaches essentially assumes that variation with X in the causal effect of R on Y can be explained by variation across levels of the baseline covariate X in the expected value of the post-treatment variable M in the treatment group. The reasonableness of this assumption should be assessed in any application. This assumption may be relaxed by allowing the effect of treatment to vary across levels of some subset of covariates; nonetheless, the general approach requires that the variation in the effect of R on Y across some subset of covariates X be explained by the variation in the effect of R on M across those covariates. Under our design and other assumptions, this assumption is not fully testable and its reasonableness must be assessed in each setting.

In the context of our example, estimation of effect modification depends on having pre-treatment predictors of the expected value of early outcomes (e. g., depression severity at one month) that in turn predict the subsequent effects of the intervention on outcomes at three to six months. The efficiency of the estimation procedure depends on the strength of these relationships.

A note on standard regression

The standard regression estimates of the parameters β or γ will in general be biased as estimators of causal effects θ if M is affected by R , since they condition or adjust on a variable M affected by the main treatment R [30, 31]. We note that M is likely to be affected by R in our data given the theory of the common treatment factor that predicts an early effect of the CT intervention on one month depression severity, which in turn will modify subsequent effects of the intervention on three and six month outcomes.

Results

Data analyses

Overall, the intent-to-treat (ITT) effect of CT on the outcomes of interest, three month suicide ideation and six month impulsive social problem solving style, is of borderline significance ($\hat{\beta} = -3.31$, $t(105) = -1.46$, $p = 0.091$, 95 % CI $[-7.17, 0.54]$) and ($\hat{\beta} = -1.54$, $t(99) = -1.68$, $p = 0.097$, 95 % CI $[-3.37, 0.28]$), respectively, which was not surprising given the small sample size, but worthy of further investigation. Though the ITT effect on either outcome was only marginally statistically significant, the CT group fared better at

the end of the outcome window, which is reflected in a lower mean outcome score (with standard deviation in parentheses) at three (six) months for the CT group than the usual care group, 9.22 (1.40) and 11.88 (1.37), respectively for three month SSIW and 7.20 (0.67) and 8.69 (0.65), respectively for six month ICS. Under the common treatment factor model, one may expect that one month depression severity would modify this ITT effect on three to six month outcomes. To help assess these interactions under the SNMM, we adjusted for the same set of baseline covariates under both models since the same post-randomization factor was used in both models, but also adjusted for the baseline level of the outcome variable in each model. These covariates included baseline versions of the effect modifiers themselves, history of suicidal behavior, depression, and numerous variables involving problem-solving characteristics such as “avoidant style,” “impulsive-careless style,” and “positive” and “rational problem solving.” Under the common treatment factor model, one would expect some effect of CT on depression severity. The corresponding effect size for the observed ITT effect on one month depression severity (mean group difference divided by the standard deviation of one month depression severity) is negligible (effect size = 0.0089).

Figures 2 and 3 show SNMM and standard regression-based estimates of the CT effects on three month SSIW (Figure 2) and six month ICS (Figure 3) across values of one month depression severity. The standard regression models (2) and (3) are labeled Standard (2) and Standard (3), respectively.

Figure 2: Stratified effect of CT on SSIW.

The ensuing results show that the SNMM and standard approach without a main effect of depression severity (model 2) results disagree to some extent in supporting the common treatment factor hypothesis by showing varying effects of CT across levels of one month depression severity under the SNMM, but not standard regression models. The two modeling approaches disagree with respect to the CT-depression severity interaction in both SSIW and ICS models. However, the standard regression approach including a main effect of depression severity (standard 3) yielded results that agree more qualitatively with the SNMM results for both the main and interaction effects; however, the standard estimates were more attenuated to 0, possibly due to unobserved confounding not taken into account in the model or departures from other SNMM assumptions.

Figure 3: Stratified effect of CT on ICS.

For the SSIW and ICS outcome models results in Figures 2 and 3 (and associated Tables 1 and 2), respectively, the two approaches (SNMM and standard 2) agree with respect to the main effect of CT on outcome, but not for the interaction of CT and depression severity.

Table 1: Main effect of CT and interaction term on SSIW.

Statistical method	Effect	$\hat{\beta}$	SE	95 % CI	P
SNMM	CT Main Effect	5.47	4.90	(-4.14, 15.07)	0.27
	Interaction	-0.45	0.20	(-0.85, -0.057)	0.027*
Standard Regression (2)	CT Main Effect	-4.51	2.90	(-10.27, 1.25)	0.12
	Interaction	0.0050	0.098	(-0.19, 0.20)	0.96
Standard Regression (3)	CT Main Effect	2.07	3.38	(-4.63, 8.78)	0.54
	Interaction	-0.27	0.12	(-0.52, -0.026)	0.030*

Note: $\hat{\beta}$ = Unstandardized Beta Coefficient; SE = Estimated Standard Error; 95 % CI = 95 % Confidence Interval for $\hat{\beta}$.

* $p < 0.05$, two-tailed test.

Table 2: Main effect of CT and interaction term on ICS.

Statistical method	Effect	$\hat{\beta}$	SE	95 % CI	P
SNMM	CT Main Effect	2.25	2.11	(-1.89, 6.38)	0.29
	Interaction	-0.18	0.089	(-0.36, -0.0077)	0.044*
Standard Regression (2)	CT Main Effect	-1.58	1.38	(-4.31, 1.16)	0.26
	Interaction	0.0016	0.049*	(-0.095, 0.098)	0.97
Standard Regression (3)	CT Main Effect	2.44	1.55	(-0.65, 5.52)	0.12
	Interaction	-0.17	0.059	(-0.28, -0.051)	0.0054**

Note: $\hat{\beta}$ = Unstandardized Beta Coefficient; SE = Estimated Standard Error; 95 % CI = 95 % Confidence Interval for $\hat{\beta}$.

* $p < 0.05$, two-tailed test.

** $p < 0.01$, two-tailed test.

In Figure 2 and Table 1, the interaction estimate under the SNMM model shows that the reduction in three month suicide ideation by CT is enhanced by 0.45 SSIW (95 % CI [−0.85, −0.057]) units for every unit increase in one month depression. This result suggests that CT has a stronger effect on reducing suicide ideation at higher levels of depression that is not resolved by the first stage of the CT intervention. The standard regression approach which includes a main effect of depression severity (standard 3) agreed qualitatively with the SNMM-based results, but was of a lower magnitude: the reduction in three month suicide ideation by CT is enhanced by 0.27 units for every one unit increase in depression (95 % CI [−0.52, −0.026]). However, the standard regression (standard 2) indicates that the CT effect is reduced by only 0.0050 SSIW units (95 % CI [−0.19, 0.20]) for every unit increase in depression severity. In accordance with the non-significant interaction under the standard regression model (2), the corresponding estimates of the CT effect for the 25th, 50th and 75th depression severity percentile groups are all of similar magnitude in Figure 2.

Although the 95 % confidence interval for the effect of CT on SSIW and ICS is comparable near the mean value of depression severity under the SNMM and Standard approaches, the SNMM confidence interval is much wider away from the mean (Figures 2 and 3). For most parameters, the SNMM estimates are less precise than the corresponding standard regression estimates, with one exception in Table 2.

Figure 3 and Table 2 also show that the SNMM and standard regression (standard 2) approaches agree more with respect to inference for the main effect of CT than for depression severity-CT interaction on ICS. The SNMM model shows that the reduction in six month impulsive/careless problem solving style by CT is enhanced by 0.18 ICS units for every unit increase in one month depression severity (95 % CI [−0.36, −0.0077]). The Standard regression approach which includes a main effect of depression severity (standard 3) yielded a similar point estimate to the SNMM estimate of the effect of CT ($\hat{\beta} = -0.17$, 95 % CI [−0.28, −0.051]). In contrast, the corresponding standard regression (standard 2) interaction estimate reveals that the CT effect is slightly enhanced by 0.0016 ICS units (95 % CI [−0.095, 0.098]).

The differences in the above results between the SNMM and standard regression (standard 2) approaches, especially with respect to the interaction terms, can be attributed to several factors including the effect of CT on the level of the effect modifier measured at one month, unobservable factors involving unmeasured confounders, and deviation from other SNMM assumption. We did not observe a large or statistically significant effect of CT on one month depression severity ($\hat{\beta} = -1.12$, 95 % CI [−5.69, 3.45]).

The absence of such an effect, which is consistent with our data, would make standard methods acceptable to use, and, when there is no such effect, standard estimates and our SNMM estimator should, in expectation, produce the same results. Nonetheless, the parameter estimates based on our new methods were larger in magnitude. Thus, it is of some interest to investigate possible sources of the discrepancy, which we do using simulation.

Simulation studies

We present simulation results in an attempt to understand how the degree of unmeasured confounding of the depression severity-outcome association (i. e., the strength of association of M and Y_0) might explain the discrepancies present in Table 1, where the magnitude of effect estimates by the SNMM approach is larger than that of standard regression (standard 3). However, we do not test other SNMM assumptions (i. e. the no interaction assumption) which also might contribute to the discrepancy found between the two methods in the data analyses.

We performed Monte Carlo simulations to measure the accuracy and precision of the SNMM 2SLS and standard regression estimators by varying the following design factors: (1) the magnitude of the contribution of the set of predictor variables in explaining the post-randomization variable (with R^2 either 0.2 or 0.8); (2) the magnitude of the departure from sequential ignorability (i. e., the assumption that, in addition to randomization of the initial treatment R , the post-randomization variable is randomly assigned in the

treatment group, or that $M \perp Y_0 | R=1$); and (3) the strength of the association of the baseline covariate-randomization interaction terms with the post-randomization variable (with these variables either explaining about half of the variation detected by R^2 or a negligible amount of the variation in M). In each simulation, we generated 1000 independent datasets of sample size 500. We created synthetic baseline variables (X_1 - X_5), a dichotomous randomization indicator (R), a linear function of the observed baseline variables (Y_0) to represent unmeasured confounding of the relationship between the interaction term and the outcome, a post-randomization continuous variable (M), an interaction between the post-randomization variable and randomization indicator ($R \cdot M$), and the outcome variable (Y).

Simulation description

Baseline variables

Five baseline variables were generated as uncorrelated random normal variables.

Measure of baseline confounding

Y_0 was specified as the parameter representing unobserved confounding between Y_r and M . This parameter was generated as a linear function of the baseline variables and normally distributed errors with a mean of 0 and constant variance (eq. (5)). We created Y_0 such that the set of baseline variables explained 50 % of the variability in Y_0 . Y_0 is included as a component in both the linear functions used to create both the M (eq. (6)) and the Y (eq. (7)) as a baseline measure associated with both M and Y .

$$Y_0 = \beta_0 + \beta_1 X_1 + \dots + \beta_5 X_5 + \epsilon \quad (5)$$

Post-randomization variable

The post-randomization variable (M) was created as a linear function of an intercept, the baseline variables, interactions between baseline variables and randomization ($X_1 \cdot R$, ..., $X_5 \cdot R$), and error. Specifically, this error parameter was incorporated by decomposing the error term in eq. (1) ϵ into two additive parts: $\epsilon = \epsilon^* + a Y_0 \cdot Y_0$ where ϵ^* is an independent normal random variable with a mean of zero and variance of one.

$$M = \beta_0 + \beta_1 X_1 + \dots + \beta_5 X_5 + \beta_6 (X_1 \cdot R) \dots + \beta_{10} (X_5 \cdot R) + \beta_{11} R + \beta_{12} Y_0 + \epsilon \quad (6)$$

The coefficient of Y_0 in eq. (5) above represents the level of degree of the departure from sequential ignorability. Under sequential ignorability, the coefficient of Y_0 is zero. A non-zero coefficient of β_{12} in (6) indicates a departure from sequential ignorability.

Outcome model:

The outcome is a function of the effect of randomization, the interaction term, and the measure of baseline confounding of the relationship between the post-randomization variable and the outcome. All coefficients in this model were set equal to one.

$$Y = Y_0 + \varphi_1 R + \varphi_2 (M \cdot R) \quad (7)$$

Simulation conditions to be varied:

1. Strength of associations between predictors and the post-randomization variable

The SNMM approach is expected to perform better if the interactions terms between the set of baseline covariates X and the randomized interaction R in predicting levels of the post-randomization variable M are large. In the presence of sequential ignorability, we obtained coefficient values for the baseline variables, randomization, and the baseline covariate-randomization interactions such that the contribution of the entire set of variables on the effect of the post-randomization variable was *low* ($R^2 = 0.2$) or *high* ($R^2 = 0.8$).

2. Departure from Sequential Ignorability

The performance of both methods under departures from sequential ignorability is examined since in reality the assumption of sequential ignorability is unlikely to be met and we do not know the true level of departure from it. Since the 2SLS method is asymptotically unbiased in principle, given certain necessary assumptions are met, and observed to be unbiased in our simulations, per our simulation conditions, we present the bias in standard regression estimators only and the precision of both the 2SLS and standard regression estimators.

We varied the departure from sequential ignorability by varying the magnitude of the coefficient of Y_0 in (6); the degree of departure was the same in both treatment groups (i. e., $R=0$ and $R=1$). A *moderate* departure from sequential ignorability was indicated by a coefficient of Y_0 such that the regression predicting the M explains an additional 12.5 % of the unexplained variability in the M ($R^2 = 0.3$ (under the low condition described in design factor 1 above) or $R^2 = 0.825$ (in the high condition from design factor 1)). A *larger* departure from sequential ignorability was indicated by a coefficient of Y_0 such that the regression predicting M explained an additional 50 % of the unexplained variability in M ($R^2 = 0.6$ in the low condition or $R^2 = 0.9$ in the high condition).

3. Strength of the contribution of the baseline covariate- randomization interaction terms

We varied the extent to which the baseline covariate-randomization interactions explained the variability in R^2 . For R^2 values for the regression of the post-randomization variable on X_1 - X_5 , $X_1 \cdot R$, ..., $X_5 \cdot R$, and Y_0 , the interaction terms explained either half of the variation (strong contribution) or a negligible amount (less than 5 % of that variation, which was considered a weak contribution).

Simulation outcome measures

We then analyzed each simulated dataset with the 2SLS estimation under the SNMM in (1) and ordinary least squares estimation under the standard regression model in (2). For each of the 2SLS and standard regression estimators of the main effect of R and $R \cdot M$ interaction, Tables 3–6 report the absolute bias, standard error, and confidence interval coverage (percentage of the iterations for which the 95 % confidence interval included the true values of θ_R or θ_{RM}). In addition, we examined the Type I error made for the interaction effect: the percentage of simulations in which the interaction term was found to be significantly different from zero when the true effect is zero.

Table 3: Simulation results: Weak association between X and M ($R^2 = 0.2$) and weak set of $X \cdot R$ interaction terms in model for post-randomization variable.

		Lower level of confounding			Higher level of confounding		
	Method (rows)	Bias	% Coverage	Standard Error	Bias	% Coverage	Standard Error
Effect = R	SNMM	−0.0018	95.0	0.12	−0.0058	95.4	0.11
	Standard (2)	−0.096	65.3	0.062	−0.38	0	0.067
	Standard (3)	−0.015	95.3	0.069	−0.036	91.9	0.069
Effect = $R \cdot M$	SNMM	0.0010	95.3	0.068	0.0023	95.1	0.042
	Standard (2)	0.060	43.4	0.029	0.15	0	0.022
	Standard (3)	−0.0054	93.7	0.039	−0.014	90.4	0.026

Note: True values for R and $R \cdot M$ are 1; Cov = Coverage; Std Err = Standard Error.

The main findings from the simulations are as follows. Our 2SLS approach is essentially unbiased in all settings. The efficiency of our 2SLS approach is much improved when the association of X and M is strong. Nonetheless, the variance of the 2SLS approach is larger than that of the standard approaches.

Table 4: Simulation results under weak association between X and M ($R^2 = 0.2$) and strong set of X*R interaction terms in model for post-randomization variable.

		Lower level of confounding			Higher level of confounding		
	Method (rows)	Bias	% Coverage	Standard Error	Bias	% Coverage	Standard Error
Effect = R	SNMM	-0.00063	95.7	0.091	-0.0040	95.8	0.094
	Standard (2)	-0.078	72.0	0.056	-0.32	0	0.062
	Standard (3)	0.0091	96.5	0.066	0.059	86.1	0.067
Effect = R*M	SNMM	0.00031	96.0	0.059	0.0017	95.4	0.038
	Standard (2)	0.056	46.8	0.028	0.14	0	0.022
	Standard (3)	-0.013	93.0	0.040	-0.037	69.9	0.026

Note: True values for R and R*M are 1.

Table 5: Simulation results under strong association between X and M ($R^2 = 0.8$) and weak set of X*R interaction terms in model for post-randomization variable.

		Lower level of confounding			Higher level of confounding		
	Method (rows)	Bias	% Coverage	Standard Error	Bias	% Coverage	Standard Error
Effect = R	SNMM	0.00031	95.5	0.056	-0.0012	95.8	0.065
	Standard (2)	-0.047	85.9	0.052	-0.24	0.0090	0.057
	Standard (3)	-0.035	91.3	0.051	-0.095	44.8	0.045
Effect = R*M	SNMM	-0.00027	95.8	0.021	0.00028	95.7	0.016
	Standard (2)	0.025	71.4	0.018	0.075	0	0.013
	Standard (3)	-0.0055	93.5	0.020	-0.012	81.5	0.011

Note: True values for R and R*M are 1.

Table 6: Simulation results under strong association between X and M ($R^2 = 0.8$) and strong set of X*R interaction terms in model for post-randomization variable.

		Lower level of confounding			Higher level of confounding		
	Method (rows)	Bias	% Coverage	Standard Error	Bias	% Coverage	Standard Error
Effect = R	SNMM	0.00026	96.1	0.050	-0.0020	96.1	0.067
	Standard (2)	-0.035	89.9	0.048	-0.34	0	0.056
	Standard (3)	0.061	81.4	0.055	0.29	0	0.051
Effect = R*M	SNMM	-0.00024	95.4	0.016	0.00034	95.5	0.010
	Standard (2)	0.019	74.3	0.015	0.065	0	0.0079
	Standard (3)	-0.062	37.7	0.027	-0.14	0	0.015

Note: True values for R and R*M are 1.

However, standard approaches (standard (2)) which do not include main effects terms for the post-treatment variable M are generally biased throughout. Standard approaches (standard (3)) which do include a main effects term are also generally biased, but typically less than standard (2); this is in line with the usual convention that when interaction terms are included in a regression that the main effects should also be included. The bias of standard methods (especially (3)) is greatest when there are large departures from sequential ignorability, the association of covariates with M is large, and the $X*R$ interaction in the outcome model is large. The simulation results inform how the methods work under departures from sequential ignorability and the strength of the association of X and M , but do not directly explain any discrepancies found between the two methods in the data analyses.

Discussion

We focus on estimating the effect of a randomized behavioral intervention stratified on a behavioral post-randomization factor that might be impacted by the intervention in its early stages. This focus is motivated by the common treatment factor model, under whichsw factors influenced early in the intervention modify subsequent effects of the intervention on outcome. While the effect modifier under consideration in our example, one month depression, was not strongly impacted by early stages of CT, it appears to be associated with effects of CT on subsequent suicide ideation and social problem solving style, based on our analyses.

In our approach, we do not attempt to estimate the effect of the post-treatment variable M . Such effects may not be well-defined if it is not clear how one can intervene on the variable M or if different ways of affecting M might have different effects on the outcome Y [32, 33]. Our approach is agnostic as to whether the variable M (one month depression severity) plays such a role as a mediator or whether such mediation effects are estimable. Nonetheless, we have shown that, just as standard regression estimates of the effects of mediators may be biased if confounders of the mediator-outcome association are not properly adjusted, standard regression estimates of how effects of the main treatment are modified by a post-treatment variable M may be biased if there are confounders of the post treatment variable M -outcome association (i. e., if the association between M and outcome Y cannot be explained by baseline variables X) and M is affected by treatment.

In our application, there is the potential for unmeasured confounders of the effect modifier-outcome (one month depression severity versus three month suicide ideation (or six month impulsive/careless social problem solving)) relationship and also the effects of CT on the effect modifiers to produce bias for the standard regression approach to assessing interactions; nonetheless, the 2SLS approach we have developed can be much less precise than standard regression approaches. Consequently, we have performed parallel analyses involving both the standard regression approach and our new approach involving two stage least squares approach under the SNMM.

While this causal approach is not biased by the possible violation of the sequential ignorability assumption and the requirement of no effect of treatment on the effect modifier, it does make untestable assumptions about the absence of other interactions. Efficiency under the two stage least squares approach is enhanced by variation across baseline covariates in the effect of the randomized intervention on the post treatment variable M . Even when such variation is small, our simulations suggest that the two stage least squares estimation approach is less biased than the standard approach.

In analysis of the data, the two approaches (2SLS and standard regression (3)) give qualitatively similar results for both outcomes (ICS and SSIW). For SSIW, the point estimates based on the two methods are close, and the precision of the standard estimator is better; for ICS, the standard regression approach provides estimates (compared to 2SLS) that are attenuated, but qualitatively similar (with stronger negative effects of CT on outcome for larger values of depression severity).

The validity of the two stage least squares approach relies on a complex set of assumptions and conditions involving several different types of interactions among baseline covariates, the randomized intervention, and the post-treatment variable. In particular, the SNMM approach assumes the absence of certain structural interactions; i. e., it assumes that the effect of the main treatment may be modified by M but that neither the main effect of R nor the $R*M$ interaction are modified by baseline covariates X . Such assumptions may be relaxed, but are not fully testable (e. g., [34]) even in large samples. In small samples, the ability to examine additional interactions is further limited. In contrast to the absence of these structural interactions, the precision of the SNMM approach is dependent through its weights on the presence and strength of testable interactions between the baseline covariates and randomized intervention on the post-randomization factor as the dependent variable. We note further that, as is the case with pretreatment effect modifiers, that the presence of effect modification is dependent on the scale in which the outcome is modeled, and that rescaling the outcome may lead to different assessments about the magnitude and direction of such modification.

Given the smaller bias for the 2SLS SNMM method relative to the standard regression estimators for all effects under all simulation conditions, it or other variants (e. g., G-estimation) should at least be used in conjunction with the standard regression in assessing causal interaction effects. Discrepancy of results, as in the case of the estimates and tests of the randomized intervention–post-randomization factor interaction would clearly require close examination of assumptions under each method, which in itself may lead to important insights into the observed data.

There are several potentially fruitful directions for research. First, analyses such as ours may help identify subgroups in which the initially assigned treatment is not working. When the treatment initially assigned continues over an extended period, this might sometimes suggest that the initially assigned treatment be discontinued. In other settings, it might suggest additional or different therapies be adopted. Our approach for identifying subgroups defined by post-treatment variables in which treatment is failing is dependent on the modeling assumptions adopted, and so we view these analyses as at least somewhat exploratory in nature. In some settings, they could be used to suggest interventions to be tested in a within-person sequential randomized design for adaptive treatment regimes [1]. Under this approach, the intervention is decomposed into step-wise components such that the early stage effect modifiers are used to subgroup patients after the randomization of the first stage but before the next randomization of the second stage. The type of second stage interventions to be randomized would depend on the strata of the post-first stage effect modifier. Murphy (2005, 1) describes methods for identifying the optimal sequences of first and second stage intervention components based on this design as well as estimating individual effects within different combinations of treatments. The methods that we have developed are based on linear models. It would be valuable future work to extend these methods to binary and survival outcomes, which present problems for causal models because of the nonlinear link functions [34], to multiple post-randomization factors potentially corresponding to multiple stages of the intervention and finally to longitudinal outcomes.

Funding: This research was supported in part by grants from the National Institute of Health including R01MH078016 Causal Methods for Mediation and Interaction and T32MH065218 “Mental Health Biostatistics training grant.”

References

1. Murphy SA. An experimental design for the development of adaptive treatment strategies. *Stat Med* 2005;24:1455–1481.
2. Neyman J. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 1923;10:1–51.
3. Rubin D. Estimating causal effects of treatment in randomized and nonrandomized studies. *J Educ Psychol* 1974;66:688–701.
4. Emsley R, Dunn G, White IR. Mediation and moderation of treatment effects in randomized controlled trials of complex interventions. *Stat Methods Med Res* 2010;19:237–270.
5. Emsley R. How do treatments work and who for? Efficacy and mechanisms evaluation in complex interventions and personalized medicine. Symposium on Causal Mediation Analysis, 29 January, 2013, Gent, Belgium.
6. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry* 2002;59:877–883.
7. Kraemer HC, Kiernan M, Essex M, Kupfer DJ. How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychol* 2008;27(2):S101–S108.
8. Zhang R, Faerber J, Joffe M, Ten Have T Post-randomization interaction analyses in clinical trials with standard regression. Manuscript submitted for publication.
9. Almirall D, Ten Have T, Murphy SA. Structural nested mean models for assessing time-varying effect moderation. *Biometrics* 2010;66:131–139.
10. Normand SL. Evaluating the optimal timing of angiography: landmark or off the mark? *Circulation* 2007;116:2656–2657.
11. Joffe MM, Hoover DR, Jacobson LP, Kingsley L, Chmiel JS, Fischer BR, et al. Estimating the effect of Ziduvodine on Kaposi’s sarcoma from observational data using a rank preserving failure time model. *Stat Med* 1998;17:1073–1102.

12. Ten Have TR, Joffe MM, Lynch K, Maisto S, Brown G, Beck T. Causal mediation analyses with rank preserving models. *Biometrics* 2007;63:926–934.
13. Ilardi SS, Craighead WE. The role of nonspecific factors in cognitive behavioral therapy for depression. *Clin Psychol Sci Pract* 1994;1:138–156.
14. Tang TZ, DeRubeis RJ. Reconsidering rapid early response in cognitive behavioral therapy for depression. *Clin Psychol Sci Pract* 1999;6:283–288.
15. Haas E, Hill R, Lambert M, Morrell B. Do early responders to psychotherapy maintain treatment gains?. *J Clin Psychol* 2002;58:1157–1172.
16. Lambert MJ. Early response in psychotherapy: further evidence for the importance of common factors rather than placebo effects. *J Clin Psychol* 2005;61:855–869.
17. Young MA, Fogg LF, Scheftner W, Fawcett J, Akiskal H, Maser J. Stable trait components of hopelessness: baseline and sensitivity to depression. *J Abnorm Psychol* 1996;105:155–165.
18. Kuyken W. Cognitive therapy outcome: the effect of hopelessness in a naturalistic outcome study. *Behav Res Ther* 2004;42:631–646.
19. Feldman G. Cognitive and behavioral therapies for depression: overview new directions and practical recommendations for dissemination. *Psychiatr Clin N Am* 2007;30:39–50.
20. Murphy S, Lynch KG, Oslin D, McKay JR, TenHave T. Developing adaptive treatment strategies in substance abuse research. *Drug Alcohol Depend* 2007;88(Suppl 2):S24–S30.
21. Brown GK, Ten Have T, Henriques GR, Xie SX, Hollander JE, Beck AT. Cognitive therapy for the prevention of suicide attempts: A randomized controlled trial. *J Am Med Assoc* 2005;294:2847–2848.
22. Beck AT, Steer RA, Brown GK. Manual for Beck depression inventory-II. Texas: Psychological Corporation, 1996.
23. Beck AT, Brown GK, Steer RA. Psychometric characteristics of the scale for suicide ideation with psychiatric outpatients. *Behav Res Ther* 1997;35:1039–1046.
24. D'Zurilla TJ, Nezu AM, Maydeu-Olivares A. Social Problem-Solving Inventory-Revised (SPSI-R): technical manual. New York: Multi-Health Systems, 2002.
25. Small DS, Joffe MM, Lynch KG, Roy JA, Localio AR. Tom Ten Have's contributions to causal inference and biostatistics: review and future research directions. *Stat Med* 33:3421–3433. (correction on page 3600).
26. Rubin D. Statistics and causal inference: comment: which ifs have causal answers. *J Am Stat Assoc* 1986;81:961–962.
27. Wooldridge JM. Econometric analysis of cross section and panel data, 2nd ed. The MIT Press, 2010. Chapter 5.
28. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications Stat* 1994;23:2379–2412.
29. Robins JM, Rotnitzky A, Scharfstein D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D, editors. *Statistical models in epidemiology: the environment and clinical trials*, vol. IMA ume 116. NY: Springer-Verlag, 1999:1–92.
30. Rosenbaum PR. The consequence of adjustment for a concomitant variable that has been affected by the treatment. *J Royal Statistical Society Ser Gen* 1984;147:656–666.
31. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods- application to control of the healthy worker survivor effect. *Math Model* 1986;7:1393–1512.
32. Cole SR, Frangakis CE. The consistency statement in causal inference: A definition or an assumption?. *Epidemiology* 2009;20(1):3–5.
33. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009;20(6):880–883.
34. Robins J, Rotnitzky A. Estimation of treatment effects in randomized trials with non compliance and a dichotomous outcome using structural mean models. *Biometrika* 2004;91:763–783.