

Causal Mediation Analyses with Rank Preserving Models

Thomas R. Ten Have,^{1,*} Marshall M. Joffe,¹ Kevin G. Lynch,² Gregory K. Brown,²
Stephen A. Maisto,³ and Aaron T. Beck²

¹Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine,
Philadelphia, Pennsylvania 19104, U.S.A.

²Department of Psychiatry, University of Pennsylvania School of Medicine,
Philadelphia, Pennsylvania 19104, U.S.A.

³Department of Psychology and Center for Health and Behavior,
Syracuse University, Syracuse, New York 13244, U.S.A.

*email: ttenhave@cceb.upenn.edu

SUMMARY. We present a linear rank preserving model (RPM) approach for analyzing mediation of a randomized baseline intervention's effect on a univariate follow-up outcome. Unlike standard mediation analyses, our approach does not assume that the mediating factor is also randomly assigned to individuals in addition to the randomized baseline intervention (i.e., sequential ignorability), but does make several structural interaction assumptions that currently are untestable. The G-estimation procedure for the proposed RPM represents an extension of the work on direct effects of randomized intervention effects for survival outcomes by Robins and Greenland (1994, *Journal of the American Statistical Association* **89**, 737–749) and on intervention non-adherence by Ten Have et al. (2004, *Journal of the American Statistical Association* **99**, 8–16). Simulations show good estimation and confidence interval performance by the proposed RPM approach under unmeasured confounding relative to the standard mediation approach, but poor performance under departures from the structural interaction assumptions. The trade-off between these assumptions is evaluated in the context of two suicide/depression intervention studies.

KEY WORDS: Baseline randomization; Direct effects; G-estimation; Sequential ignorability; Structural mean model; Unmeasured confounding.

1. Introduction

We present a causal rank preserving model (RPM; e.g., Joffe et al., 1998) approach for investigating whether a randomized baseline intervention effect on a continuous outcome occurs through (indirect effect) or around a postrandomization intermediate factor (direct effect) in the context of randomized behavioral health trials (i.e., mediation analysis). Using the approach of potential outcomes (Neyman, 1923; Rubin, 1978), this work attempts to respond to the problem that direct effects are not nonparametrically identifiable from the data (Robins and Rotnitzky, 2005), as the number of parameters exceeds the number of identifying estimating equations. Model-based assumptions achieve such identifiability by increasing the number of identifying estimating equations. One such assumption that is commonly made for standard mediation analysis methods (Baron and Kenny, 1986) is the untestable no unmeasured confounding assumption for the intermediate factor. This assumption is equivalent to randomization of the baseline intervention and of subsequent intermediate variables (i.e., a strong form of “sequential ignorability”; e.g., Robins and Rotnitzky, 2005). Alternative approaches to increasing the number of identifying equations entail specifying baseline covariate relationships with the outcome (e.g.,

Vansteelandt and Goetghebeur, 2004) or using principal stratification (PS, e.g., Frangakis and Rubin, 2002; Mealli and Rubin, 2003). Under still another approach, we extend a weighted estimating equation method by Robins and Greenland (1994) for the survival context to the continuous outcome context under certain no-interaction assumptions. Accordingly, this article compares the proposed RPM approach to the standard approach in terms of a trade-off between these interaction assumptions and sequential ignorability. We use data from one of the behavioral intervention studies of interest in this article to compare inference under divergent assumptions.

The first study for which we use clinical insights to help compare the standard and proposed approaches is a suicide therapy study. It evaluated the effect of a specific type of psychotherapy (cognitive therapy) versus usual care in the treatment of suicide attempts, suicide ideation, hopelessness, and depression in 120 patients who had recently attempted suicide (Brown et al., 2005). The sample size for this investigation at 6 months is 101 because of drop-out, which appears to be weakly associated with the factors used in this analysis as well as others ($p > 0.35$; Brown et al., 2005). We assess if

the significant intent-to-treat effect of the cognitive therapy on 6-month depression outcome as measured by the Beck Depression Inventory-II (BDI) was due to a direct effect apart from the use of nonstudy therapy (mediator) between 4 and 6 months. Potential confounders of the mediator–outcome relationship include economic and personal stress reducing the motivation for nonstudy therapy and increasing the likelihood of depression in suicide attempters. Additionally, the potential for baseline depression to have modified significant cognitive therapy effects on follow-up depression needs to be addressed as a structural interaction under the proposed RPM approach.

The second study, a suicide prevention study, compared collaborative care management for treating depression (and thus reducing the risk of suicide) with usual care in 293 elderly depressed primary care patients (Bruce et al., 2004). The collaborative care management program in the intervention group was based on patient, primary care, and staff and physician interactions with a nurse-level behavioral health specialist (BHS). We evaluate if the significant intent-to-treat effect of the intervention on the 4-month Hamilton depression outcome was due to a direct effect apart from the use of prescribed antidepressant medication (mediator) between baseline and 4 months. Potential unmeasured confounders of the medication–depression relationship include medical comorbidities at follow-up, which deter elderly depressed patients from taking antidepressant medications because of so many other medications necessitated by their medical comorbidities, which also predispose patients to depression. As with the first study, potential baseline factors such as baseline depression and suicide ideation may have modified the significant effect of the BHS intervention and also the mediator, antidepressant medication, on the follow-up depression outcome.

The article now proceeds to Section 2 for notation, Section 3 for models, Section 4 for assumptions, Section 5 for estimation, Section 6 for the simulation results, Section 7 for the case study analyses, and Section 8 for the discussion.

2. Notation

We define the observed and potential variables for participant i . However, we suppress the index i to simplify the notation resulting from the addition of indices for the randomized intervention and mediators when defining the potential outcome variables.

For the observed variables, Y is the observed continuous outcome; R is the observed randomized zero-one variable; \mathbf{X} is the vector of observed baseline covariates other than randomization; and M is the observed mediation variable. Without loss of generality, we assume M is binary. The RPM approach and the corresponding G-estimation equations procedure that we present can accommodate continuous M in a straightforward way.

In contrast to the observed outcome variable, Y , we define Y_{rm} to be the outcome variable that would be observed if participant i were randomized to level r of the intervention and through some hypothetical mechanism were to receive or exhibit level m of the mediator. To establish a unique potential outcome, this approach assumes that all such hypothetical mechanisms lead to the same potential outcome. With r and m binary, we define four separate potential outcome vari-

ables: Y_{00} , Y_{10} , Y_{01} , and Y_{11} . With these four potential outcome variables, one can define the causal contrasts for the direct effect of the baseline intervention and the effect of the mediator on outcome. These effects are defined more formally in Section 3.1.

The indices of the potential outcome, which represent levels of the baseline intervention and mediator “set” or manipulated by those in control of these factors (e.g., investigators or clinicians) need to be distinguished from the observed levels of these factors for patient i . Given that the set levels of randomized baseline and mediators are denoted by r and m , respectively, in the definition of Y_{rm} , we denote the observed levels of R and M by \tilde{r} and \tilde{m} , respectively. To be consistent, we also denote the corresponding observed level of the baseline covariates, \mathbf{X} , as $\tilde{\mathbf{x}}$.

3. Models

In our context, a RPM may be used to model jointly the causal effects of the randomized baseline intervention and the mediator. We present such an RPM in Section 3.1 and the standard model in Section 3.2.

3.1 RPM

One specification of the RPM is

$$Y_{rm} = g(\tilde{\mathbf{x}}) + \theta_M m + \theta_R r + \epsilon, \quad (1)$$

for all participants regardless of what is actually observed in terms of R and M ; where $\theta_M = Y_{r1} - Y_{r0}$; $\theta_R = Y_{1m} - Y_{0m}$; $g(\cdot)$ is an unspecified function; and ϵ is a mean zero (given R and \mathbf{X}) error term with unspecified common distribution with finite variance. Here, θ_M represents the effect of the mediator on the outcome holding the baseline intervention fixed at any level r ; and θ_R represents the direct effect of the randomized intervention on the outcome, holding the mediator fixed at any level m .

The error term ϵ is not a function of m or r ; this results in a RPM model for intervention effects, where the effects of both R and M are the same for all subjects. This rank-preserving assumption allows simpler presentation of our estimators than analogous more general and less restrictive structural distribution models (SDMs; e.g., Robins et al., 1992) but imposes no additional restrictions on the observable data. The assumption of a common error distribution across covariate levels imposes some restrictions on the observable data and makes this model also more restrictive than a structural mean model (SMM; e.g., Vansteelandt and Goetghebeur, 2004) with an analogous form. Nonetheless, the estimators discussed below apply to SMMs as well.

3.2 Standard Regression Model

For comparison with the RPM in (1), we present the corresponding standard linear regression model as presented by a number of authors (e.g., Baron and Kenny, 1986). This standard linear regression model is defined as

$$Y = \beta_S \tilde{\mathbf{x}} + \theta_{MS} \tilde{m} + \theta_{RS} \tilde{r} + \epsilon_S, \quad (2)$$

for all participants, and where $\theta_{RS} = E(Y | R = 1, M = \tilde{m}, \mathbf{X} = \tilde{\mathbf{x}}) - E(Y | R = 0, M = \tilde{m}, \mathbf{X} = \tilde{\mathbf{x}})$; $\theta_{MS} = E(Y | R = \tilde{r}, M = 1, \mathbf{X} = \tilde{\mathbf{x}}) - E(Y | R = \tilde{r}, M = 0, \mathbf{X} = \tilde{\mathbf{x}})$; β_S is a vector of effects for baseline covariate values $\tilde{\mathbf{x}}$; and ϵ_S is a mean zero error term with a normal distribution and variance equal

to σ_S^2 . The parameters θ_{RS} and θ_{MS} are defined as comparisons of observed outcome expectations from different sample subgroups defined by \tilde{r} and \tilde{m} but not as causal contrasts of expectations under different conditions defined by r and m for the same individual. The comparisons of such subgroups will only equal the causal contrasts for an individual under certain conditions listed below for the standard approach (e.g., sequential ignorability).

4. Model Assumptions

We now address the assumptions of each of the RPM and standard approaches in Sections 4.1 and 4.2, respectively. Of special interest will be sequential ignorability for the standard approach and the structural interaction assumptions under the RPM approach.

4.1 RPM Assumptions

For the RPM model, the assumptions necessary for unbiased inference are: (1) Stable Unit Treatment Value Assumption (SUTVA); (2) randomization (i.e., ignorability) of baseline intervention assignment; (3) independence of observations for standard error estimation; and (4) model assumptions including no-interaction assumptions among baseline covariates, the baseline randomized intervention, and the mediator.

SUTVA consists of two sub-assumptions. First, there is a single value for each of the potential random outcome variables (Y_{rm}) for a given patient i regardless of the randomization assignment or mediation behavior of any other patient i' . Notationally, this assumption implies that Y_{rm} is defined with scalar indices for a given participant i , rather than vectors of indices representing baseline intervention assignments and mediator levels of all patients.

Second, there is a single value for each of the potential outcome random variables (Y_{rm}) for a given patient i regardless of the method of administration of the randomized baseline intervention or the administration or occurrence of the mediator, such that for patient i with observed levels \tilde{r} and \tilde{m} for R and M , respectively, $Y = \tilde{r}\tilde{m}Y_{\tilde{r}\tilde{m}} + (1 - \tilde{r})\tilde{m}Y_{1-\tilde{r}\tilde{m}} + \tilde{r}(1 - \tilde{m})Y_{\tilde{r}1-\tilde{m}} + (1 - \tilde{r})(1 - \tilde{m})Y_{1-\tilde{r}1-\tilde{m}}$. Such an identity only holds for binary r and m , but extends in a straightforward way to continuous m .

The randomization assumption for the RPM in (1) implies stochastic independence between the randomized baseline intervention, R , and potential outcomes. Stochastically, this means for the potential outcomes: $\Pr(Y_{1,1}, Y_{1,0}, Y_{0,1}, Y_{0,0} | R = \tilde{r}, \mathbf{X} = \tilde{\mathbf{x}}) = \Pr(Y_{1,1}, Y_{1,0}, Y_{0,1}, Y_{0,0} | \mathbf{X} = \tilde{\mathbf{x}})$. Such an assumption implies no imbalance between randomization groups with respect to unmeasured confounders, i.e., no unmeasured confounding. We note that for the suicide prevention study, primary care practices were randomized. However, because the within-practice design effect is so small for the outcome, the Hamilton depression scale, we ignore the clustering due to primary care practice (Bruce et al., 2004).

Additional assumptions for the RPM in (1) include the additive structure of the baseline intervention and mediating factors and the following structural no-interaction assumptions. First, the causal effects of treatment and intermediate factors are assumed to be the same for all subgroups of patients defined by baseline covariates (i.e., no $\mathbf{X} * R$ and $\mathbf{X} * M$ interactions on $Y_{r,m}$). Second, we assume the absence

of a structural interaction between the baseline intervention and intermediate factors ($R * M$) on Y_{rm} . These structural no-interaction assumptions are not fully testable, but future research will focus on assessing and relaxing these assumptions under the RPM without the sequential ignorability assumption.

An additional but fully testable assumption is needed for obtaining weights that are sufficiently noncollinear for achieving unique identifying estimating equations. Such an assumption requires a significant intent-to-treat effect on the mediator that is modified by baseline covariates (i.e., $\mathbf{X} * R$ on M).

Finally, the consistency of the proposed estimators of θ_M and θ_R does not rely on the correct specification of $g(\tilde{\mathbf{x}})$ or the distribution of ϵ in (1). However, efficiency depends on how well $g(\tilde{\mathbf{x}})$ approximates the true relationship between \mathbf{X} and Y_{rm} (e.g., Fischer-Lapp and Goetghebeur, 1999). In contrast, the standard regression approach does rely on correct specification of the relationship between the outcome and baseline covariates, which is testable under sequential ignorability.

4.2 Standard Regression Assumptions

The standard regression model assumptions are: (1) sequential ignorability of both the baseline intervention and mediator given baseline covariates; (2) independence among participants; and (3) model assumptions including the correct form of $g(\tilde{\mathbf{x}})$ and a no-interaction assumption. We focus on the sequential ignorability and no-interaction assumptions. While the RPM in (1) requires ignorability of R under randomization of the baseline intervention assignment, the standard regression model in (2) requires ignorability for both the baseline intervention and mediator. The sequential ignorability assumption implies stochastic independence of these two factors relative to the potential outcomes, conditional on baseline covariates. Stochastically, this implies $\Pr(Y_{1,1}, Y_{1,0}, Y_{0,1}, Y_{0,0} | R = \tilde{r}, M = \tilde{m}, \mathbf{X} = \mathbf{x}) = \Pr(Y_{1,1}, Y_{1,0}, Y_{0,1}, Y_{0,0} | \mathbf{X} = \tilde{\mathbf{x}})$. The no confounding assumption for the mediator that is made in the literature on standard mediation methods (e.g., Baron and Kenny, 1986) requires the above identity. Although there is no interaction in the standard model in (2), any of the candidate interactions, $\mathbf{X} * R$, $\mathbf{X} * M$, and $R * M$ can be identified and estimated under the above assumptions, especially the sequential ignorability assumption for both the baseline intervention and mediator.

5. Estimation

Under the assumptions in Section 4.1 and RPM in (1), consistent estimators of θ_R and θ_M can be obtained by solving the following weighted G-estimation equations for θ_M , θ_R , and parameters of $g(\mathbf{X})$. To obtain these equations based on observed data, we first relate the observed and potential outcome variables as follows with the potential outcome indices in (1) equal to the corresponding observed outcome indices, $m = \tilde{m}$ and $r = \tilde{r}$: $Y = Y_{0,0} + \tilde{m}\theta_M + \tilde{r}\theta_R$. Based on this relation, we obtain a candidate value for $Y_{0,0}$ for each combination of \tilde{m} and \tilde{r} : $Y_{0,0}(\theta^*) = Y - \tilde{m}\theta_M^* - \tilde{r}\theta_R^*$ where $\theta^{*T} = (\theta_M^* \theta_R^*)$ and the elements of which are putative or candidate values for θ_M and θ_R . When $\theta_R^* = \theta_R$ and $\theta_M^* = \theta_M$ under the RPM and $\hat{\beta}$ is some estimate of β under the working specification $g(\mathbf{X}) = \beta^T \mathbf{X}, Y_{0,0}(\theta^*) - \hat{\beta}^T \tilde{\mathbf{x}}$ and the

randomized baseline intervention, R , are uncorrelated. Hence, we can obtain consistent estimators of θ_M and θ_R by iteratively solving the following unbiased estimating equation using a Newton–Raphson routine,

$$\sum (R - q) \mathbf{W}(\tilde{\mathbf{x}}) (Y_{0,0}(\boldsymbol{\theta}) - \hat{\boldsymbol{\beta}}^T \tilde{\mathbf{x}}) = 0, \quad (3)$$

where $\boldsymbol{\theta}^T = (\theta_M \theta_R)$; $q \equiv \Pr(R = 1)$, the proportion randomized to the baseline intervention; $\hat{\boldsymbol{\beta}}$ is obtained from a linear regression of $Y_{0,0}(\hat{\boldsymbol{\theta}})$ on \mathbf{X} given an estimate of $\boldsymbol{\theta}$ from the previous iteration; and $\mathbf{W}(\tilde{\mathbf{x}})$ is a weight vector function of the observed elements of \mathbf{X} .

Assuming that they are not perfectly collinear, each element of $\mathbf{W}(\tilde{\mathbf{x}})$ corresponds to a separate identifying equation for the corresponding structural parameter under certain assumptions. These weight elements can be derived from efficient score criteria in Robins et al. (1992) under the linear structural distribution model given additional assumptions such as sequential ignorability and normal errors with constant variance. These additional assumptions impact efficiency but not the consistency of the resulting estimators. Accordingly, $\mathbf{W}(\tilde{\mathbf{x}})^T = [1 \ \eta(\tilde{\mathbf{x}})]$ with the two elements corresponding to θ_R and θ_M , respectively. In the context of intervention nonadherence, the element corresponding to θ_M is the “compliance score,” $\eta(\tilde{\mathbf{x}}) = \Pr(M = 1 | R = 1, \mathbf{X} = \tilde{\mathbf{x}}) - \Pr(M = 1 | R = 0, \mathbf{X} = \tilde{\mathbf{x}})$ (e.g., Ten Have et al., 2004). To preclude collinearity with the element of “1” in $\mathbf{W}(\tilde{\mathbf{x}})$, variation across covariates \mathbf{X} is needed for this score, which is a measure of the interaction between baseline covariates and the randomized intervention factor with the mediator as the dependent variable. In estimating the effect of the mediator on outcome (θ_M), $\eta(\tilde{\mathbf{x}})$ upweights participants characterized by $\mathbf{X} = \tilde{\mathbf{x}}$ for whom the randomized intervention effect on the mediator is large, thus contributing information to estimating the path through the mediator to outcome. For this article, estimation of $\Pr(M | R, \mathbf{X})$ was based on the logistic model. The consistency of the resulting estimator of $\boldsymbol{\theta}$ is not impacted by the form of either $\mathbf{W}(\tilde{\mathbf{x}})$ or $Y_{0,0}(\boldsymbol{\theta}) - \hat{\boldsymbol{\beta}}^T \tilde{\mathbf{x}}$, although the efficiency of $\hat{\boldsymbol{\theta}}$ is.

The variance–covariance for $\hat{\boldsymbol{\theta}}$ is estimated after convergence of the G-estimation algorithm with a sandwich estimator based on (3) as follows: $\text{VarCov}(\hat{\boldsymbol{\theta}}) = \mathbf{D}^{-1} \mathbf{H}^{-1} \mathbf{D}^{-1T}$, where \mathbf{D} is a symmetric 2×2 matrix: $\mathbf{D} = \sum \frac{\partial \mathbf{S}}{\partial \boldsymbol{\theta}}$; \mathbf{S} is a 2×1 column vector for patient i : $\mathbf{S} = (R - q) (Y_{0,0}(\boldsymbol{\theta}) - \boldsymbol{\beta}^T \tilde{\mathbf{x}}) \mathbf{W}(\tilde{\mathbf{x}})$; and \mathbf{H} is a 2×2 submatrix corresponding to the $\boldsymbol{\theta}$ elements in the $(2 + k) \times (2 + k)$ matrix: $\sum \mathbf{S}^* \mathbf{S}^{*T}$ with \mathbf{S}^* equal to \mathbf{S} but augmented with score functions for $g(\mathbf{X})$ in general and $\boldsymbol{\beta}^T \tilde{\mathbf{x}}$ in particular. The resulting estimate of $\text{VarCov}(\hat{\boldsymbol{\theta}})$, evaluated at $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$, is used in Wald statistics for hypothesis testing and Wald confidence intervals for θ_M and θ_R . We assume that $q \equiv \Pr(R = 1)$ is fixed by design but estimation of q may improve efficiency (Robins et al., 1992; Fischer-Lapp and Goetghebeur, 1999).

6. Simulations

Under different combinations of sequential ignorability and structural no-interaction assumptions, we now present simulation results for the effects of the randomized baseline intervention and mediation factors given the conditions of the two example trials. Each data set for each set of simulations

was based on the corresponding characteristics of the respective example data set and fitted RPMs: (1) sample size of the data set (293 for the suicide prevention study and 101 for the suicide therapy study); (2) observed values of \mathbf{X} and R for each subject in each study; and (3) study-based estimates of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\eta(\tilde{\mathbf{x}})$. Given these specifications, we simulated Y_{rm} and M .

First, in Table 1 under no sequential ignorability but still under the assumption of no structural $\mathbf{X} * R$, $\mathbf{X} * M$, and $R * M$ interactions for Y_{rm} , we specified that the error term for Y_{rm} in (1), ϵ , was decomposed into two components, one of which was related to a model for M . That is, we specified the following shared parameter framework: $\epsilon = \epsilon^* + \nu_Y \gamma$ and $\Pr(M = 1 | R = \tilde{r}, \mathbf{X} = \tilde{\mathbf{x}}) = \text{expit}(\boldsymbol{\delta}^T \tilde{\mathbf{x}} + \alpha \tilde{r} + \nu_M \gamma)$, where γ is a normal random variable with mean zero and variance equal to one; $\nu_Y = \nu_M = 1$; ϵ^* is a normal random variable with mean zero and variance equal to the variance of the observed outcome variable in the particular example data set; and $\boldsymbol{\delta}$ and α are specified to be equal to the corresponding naive estimates from the logistic regression of M on \mathbf{X} and R interaction without consideration of γ . The values of \mathbf{X} used for the simulations were obtained from the samples of the two studies, such that on Y_{rm} , $\eta(\tilde{\mathbf{x}})$ varied by more than 0.80 on the probability difference scale across $\mathbf{X} = \tilde{\mathbf{x}}$.

Second, Table 2 presents an assessment of the sensitivity of the RPM approach to departures from the assumptions of no $\mathbf{X} * R$ interaction on Y_{rm} . Specifically, we included an $X_k * R$ interaction term in the RPM based on (1) in an additional simulation. Finally, in Table 3 we present a simulation under both sequential ignorability and no structural $\mathbf{X} * R$, $\mathbf{X} * M$, and $R * M$ interactions for Y_{rm} .

For each simulation specification, we simulated 1000 sets of data for Y and M . From the corresponding 1000 sets of fitted RPMs under (1), we computed for θ_R and θ_M , the absolute and relative bias of the RPM estimate, the mean squared error (MSE), and confidence interval coverage (a proportion of iterations for which the 95% confidence interval included the true value of θ_R or θ_M). We computed the same simulations statistics for the standard regression model in (2). The results of the simulations for the suicide prevention and therapy studies are summarized below.

Table 1 shows that under the absence of sequential ignorability but under no structural $\mathbf{X} * R$, $\mathbf{X} * M$, and $R * M$ interactions for Y_{rm} , the RPM approach yields smaller bias and more accurate confidence intervals than the standard regression procedure, but at the expense of larger MSE due to greater variability. This result is consistent for MSE and bias between the two example-based simulations and between θ_R and θ_M . There is somewhat less consistency for the improvement in 95% confidence interval coverage under the RPM approach relative to the standard regression procedure. Also, the MSE tends to be larger for RPM estimators of θ_M than of θ_R .

The simulation results in Table 1 for the RPM approach under the suicide prevention study conditions are somewhat better than the analogous results for the suicide therapy study conditions with the smaller sample size. Based on additional simulations, this difference in results between study conditions appears to be partly attributable to differences in sample size and the magnitude of the $\mathbf{X} * R$ interaction on

Table 1

*Simulation results based on the suicide prevention (“prevention”) study ($N = 293$; $\theta_M = -1.43$ and $\theta_R = -2.58$) and on the suicide therapy (“therapy”) study ($N = 101$; $\theta_M = 14.59$ and $\theta_R = -3.93$) without sequential ignorability but assuming no $R * M$, $X * R$, or $X * M$ structural interactions for Y_{rm}*

Suicide study	Method	Simulation statistic	Mediation effect (θ_M)	Direct effect (θ_R)
Prevention	Standard	Bias (%)	3.68 (258%)	0.79 (31%)
	RPM		0.02 (1%)	0.00 (0%)
Therapy	Standard	% Coverage	-4.05 (-28%)	-0.41 (-10%)
	RPM		0.40 (3%)	-0.18 (5%)
Prevention	Standard	MSE	2%	84%
	RPM		99%	95%
Therapy	Standard	MSE	73%	95%
	RPM		90%	97%
Prevention	Standard	MSE	14.37	1.46
	RPM		23.43	1.92
Therapy	Standard	MSE	25.11	7.10
	RPM		326.95	11.60

Table 2

*Simulation results based on the suicide prevention (“prevention”) study ($N = 293$; $\theta_M = -1.43$ and $\theta_R = -2.58$) and on the suicide therapy (“therapy”) study ($N = 101$; $\theta_M = 14.59$ and $\theta_R = -3.93$) without sequential ignorability and with a $X_1 * R$ (10% of θ_R) structural interaction for Y_{rm}*

Suicide study	Method	Simulation statistic	Mediation effect (θ_M)	Direct effect (θ_R)
Prevention	Standard	Bias (%)	3.65 (255%)	-3.88 (-150%)
	RPM		-1.14 (-80%)	-4.91 (-190%)
Therapy	Standard	% Coverage	-3.26 (-23%)	-12.99 (-333%)
	RPM		12.35 (864%)	-11.67 (-299%)
Prevention	Standard	MSE	0%	0%
	RPM		99%	7%
Therapy	Standard	MSE	82%	0%
	RPM		86%	21%
Prevention	Standard	MSE	14.16	15.85
	RPM		24.88	26.11
Therapy	Standard	MSE	19.70	175.75
	RPM		482.99	156.50

M . First, the bias of $\hat{\theta}_R$ appears to be impacted slightly by the magnitude of the $X * R$ interaction on M , although the variability of $\hat{\theta}_R$ is increased substantially, as are the bias and variability of $\hat{\theta}_M$. Second, halving the sample size for each study has a differential impact. For the RPM estimators under the suicide therapy study conditions, halving the overall sample size to 50 did not adversely impact bias, but did increase MSE, especially for θ_R and decreased the coverage for θ_M . Under the suicide prevention study conditions, halving the overall sample size to 150 adversely impacted bias and MSE but not the confidence interval coverage for the RPM estimators of θ_R and θ_M .

Additionally, Table 2 assesses the RPM approach relative to the standard regression method without sequential ignorability and with a structural $X_k * R$ interaction on Y_{rm} . Table 2 does not show the Table 1 superiority of the RPM approach over the standard regression method for bias and confidence interval coverage. Specifically, the RPM and standard regres-

sion estimators of the direct effect of the baseline intervention exhibit similarly severe bias and lack of confidence interval coverage for both study contexts. Nonetheless, the RPM approach does provide consistently better coverage than the zero coverage of the standard approach.

Table 3 presents very different simulation results under sequential ignorability and without any structural interactions. The RPM and standard regression approaches perform similarly well for bias and confidence interval coverage, but with the RPM approach exhibiting larger MSEs and somewhat worse confidence interval coverage.

In summary, the simulations show the trade-off between the sequential randomization and structural no-interaction assumptions in terms of comparisons between the RPM and standard approaches. For bias and coverage but not MSE, the RPM approach performs better than the standard approach under the structural no-interaction setting. However, they both perform poorly when the structural

Table 3

Simulation results based on the suicide prevention (“prevention”) study ($N = 293$; $\theta_M = -1.43$ and $\theta_R = -2.58$) and on the suicide therapy (“therapy”) study ($N = 101$; $\theta_M = 14.59$ and $\theta_R = -3.93$) with sequential ignorability and without any structural interactions (e.g., $X_1 * R$) for Y_{rm}

Suicide study	Method	Simulation statistic	Mediation effect (θ_M)	Direct effect (θ_R)
Prevention	Standard	Bias (%)	0.03 (2%)	0.02 (1%)
	RPM		0.08 (6%)	0.02 (1%)
Therapy	Standard	% Coverage	-0.10 (-1%)	-0.10 (-2%)
	RPM		-0.13 (-1%)	-0.12 (-3%)
Prevention	Standard	MSE	96%	95%
	RPM		99%	94%
Therapy	Standard	MSE	95%	95%
	RPM		90%	96%
Prevention	Standard	MSE	0.73	0.73
	RPM		20.06	1.64
Therapy	Standard	MSE	8.51	6.76
	RPM		308.87	16.28

no-interaction assumption is relaxed, although the standard approach may yield worse confidence interval coverage. Finally, when both sets of assumptions hold, the two models perform well, although the standard approach performs somewhat better in terms of confidence interval coverage. Overall, the standard approach yields better MSE than the proposed RPM procedure.

7. Data Analyses

The descriptive statistics in Table 4 suggest similarities between the two examples in terms of the intent-to-treat (ITT) comparisons of outcome but not in terms of the ITT comparison of the mediator factor. First, the ITT contrasts for outcome and mediator are significant in both studies. Hence, an analysis of the mediation of these significant ITT effects is justified. Second, Table 4 also indicates differences between the two examples in terms of the level of use of the mediator factor by patients and also the sign of the ITT effect on the mediator factors. Most of the depressed patients in the suicide prevention study used medication regardless of whether they were in the BHS arm or not. In contrast, in the suicide therapy study, fewer of the suicidal patients used nonstudy therapy in either arm, although a higher proportion of the usual care group used nonstudy therapy than the randomized study therapy group. Given the differences between the two examples with respect to the mediator results in Table 4, we now compare the RPM and standard regression results in Table 5.

7.1 Suicide Prevention Study

The RPM and standard regression estimates for the suicide prevention study in Table 5 are in agreement in estimating a statistically significant direct effect of the BHS intervention on the 4-month Hamilton outcome apart from increasing antidepressant use among the depressed patients. The estimated direct effect of this intervention under both the RPM and standard regression approaches is an approximate reduction of 2.5 Hamilton units. However, the RPM confidence interval is wider than the standard regression confidence

Table 4

For the suicide prevention (“prevention”) and therapy (“therapy”) studies, means (standard deviations in parentheses) and proportions for the Hamilton or BDI depression outcomes, respectively, and proportion of patients taking antidepressant medication or nonstudy therapy, respectively, by randomized intervention group or by whether they took antidepressant medication or nonstudy therapy

Suicide study	Group	Hamilton	Medication
Prevention	Usual care	13.55 (8.35)	0.45
	Intervention	11.50 (7.38)	0.85
	No medication	13.14 (8.09)	
	Medication	12.23 (12.23)	
		BDI	Nonstudy therapy
Therapy	Usual care	19.33 (12.07)	0.25
	Study therapy	14.02 (14.77)	0.08
	No nonstudy therapy	17.08 (14.78)	
	Nonstudy therapy	15.11 (12.07)	

intervals, as one would expect from the MSE results in the simulations. The significant direct effect of the presence of BHS on reducing depression could be the result of the impact of this specialist on the staff and physicians of the practices. That is, one would expect that the presence of the BHS in the intervention practices raised the sensitivity of the staff and providers in treating depression. We also see that both the RPM and standard regression approaches indicate a nonsignificant effect of the mediator (antidepressant use) on outcome.

Estimating the direct effect of the BHS intervention under the RPM approach required covariates that interact with the significant randomized intervention factor on the mediator, i.e, varying the compliance score-based weight element, $\eta(\tilde{x})$. One strategy for identifying such predictors is to perform

Table 5

For the suicide prevention (“prevention”) and therapy (“therapy”) studies, ITT, standard regression, and RPM estimates are presented for the direct effects of the randomized baseline intervention (BHS or study cognitive therapy) and the mediator (antidepressant medication or nonstudy therapy). Standard errors and nominal 95% confidence intervals are in parentheses.

Suicide study	Method	Direct effect	Mediator effect
Prevention	ITT	-3.12 (0.82) (-4.72, -1.51)	
	Standard	-2.67 (0.89) (-4.41, -0.93)	-1.19 (0.94) (-3.03, 0.65)
	RPM	-2.58 (1.27) (-5.07, -0.10)	-1.43 (2.34) (-6.01, 3.15)
Therapy	ITT	-6.35 (2.53) (-11.37, -1.33)	
	Standard	-6.86 (2.60) (-12.01, -1.70)	-3.05 (3.46) (-9.92, 3.82)
	RPM	-3.93 (3.09) (-9.98, 2.12)	14.59 (15.87) (-16.52, 45.69)

logistic regression of medication use on baseline covariates stratified by the randomization arm. For the group not randomized to the BHS, we did not find any significant predictors of taking medication ($p > 0.50$), except for baseline antidepressant medication status ($p = 0.03$). For the group randomized to the BHS, site, past medication history and baseline medication status are strongly predictive of the mediator medication factor ($p < 0.001$). Comparing these predictive relationships between the two randomization arms, the test of the overall $\mathbf{X} * R$ interaction on M yielded a p -value of 0.006. Accordingly, the distribution of the estimated compliance scores based on these predictive factors, $\hat{\eta}(\bar{\mathbf{x}})$, appears to be sufficient, as evidenced by the range of scores (-0.08 to 0.72) and quartiles (-0.06, 0.55, and 0.70).

7.2 Suicide Prevention Study

In contrast to the suicide prevention study, the RPM and standard regression estimates for the suicide therapy study in Table 5 are not in agreement, indicating possible unmeasured confounding of the standard regression results and/or a violation of the no $M * R$, $\mathbf{X} * R$, and $\mathbf{X} * M$ interactions assumption for Y_m . Specifically, for the suicide therapy study, the estimate of θ_R under the RPM is smaller than the standard regression estimate of θ_{RS} . Hence, under the standard approach there is a significant direct effect of the study therapy on the 6-month depression outcome, apart from any impact on this outcome through the use of nonstudy therapy, whereas the RPM approach indicates that there is not sufficient evidence for such inference. There are three alternative explanations for this discrepancy in the direct effect estimates between the RPM and standard approaches: (1) confounding of the nonstudy therapy versus depression outcome relationship; (2) effect modification of the nonstudy therapy mediator on an outcome by study cognitive therapy; and (3) modification of the effect of baseline cognitive therapy on an outcome by baseline depression or suicide ideation.

First, the smaller direct effect of the study therapy intervention on 6-month depression under the RPM approach relative to the standard approach may reflect potential confounding of the effect of nonstudy therapy on depression. While not significant under either model, the conditional association between nonstudy therapy and depression under the standard approach is negative (helps reduce depression) while the corresponding effect under the RPM approach is positive (helps increase depression). It is possible that the negative association under the standard approach may be due to confounding by environmental stresses (family and or financial) that reduced the likelihood of use of nonstudy therapy and also increased depression. However, when potentially controlling for this unmeasured stress confounder under the RPM, nonstudy therapy increases depression because of the ineffectiveness of these therapies in dealing with suicidal thoughts (Brown et al., 2005). Not being vulnerable to such unmeasured confounding, the RPM approach suggests that some of the intent-to-treat effect of the baseline therapy intervention occurs by reducing the reliance on nonstudy therapy and thus reducing depression.

Second, the reduced direct effect of study therapy under the proposed RPM may reflect that the study therapy enhanced the effectiveness of the nonstudy therapy on depression (effect modification based on $R * M$). For example, it is possible that patients learned to utilize information and problem-solving skills obtained from the study therapy and applied them to the other therapy that they had received. The clinical investigators for this study suggested that this scenario was likely.

Third, the difference in inference between the RPM and standard regression approaches could also reflect departures from the assumption of no baseline covariate effect modification of the direct effect of the cognitive therapy intervention ($R * \mathbf{X}$). The cognitive therapists may have provided more intensive therapy for patients with more suicide ideation or depression at baseline, which would have resulted in such interaction. Nonetheless, the investigators believed that the cognitive therapy approach is standardized enough that such an interaction was not very likely. A similar scenario may have existed with baseline depression severity impacting the way external therapists provide nonstudy therapy ($M * \mathbf{X}$). However, this interaction may also be less likely given that nonstudy therapists mostly saw patients after the study started and may not have been aware of their respective patients’ baseline status.

The above subject matter discussion of possible violations of the sequential ignorability and structural no-interaction assumptions has provided information to weigh these two sets of assumptions against each other. The study investigators believe that the unmeasured stress-based source of confounding violating sequential ignorability was as likely as the possibility of effect modification of the nonstudy therapy effect on depression outcome by the baseline cognitive intervention. Hence, clinical information and statistical evidence suggests that departures from sequential ignorability and/or departures from the assumption of no $R * M$ interaction on Y_m may be leading to differences between the standard and RPM approaches with respect to the direct effect of the baseline cognitive intervention.

Inferentially, the RPM and standard approaches also disagree with respect to the sign of the effect of nonstudy therapy on the depression outcome, although both approaches yielded confidence intervals surrounding one. Moreover, the RPM-based estimate of θ_M and corresponding standard error are much larger in magnitude than the analogous standard regression estimates. This result conforms to the large simulation-based MSE for θ_M in Table 1. Nonetheless, Table 1 indicates that such variability in the θ_M estimate does not preclude more accurate inference of the G-estimation estimate of θ_R under the structural no-interaction assumption.

In assessing the effectiveness of the compliance score-based weight element, $\eta(\tilde{\mathbf{x}})$, for G-estimation, we again evaluate the predictors of the mediator, taking nonstudy therapy, stratified by randomization arms. For the group not randomized to the study therapy, we did not find any significant predictors of nonstudy therapy before 6 months ($p > 0.45$), except for the positive association with baseline suicide ideation status ($p = 0.03$). For the group randomized to the study therapy, we did not find any significant predictors ($p > 0.30$) of nonstudy therapy use before 6 months. The corresponding test of the overall $\mathbf{X} * R$ interaction on M yielded a p -value of 0.59, which is much less significant than the p -value of 0.006 for the larger suicide intervention study. Nonetheless, the suicide therapy study appeared to have a wider range of estimated compliance scores (-0.99 to 0.21) than did the suicide prevention study (-0.72 to 0.08). The spread of the quartiles for the suicide therapy study compliance scores (-0.20 , -0.10 , and 0.01) indicates some skewness but with higher mass toward zero.

8. Discussion

We have proposed a new approach to analyzing direct effects of randomized baseline interventions in the presence of a post-randomization mediation factor. This approach is based on a linear model extension of a weighted test-based approach by Robins and Greenland (1994) for testing direct and mediator effects with respect to survival outcomes. The multielement weight vector with separate weights for each parameter leads to separate identifying equations for each parameter. The absence of perfect collinearity among the weight elements ensures that the identifying equations achieve a full rank-identifying matrix. A similar approach was implemented by Ten Have et al. (2004) but in a different context, that of intervention noncompliance with intervention-received as the postrandomization factor. In contrast to Ten Have et al. (2004), who investigated the estimation of the intermediate factor (adherence to randomized intervention), we focus this approach on estimating the direct effect of the randomization factor when the postrandomization factor is a mediator.

In this article, we relate the proposed RPM approach to the standard regression mediation approach through simulations and analyses of data from two behavioral intervention trials. These two approaches require different sets of potentially untestable assumptions, which lead to tradeoffs. In practice, these tradeoffs of assumptions require assessment using clinical or subject matter information.

In addition to the bias versus variability trade-off shown in the simulations and example, the RPM approach exchanges the untestable sequential ignorability assumption for not fully

testable no-interaction assumptions among baseline covariates and the baseline intervention and mediator. In both of our studies, there was clinical conjecture about potential unmeasured confounders that would violate the sequential ignorability assumption. However, there is also clinical weight given to interactions between baseline study interventions and follow-up nonstudy therapies on the follow-up depression outcome. Balancing these assumptions is a clinical judgment.

Future research will focus on assessing the structural $R * M$, $\mathbf{X} * R$, or $\mathbf{X} * M$ interactions under (1). An additional element involving \mathbf{X} will be added to $\mathbf{W}(\tilde{\mathbf{x}})$ for each additional structural interaction parameter based on the criteria of Robins et al. (1992). The difficulty of testing these structural interactions arises because \mathbf{X} would be required to satisfy several strong constraints. For example, for $R * M$, \mathbf{X} (e.g., baseline depression) would need to satisfy two conditions: (1) \mathbf{X} leads to strong interaction with R on M (i.e., variation in compliance score across $\tilde{\mathbf{x}}$); and (2) $\Pr(M = 1 | R = 1, \mathbf{X})$ is not perfectly collinear with the compliance score. For assessing $R * \mathbf{X}$, condition (2) would need to be that \mathbf{X} itself is not perfectly collinear with the compliance score. Our future research will focus on determining such baseline covariates satisfying these conditions for either of the two example studies. While the above weights yield consistent estimators under departures from sequential ignorability, they are not efficient under these departures. Additional future research will develop weights leading to consistent estimators that are also efficient under departures from sequential ignorability.

An alternative to the proposed RPM approach is principal stratification (Frangakis and Rubin 2002; Mealli and Rubin, 2003), which is based on estimating baseline intervention effects within latent strata defined by potential mediator outcomes in the different randomization arms. This approach allows an alternate definition of the direct effect as the effect of the baseline intervention for certain subgroups for which baseline intervention does not affect the mediator (Mealli and Rubin, 2003). This PS method does not attempt to parse out the separate effects of the baseline intervention and mediator factors, which is the goal of the standard mediation approaches (Joffe, Small, and Hsu, 2007). Hence, to provide a parallel way to do this parsing while relaxing sequential ignorability but under not fully testable no-interaction assumptions, we have proposed the RPM approach.

Ten Have et al. (2004) and at a conceptual level Joffe et al. (2007) expressed the PS direct effect parameters in terms of RPM parameters but with a focus on the effect of adherence as the mediator. Without a direct effect of the randomized baseline intervention, the equivalence of the mediation effects between PS and RPM approaches was empirically supported. However, with a direct effect, the mediation effects were very different under the two approaches. Future research will focus on whether the direct effects also differ between the two approaches and on the nature of this relationship under the no-interaction assumptions.

Finally, a reviewer notes that the principal stratification approach to mediation does not require the structural no-interaction assumption between R and M , although it still requires no-interaction between the effect of R and baseline covariates \mathbf{X} within the principal strata. The separate direct effects would be identified and estimated consistently within

each principal strata for which the potential mediator M_r is constant for both levels of r (i.e., never-takers and always-takers in the adherence context). The required assumptions are that the covariates predict and distinguish among these principal strata, but assume constancy of the effect of R across covariate levels within these two strata. Comparing the separate effects of R on Y identified within each of these strata represent a way for assessing the $R * M$ interaction.

9. Supplementary Materials

The SAS Macros for the RPMI as they apply to the two example data sets in this article (Suicide Prevention and Therapy Studies in Section 7) are available under the Article Information link at the *Biometrics* website <http://www.tibs.org/biometrics>. More general software for more than two covariates will be available from the first author at ttenhave@cceb.med.upenn.edu.

ACKNOWLEDGEMENTS

The authors thank Dylan Small, Michael Elliott, Knashawn Morales, Joseph Gallo, Mark Cary, and two reviewers for very insightful comments that improved the article tremendously. Funding was provided by NIMH grants: R01-MH61892, R01-MH59380, R01-CA095415, P30-MH066270, R01-MH60915, P20-MH71905, and R37-CCR316866.

REFERENCES

- Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**, 1173–1182.
- Brown, G. K., Ten Have, T. R., Henriques, G. R., Xie, S. X., Hollander, J. E., and Beck, A. T. (2005). Cognitive therapy for the prevention of suicide attempts: A randomized controlled trial. *Journal of the American Medical Association* **294**, 2847–2848.
- Bruce, M. L., Ten Have, T. R., Reynolds, C. F., III, Katz, I. R., Schulberg, H. C., Mulsant, B. H., Brown, G. K., McAvay, G. J., Pearson, J. L., and Alexopoulos, G. S. (2004). A randomized trial to reduce suicidal ideation and depressive symptoms in depressed older primary care patients: The PROSPECT study. *Journal of the American Medical Association* **291**, 1081–1091.
- Fischer-Lapp, K. and Goetghebeur E. (1999). Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Randomized Controlled Trial Controlled Clinical Trials* **20**, 531–546.
- Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.
- Joffe, M. M., Hoover, D. R., Jacobson, L. P., Kingsley, L., Chmiel, J. S., Fischer, B. R., and Robins, J. M. (1998). Estimating the effect of Zidovudine on Kaposi's sarcoma from observational data using a rank preserving failure time model. *Statistics in Medicine* **17**, 1073–1102.
- Joffe, M. M., Small, D., and Hsu, C. (2007). Defining and estimating intervention effects for groups who will develop an auxiliary outcome. *Statistical Science*. Accepted for publication.
- Mealli, F. and Rubin, D. B. (2003). Commentary: 'Assumptions allowing the estimation of direct causal effects'. *Journal of Econometrics* **112**, 79–87.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Translated by D. M. Dabrowska and edited by T. P. Speed (1990). *Statistical Science* **5**, 465–472.
- Robins, J. M. and Greenland, S. (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association* **89**, 737–749.
- 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.
- Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3**, 319–336.
- Rubin, D. (1978). Bayesian inference for causal effects. *The Annals of Statistics* **6**, 34–58.
- Ten Have, T. R., Elliott, M., Joffe M. M., Zanutto, E., and Datto, C. (2004). Causal models for randomized physician encouragement trials in treating primary care depression. *Journal of the American Statistical Association* **99**, 8–16.
- Vansteelandt, S. and Goetghebeur, E. (2004). Using potential outcomes as predictors of treatment activity via strong structural mean models. *Statistica Sinica* **14**, 907–925.

Received February 2006. Revised October 2006.

Accepted November 2006.