Causal Mediation Analysis in multi-level non-randomized exposure and multi-component mediator case

Cheng Zheng* Xiao-hua Andrew Zhou†

Abstract

Mediation analysis is an important tool in social and behavirol sciences as it helps to understand why a behavioral intervention works. To yield a causal interpretation the most common approach (e.g., Baron and Kenny, 1986), as discussed by (Imai, Keele, and Tingley 2010), needs a sometime unrealistic assumption of "sequential ignorability". Rank preserving model (RPM; Ten Have et al., 2007) was proposed to relax this assumption. However, RPM is restricted to the case with binary intervention and single mediator. Also, it needs the strong "rank preserve" assumption. We proposed a new model that can handle multi-level intervention and a multi-component mediator with a weaker assumption. Also, our model has the ability to handle correlated data and missing data. Finally our method can also be used in many other research settings, which have a similar model as mediation analysis such as treatment compliance, post randomized treatment component analysis. For the proposed causal mediation model, we first showed identifiability for the parameters in the model. We then proposed a semi-parametric method for estimating the model parameters and derive the asymptotic results for the proposed estimators. Simulation showed that our model gave robust results. Finally we applied the proposed method to the two real-world clinical studies and our method is applied to two data analysis: (1) the college student drinking study (2) IMPACT study.

Key Words: Causal Inference, Generalized Estimating Equation, Correlated Data, Missing Data

1. Introduction

In many research projects, we hope to know not only whether the intervention can affect a certain outcome, but also, how it affects the outcome. And we are also interested in testing whether the effect is through a certain pathway and estimating the direct and indirect causal effects. Mediation analysis is a tool for answering these questions. However, the traditional regression based mediation analysis such as Baron and Kenny and its extensions do not have a valid causal interpretation unless the sequential ignorability assumption holds. The resulting estimator may be biased due to an unmeasured confounder between the mediator and the outcome. Based on Rubin's idea of potential outcome, several causal frameworks were proposed to make a valid inference on the mediation analysis with causal interpretation. One framework is principal stratification (PS), which is often used in the context of compliance for taking a drug. To make the principal stratification model identifiable, exclusion restriction is often assumed, and this assumption violates our purpose for mediation analysis in terms of estimate the direct effect. To relax such the assumption, Gallop et al. (2009) proposed a model based on strong normality assumption of the outcome within each strata. One additional limitation of the PS model, as pointed out by Vanderweele (2011), is that the parameter estimated in the PS model only represents an associative effect and the mediation portion is quantified by the portion of associative effect, and hence the resulting mediation interpretation may be misleading. Another framework for making causal mediation analysis is the counterfactual model such as the structural mean model (SMM), which is based on the concept of potential outcomes, assuming that we can manipulate both intervention and mediator levels for the patient. Under the counterfactual

*Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195
†Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195

SMM model, we can estimate controlled mediation effects. The SMM does not assume any specific distribution for its residuals and hence is a semiparametric model, which can be estimated by a estimating equation based method. The price that the SMM pays for its semiparametric nature is that some extra assumptions on the covariates are needed to make the model identifiable; for example, for a SMM model without an interaction for the outcome, we need a covariate that modifies the effect of intervention/exposure on mediator but does not modify the direct and mediated effects. Fortunately, the existence of such an interaction is partly testable from the observed data. One such the SMM model is rank preserving model (RPM) proposed by Ten Haven. However the RPM has several limitations. First it requires the randomization of intervention and needs the intervention to be binary. Second it can handle only one mediator and requires the outcome to be continuous. Third, it can't handle correlated data and longitudinal data. Fourth the RPM does not allow missing data. In this paper, we extend the original RPM to a more general case where we can handle a more general problem, including a continuous/multi-level mediator, a vector mediator, multi-level(continuous) intervention, ignorable but non-randomized intervention. Finally, our model can also handle the correlated and missing data.

2. Motivational Examples

2.1 Interventions for College Student Drinking and Comorbid Mental Health Problems

The first example data come from a randomized trial on the effectiveness of an intervention in reducing problematic college student drinking and comorbid depression or anxiety. College student drinking spans a spectrum from occasional use to heavy and chronic use, and previous research has highlighted numerous problems associated with heavy drinking in college students, including physical injuries, legal troubles (e.g., DUI), school or work problems, and unwanted sexual encounters.

Participants in the study were randomized to one of three conditions: a) relaxation control in which participants were provided a calm environment to relax for an hour (i.e., no relaxation skills were taught), b) brief alcohol screening and intervention for college students (BASICS), or c) DBT-BASICS, a condition that combines BASICS with specific skills for emotion regulation from dialectical behavior therapy. It was hypothesized that DBT-BASICS might be particularly helpful for the subset of college drinkers who drink in part to cope with negative affect, as the DBT skills specifically target strategies for emotion regulation. The active interventions consisted of a single, one hour session with a facilitator, and participants were assessed at one month and three months following their baseline assessment. Based on the theory underlying the intervention, DBT-BASICS should affect depression (i.e., BDI) through its impact on emotion regulation (i.e., DERS). Specifically, there is a notable improvement in DERS following the intervention, between baseline and one month assessments, and this is in sharp contrast to the pattern seen in the control condition. Thus, our goal in the present article is to use mediation analysis to assess how effect of BASICS and DBT-BASICS on the outcome is mediated through their influence on emotion regulation. This data raise the question how we should incorporate a multi-level intervention. Although the existing RPM may be used to compare each level of the intervention to the control separately, the use of the RPM is not efficient.

2.2 Improving Mood-Promoting Access to Collaborative Treatment for Late Life Depression(IMPACT)

In this study, 1801 subjects were recruited from 18 primary care clinics affiliated with 8 diverse health care organizations in 7 distinct geographic regions across US. The IMPACT intervention was a mlti-modal intervention that included a care manager who assessed the initial depression and provided suggestion about antidepressant medication and psychotherapy approaches to treatment. The patients are randomized to two groups, one with assigned care manager and one without. All patients were offered a choice of either use antidepressant medication or a 6-8 session problem solving treatment in primary care (PST-PC) or both. The aim of the study is to determine the relative effects of antidepressant medication, PST-PC and their combination. We consider this as a model with two binary mediators, M_1 for the indicator of antidepressant medication use, M_2 for the indicator of PST-PC use and their interaction term. Our outcome is depression score measured by 20 depression items from the Symptom Checklist (SCL-20) at 3,6,9,12 months and the intervention Z is the indicator whether a patient is in the group with care manager assignment. Here for our primary analysis, we use the outcome measured at 12 months. The covariates measured include sex, age, marital status, ethnicity, education level, medicare coverage, prescription medication coverage, prior episodes of depression, baseline depression score, thoughts of suicide, cognitive impairment, chronic disease, significant chronic pain, health-related functional impairment, overall quality of life, antidepressant use in the past 3 months, specialty mental health visits or psychotherapy in the past 3 months. We are also interested in mediation analysis when considering the outcome at different time point together rather than just one specific time point. This data set raises the question about how to assess the relative effects of a mediator with multiple components and correlated data.

3. Method

3.1 Parameters of Interest

We define two parameters of interest: The first parameter of interest is the direct effect of treatment level z_1 comparing to treatment level z_2 when the vector of the mediators is fixed at level $\boldsymbol{m}$, which is defined as $E(Y_i^{z_1\boldsymbol{m}} - Y_i^{z_2\boldsymbol{m}})$. The second parameter of interest is the mediator effect, defined by the difference in outcome for a particular set of mediator values on multiple mediators, i.e., $\boldsymbol{m}_1$ versus mediating level $\boldsymbol{m}_2$, which is defined as $E(Y_i^{z\boldsymbol{m}_2} - Y_i^{z\boldsymbol{m}_1})$. For both of the two parameters, the effect can be modified by baseline covariate $\boldsymbol{X}$ and we can define the direct and mediator effect among subgroup with covariate $\boldsymbol{X}$ at certain level. Mathematically, we will be interested in $E[(Y_i^{z_1\boldsymbol{m}} - Y_i^{z_2\boldsymbol{m}})|\boldsymbol{X}_i = \boldsymbol{x}]$ and $E[(Y_i^{z\boldsymbol{m}_2} - Y_i^{z\boldsymbol{m}_1})|\boldsymbol{X}_i = \boldsymbol{x}]$.

3.2 Model

For participant j in cluster i, we denote Y_{ij} as the observed continuous outcome, Z_{ij} as the randomized intervention assignment indicator, $\boldsymbol{X}_{ij}$ as a vector of covariates, and $\boldsymbol{M}_{ij}$ as a vector of mediation variables. The potential outcome $Y_{ij}^{z\boldsymbol{m}}$ is defined as the outcome variable that would be observed for the jth participant in cluster i if the participant is randomized to intervention level z and receives mediation at level $\boldsymbol{m}$. In graphical model representation, we assume that $\boldsymbol{U}$ is a vector of unmeasured confounders between $\boldsymbol{M}$ and Y. Using the notation by Pearl, we can represent the potential outcome as $Y_{ij}^{z\boldsymbol{m}} = Y(do\ Z = z, do\ \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{U}_{ij}, \boldsymbol{X}_{ij})$, which means that $Y_{ij}^{z\boldsymbol{m}}$ is the outcome if we can fixed Z and $\boldsymbol{M}$ at level z and $\boldsymbol{m}$. Let $\boldsymbol{U}_{ij}$ be the value of unmeasured confounder vector subject i in

cluster j. We propose a new model that extends the original rank preserving model(RPM) as follows:

$$Y_{ij}^{z\boldsymbol{m}} = g(\boldsymbol{X}_{ij}) + \sum_{k=1}^{K} \theta_k h_k(z, \boldsymbol{m}, \boldsymbol{X}_{ij}) + \varepsilon^{z\boldsymbol{m}}(\boldsymbol{U}_{ij}, \boldsymbol{X}_{ij}), \tag{1}$$

where the effects of intervention and mediator on the potential outcomes, $h_k(\cdot), k = 1, \cdots, K$, are known functions, which satisfy $h(0, 0, \boldsymbol{X}) = 0$, the error term $\varepsilon(\boldsymbol{U}, \boldsymbol{X})$ has an unknown distribution with mean zero, $E(\varepsilon^{zm}(\boldsymbol{U}, \boldsymbol{X})|\boldsymbol{X}) = 0$, and $g(\cdot)$ is an unknown function. The key feature for this model is that we allow the existence of the vector of unmeasured confounders, $\boldsymbol{U}$, between the outcome and the mediator. However, such the confounders should not modify the effect of Z and $\boldsymbol{M}$ on outcome Y. Now under this model, we can write the specific form of our parameters of interest as below:

$$E[(Y_i^{z_1\boldsymbol{m}} - Y_i^{z_2\boldsymbol{m}})|\boldsymbol{X}_i = \boldsymbol{x}] = \sum_{k=1}^{K} \theta_k [h_k(z_1, \boldsymbol{m}, \boldsymbol{x}) - h_k(z_2, \boldsymbol{m}, \boldsymbol{x})]$$

$$E[(Y_i^{z\boldsymbol{m}_1} - Y_i^{z\boldsymbol{m}_2})|\boldsymbol{X}_i = \boldsymbol{x}] = \sum_{k=1}^{K} \theta_k [h_k(z, \boldsymbol{m}_1, \boldsymbol{x}) - h_k(z, \boldsymbol{m}_2, \boldsymbol{x})].$$

As we can see that the key is to estimate $\boldsymbol{\theta}$ since the form of $h(\cdot)$ is pre-specified.

For independent data, here we can see that the original RPM proposed by Ten Have is just a special case of our model with $K = 2$, and the functions of h_1, h_2, and ε^{zm} have following forms:

$$h_1(Z, \boldsymbol{M}, \boldsymbol{X}) = Z, h_2(Z, \boldsymbol{M}, \boldsymbol{X}) = \boldsymbol{M}, \varepsilon(\boldsymbol{U}, \boldsymbol{X}) = \varepsilon(\boldsymbol{U}).$$

Also the RPM is the special case where there is only one observation per cluster and assume $\varepsilon(\boldsymbol{U})$ is mean zero error term uncorrelated to Z given X. However, as discussed below, our model can handle many more cases beyond the original RPM setting.

3.3 Assumptions

To make parameters in the model identifiable, we need the following assumptions.

1. Assumption 1: The stable unit treatment value assumption (SUTVA). This means there are no multiple versions of the intervention, and there is no interference between participants. This allows us to use a single model for the sample, and we do not need to consider the effect of one person's intervention on others' outcome. This assumption is needed for most of the existing models to have a causal interpretation, although it is rarely stated explicitly. There are scenarios in which this may not be reasonable; for example, a group therapy intervention in which participants interact with one another may violate this assumption, or a behavioral intervention in which the provider becomes more skilled over time in delivering the intervention also could be problematic for this assumption.

2. Assumption 2: Consistency. We assume the observed outcome is just one realization of the potential outcome with observed intervention level r and mediator level m. This assumption is required to make a connection between the observed outcome and the potential outcome. A mathematic representation of this can be written as

$$Y = \sum_{z,\boldsymbol{m}} Y^{z\boldsymbol{m}} I(Z = z, \boldsymbol{M} = \boldsymbol{m}),$$

where $I(\cdot)$ is an indicator function.

3. Assumption 3: Randomization (i.e. ignorability). It means that the observed intervention assignment is independent of the mediator and all potential outcomes defined by different levels of intervention and mediator conditional on covariates. For the traditional model, it implies that the intervention is not affected by any unmeasured covariates, and thus there is no confounding between intervention and mediator or outcome. A mathematic representation of this can be written as

$$(Y^{z\boldsymbol{m}}, z \in \Omega_z, \boldsymbol{m} \in \Omega_{\boldsymbol{m}}) \perp Z|X,$$

where Ω_z is the support of Z, and the $\Omega_{\boldsymbol{m}}$ is the support of $\boldsymbol{M}$

4. Assumption 4: Mean model correctly specified. It requires that the mean model defined by equation (1) is correctly specified up to unknown parameters. The only exception is that we allow misspecification on function $g(\cdot)$. This requirement suggests that there is no interaction between unmeasured confounders and the mediators, between unmeasured confounders and the intervention. This is similar to assuming that there is no moderation of direct and mediator effect by unmeasured confounders. This assumption is just partly testable since we can not test the model on potential outcomes and we can only test the derived model for the observed outcome.

5. Assumption 5: Same conditional mean for the error term. This assumption says that the conditional mean for the error term should be the same for potential outcomes under any levels. Mathematically, this assumption means $E[\varepsilon^{z\boldsymbol{m}}(\boldsymbol{U}_{ij}, \boldsymbol{X}_{ij})|M, X, Z] = F(M, X, Z)$ for all levels of z and $\boldsymbol{m}$. Note that this is much weaker than the original assumption of "rank preserving" in the RPM, which indicates $\varepsilon^{z\boldsymbol{m}}(\boldsymbol{U}_{ij}, \boldsymbol{X}_{ij}) = \varepsilon(\boldsymbol{U}_{ij}, \boldsymbol{X}_{ij})$ for all levels of z and $\boldsymbol{m}$.

6. Assumption 6: The variance covariance matrix of that random vector is positive definite,

$$Cov([E(h_1(Z, \boldsymbol{M}, \boldsymbol{X})|\boldsymbol{X}, Z), \cdots, E(h_K(Z, \boldsymbol{M}, \boldsymbol{X})|\boldsymbol{X}, Z)]|\boldsymbol{X}) > 0,$$

where ">" means positive-definite. The interpretation of this assumption depends on the specific model. In general, the covariance matrix can be estimated from the observed data and we can test whether the smallest eigenvalue is 0. The interpretation of this assumption in some special cases is given in the Section 4.

3.4 Identifiability

Theorem.
Model (1) is globally identifiable under Assumption 1-6 with regularity conditions given in the appendix.
In this theorem, the first three assumptions are trivial and the fourth and fifth assumptions are related to our correct specification of the structural mean model form. The key assumptions that is crucial to identifying the model with our method is Assumption 6. It is worth while to note that this assumption is sufficient but not necessary.

3.5 Estimation

We denote $Z_i = (Z_{i1}, \cdots, Z_{in_i})$,$\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \cdots, \boldsymbol{X}_{in_i})$,$Y_i = (Y_{i1}, \cdots, Y_{in_i})$, and $U_i = (U_{i1}, \cdots, U_{in_i})$,where n_i is the number of participants in the ith cluster. Under model (1) with Assumption 1,2, 4 and 5, we have

$$E(Y_{ij}|Z_{ij}, M_{ij}, X_{ij}) = g(\boldsymbol{X}_{ij}) + \sum_{k=1}^{K} \theta_k h_k(Z_{ij}, \boldsymbol{M}_{ij}, \boldsymbol{X}_{ij}) + F(Z_{ij}, \boldsymbol{M}_{ij}, \boldsymbol{X}_{ij}). \quad (2)$$

With Assumption 3, we know that

$$E[F(Z_{ij}, \boldsymbol{M}_{ij}, \boldsymbol{X}_{ij})|Z_{ij}, X_{ij}] = E[F(Z_{ij}, \boldsymbol{M}_{ij}, \boldsymbol{X}_{ij})|X_{ij}] = 0.$$

Now we denote $\tilde{Y}_{ij}(\boldsymbol{\theta}) = Y_{ij} - \sum_{k=1}^{K} \theta_k h_k(Z_{ij}, \boldsymbol{M}_{ij}, \boldsymbol{X}_{ij})$, then we have

$$E(\tilde{Y}_{ij}(\boldsymbol{\theta})|Z_{ij}, X_{ij}) = E(g(\boldsymbol{X}_{ij}) + F(Z_{ij}, \boldsymbol{M}_{ij}, \boldsymbol{X}_{ij})|Z_{ij}, \boldsymbol{X}_{ij}) = g(\boldsymbol{X}_{ij}),$$

or more general $\tilde{Y}_{ij}(\boldsymbol{\theta}_0) \perp Z_{ij}|X_{ij}$ under $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is the true parameters. This suggests that any estimating equation, based on $Cov(\tilde{Y}_{ij}(\boldsymbol{\theta}), \tilde{A}(Z_{ij}, X_{ij})|X_{ij}) = \mathbf{0}$, is unbiased, where $\tilde{A}(\cdot, \cdot)$ can be arbitrary function. So to get a consistent estimator for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)^T$, we can use the following estimating equations:

$$\sum_{i=1}^{n} \mathbf{A}(Z_i, \boldsymbol{X}_i)^T [\tilde{Y}_i(\boldsymbol{\theta}) - \tilde{g}(\boldsymbol{X}_i)] = \mathbf{0}, \tag{3}$$

where

$$\mathbf{A}^T(Z_i, \boldsymbol{X}_i) = \begin{pmatrix} a_1(Z_{i1}, \boldsymbol{X}_{i1}) & \cdots & a_1(Z_{in_i}, \boldsymbol{X}_{in_i}) \\ \cdots & \cdots & \cdots \\ a_K(Z_{i1}, \boldsymbol{X}_{i1}) & \cdots & a_K(Z_{in_i}, \boldsymbol{X}_{in_i}) \end{pmatrix} \Omega_i(\boldsymbol{X}_i).$$

Here $\Omega_i(\boldsymbol{X}_i)$ is a $n_i \times n_i$ matrix and $a_j(Z_{ik}, \boldsymbol{X}_{ik})$ is any function that satisfies $E(\mathbf{a}_j(Z, \boldsymbol{X})|\boldsymbol{X}) = 0$. $\tilde{g}(\boldsymbol{X})$ is a working model of $g(\boldsymbol{X})$ and is not required to be correctly specified. So the estimator is robust to a misspecification of $\tilde{g}(X)$ for $g(X)$ and unmeasured confounder U. However, a bad choice of $\tilde{g}(\boldsymbol{X})$ may lose efficiency. When we choose a wrong working model, our estimating equation will be $Cov(\tilde{Y}_{ij}(\boldsymbol{\theta}), \tilde{A}(Z_{ij}, \boldsymbol{X}_{ij})|\boldsymbol{X}_{ij}) + E\{[g(\boldsymbol{X}) - \tilde{g}(\boldsymbol{X})]A(Z_{ij}, \boldsymbol{X}_{ij})\} = 0$. Because of our restriction on $a_k(Z_{ij}, \boldsymbol{X}_{ij})$, i.e. $E[a_k(Z_{ij}, \boldsymbol{X}_{ij})|\boldsymbol{X}_{ij}] = 0$, we have $E\{[g(\boldsymbol{X}_{ij}) - \tilde{g}(\boldsymbol{X}_{ij})]A(Z_{ij}, \boldsymbol{X}_{ij})\} = E\{[g(\boldsymbol{X}_{ij}) - \tilde{g}(\boldsymbol{X}_{ij})]E[A(Z_{ij}, \boldsymbol{X}_{ij})|\boldsymbol{X}_{ij}]\} = \mathbf{0}$. So the estimating equation (3) is still unbiased and will result in a consistent M-estimator for $\boldsymbol{\theta}$ even if $\tilde{g}(\boldsymbol{X})$ is a wrong model of $g(\boldsymbol{X})$. In practice, we often use some parametric working model for $\tilde{g}(\cdot)$, such as letting $\tilde{g}(\boldsymbol{X}) = \tilde{g}(\boldsymbol{X}, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is $q \times 1$ vector of parameters. We can use the following method to get a consistent estimator for $\boldsymbol{\theta}$. We can first arbitrarily choose a $\boldsymbol{\beta}^{(0)}$ and then solve the estimating equations to obtain $\hat{\boldsymbol{\theta}}^{(0)}$. Then we estimate $\boldsymbol{\beta}^{(1)}$ by fitting a regression model of $E[\tilde{Y}(\hat{\boldsymbol{\theta}}^{(0)})|\boldsymbol{X}_i] = \tilde{g}(\boldsymbol{X}_i, \boldsymbol{\beta})$ and solve estimating equations to obtain $\hat{\boldsymbol{\theta}}^{(1)}$ by using $\tilde{g}(\boldsymbol{X}) = \tilde{g}(\boldsymbol{X}, \boldsymbol{\beta}^{(1)})$. Then iteratively update $\hat{\boldsymbol{\theta}}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$ until $\hat{\boldsymbol{\theta}}^{(t)}$ converge. We denote

$$G_{1i}(\boldsymbol{\beta}, \boldsymbol{\theta}) \quad = \quad \mathbf{A}(Z_i, \boldsymbol{X}_i)^T [\tilde{Y}_i(\boldsymbol{\theta}) - \tilde{g}(\boldsymbol{X}_i, \boldsymbol{\beta})],$$

and

$$G_{2i}(\boldsymbol{\beta}, \boldsymbol{\theta}) \quad = \quad \mathbf{B}(\boldsymbol{X}_i)^T [\tilde{Y}_i(\boldsymbol{\theta}) - \tilde{g}(\boldsymbol{X}_i, \boldsymbol{\beta})].$$

Here $\mathbf{B}(\boldsymbol{X}_i)$ can be any matrix with the element in the j-th row and the k-th column as $b_j(\boldsymbol{X}_{ik})$. For example, we can use $b_j(\boldsymbol{X}_{ik}) = \frac{\partial \tilde{g}(\boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \beta_j}$. The third method, when converge, is equivalent to solving the following estimating equations simultaneously:

$$\sum_{i=1}^{n} G_{1i}(\boldsymbol{\beta}, \boldsymbol{\theta}) \quad = \quad \mathbf{0},$$

and

$$\sum_{i=1}^{n} G_{2i}(\boldsymbol{\beta}, \boldsymbol{\theta}) \quad = \quad \mathbf{0}.$$

A special case is that if we use a linear working model $g(X) = \boldsymbol{\beta}^T \boldsymbol{X}$, then we can get a closed form for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ without doing iteration. However, when the iteration does not converge, we will use one step updated estimator $\hat{\boldsymbol{\theta}}^{(1)}$.

3.6 Asymptotic Theorems

3.6.1 ***Theorem 1***.

We denote the true value of parameters $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$ and assume that under the working model, there is unique solution $\boldsymbol{\beta}_0$ to $EG_{2i}(\boldsymbol{\beta}, \boldsymbol{\theta}_0) = \mathbf{0}$. Under the assumption 1-6 and regularity conditions as listed in the appendix, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow_d N(\mathbf{0}, V),$$

where V is the subset matrix of

$$(E(\frac{\partial G_i(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \frac{\partial G_i(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}})^T E(G_i(\boldsymbol{\beta}, \boldsymbol{\theta}) G_i(\boldsymbol{\beta}, \boldsymbol{\theta})^{-1}) E(\frac{\partial G_i(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \frac{\partial G_i(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}}))^{-T}.$$

Here $G_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = [G_{1i}(\boldsymbol{\beta}, \boldsymbol{\theta}), G_{2i}(\boldsymbol{\beta}, \boldsymbol{\theta})]$, and the expectation is taken under $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_0$. This result directly comes from the theory of generalized estimating equations.
Note that the assumption of unique solution to $EG_{2i}(\boldsymbol{\beta}, \boldsymbol{\theta}_0) = \mathbf{0}$ is required to ensure the working model is identifiable. When using a linear working model, this assumption will always hold. However, when using a nonlinear model, it is hard to check whether this assumption hold. One way to partly check it is that if the regression model $EY|X = g(X, \boldsymbol{\beta})$ is not identifiable, then the assumption fails.

The variance covariance matrix of the estimators, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ can be estimated by the following sandwich estimator:

$$E_n[\frac{\partial G(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}, \frac{\partial G(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}]^T E_n[G(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) G(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})^T] E_n[\frac{\partial G(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}, \frac{\partial G(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}],$$

where E_n denotes the empirical expectation. Since the empirical expectation will converges to the true expectation uniformly under certain regularity condition and since the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are consistent, by uniform law of large number, the sandwich estimator is a consistent estimator for the variance covariance matrix.

3.7 Selection of the Weights

In our proposed estimating equation (3), we can select any function for $\mathbf{a}(Z, \boldsymbol{X})$ that satisfies $E(\mathbf{a}(Z, \boldsymbol{X})|\boldsymbol{X}) = \mathbf{0}$ to obtain a consistent estimator. Here we consider how to select $\mathbf{a}(Z, \boldsymbol{X})$ so that the resulting estimator is the most efficient (the smallest variance). Since our focus is on a vector of parameters, the most efficient estimator means that for any linear combination of the parameters, that estimator has the smallest variance. This is equivalent to find a covariance matrix that is smaller than other covariance matrix where "smaller" means their difference is negative definite. We have the following result.

3.7.1 ***Theorem 2***.

If the $\tilde{g}(\boldsymbol{X})$ is known and the covariance of the error term given covariates $\boldsymbol{X}$ in the model (1) (e.g. $\Omega(\boldsymbol{X}) = Var(\varepsilon^z \boldsymbol{m}(\boldsymbol{U}, \boldsymbol{X})|\boldsymbol{X})$) does not depend on z and $\boldsymbol{m}$, then the best

choice of $\mathbf{A}(Z, \boldsymbol{X})$ that gives the most efficient estimator has the following form:

$$\mathbf{A}(Z, \boldsymbol{X}) = [E(\frac{\partial \tilde{Y}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \boldsymbol{X}, Z) - E(\frac{\partial \tilde{Y}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \boldsymbol{X})]\Omega^{-1}(\boldsymbol{X}),$$

where $\Omega^{-1}(\boldsymbol{X})$ is the inverse of the covariance matrix of the error term given covariate $\boldsymbol{X}$. Under model (1), the optimally chosen weight can be written as

$$a_k(Z, \boldsymbol{X}) = (E(h_k(Z, \boldsymbol{M}, \boldsymbol{X}) | \boldsymbol{X}, Z) - E(h_k(Z, \boldsymbol{M}, \boldsymbol{X}) | \boldsymbol{X}))\Omega^{-1}(\boldsymbol{X}). \quad (4)$$

Since the optimal weight involves two unknown quantities, $E(h_k(Z, \boldsymbol{M}, \boldsymbol{X}) | \boldsymbol{X}, Z)$ and $\Omega^{-1}(\boldsymbol{X})$, we can model them by some parametric models, $f(\boldsymbol{M} | \boldsymbol{X}, Z, \boldsymbol{\alpha})$ and $\Omega^{-1}(\boldsymbol{X}, \boldsymbol{\eta})$, respectively. It is easy to estimate $\hat{\boldsymbol{\alpha}}$ by some regression of $\boldsymbol{M}$ on $\boldsymbol{X}$ and Z. And $\Omega^{-1}(\boldsymbol{X})$ can be modeled and estimated from squared residuals $Y_i^{00}(\boldsymbol{\theta}) - g(\boldsymbol{X}_i)$ v.s. $\boldsymbol{X}_i$. We can obtain more efficient estimator by either one step or iterative update of $\Omega^{-1}(\boldsymbol{X})$ with solving estimating equation for $\boldsymbol{\theta}$. After getting the estimator for $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, we obtain the estimated optimal weight $\hat{a}_k(Z, \boldsymbol{X}) = (E(h_k(Z, \boldsymbol{M}, \boldsymbol{X}) | \boldsymbol{X}, Z, \hat{\boldsymbol{\alpha}}) - E(h_k(Z, \boldsymbol{M}, \boldsymbol{X}) | \boldsymbol{X}, \hat{\boldsymbol{\alpha}}))\Omega^{-1}(\boldsymbol{X}, \hat{\boldsymbol{\eta}})$. Here we point out that the misspecification of these two models will not affect the validity of the inference. When the intervention, Z, is binary variable, we have

$$E(h_k(Z, \boldsymbol{M}, \boldsymbol{X}) | \boldsymbol{X}, Z) - E(h_k(Z, \boldsymbol{M}, \boldsymbol{X}) | \boldsymbol{X}) = -(Z - EZ)W(X),$$

where

$$W_k(X) = Eh_k(Z, \boldsymbol{M}, \boldsymbol{X} | Z = 1, \boldsymbol{X}) - Eh_k(Z, \boldsymbol{M}, \boldsymbol{X} | Z = 1, \boldsymbol{X}).$$

Under the special case as in RPM, we have $h_1(Z, M, \boldsymbol{X}) = Z$, $h_2(Z, M, \boldsymbol{X}) = M$, and then the best weight should be $(Z - EZ)W(X)$, where $W(X) = [1, E(M | Z = 1, \boldsymbol{X}) - E(M | Z = 0, \boldsymbol{X})]$, which is consistent with the result from the original RPM. When we just have $\tilde{g}(\boldsymbol{X}, \boldsymbol{\beta})$ correctly specified, the best weight will include the selection of both $A(\cdot)$ and $B(\cdot)$ and is not discussed here.

3.8 Missing Data

In this section, we consider how to make an inference when some subjects are missing the outcome under the missing at random (MAR) assumption. We denote R to be the missing data indicator for a subject and assume a parametric model for the missing data mechanism $\pi(Z, \boldsymbol{X}, \boldsymbol{M}) = P(R = 1 | Z, \boldsymbol{X}, \boldsymbol{M}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a vector of r parameters. Note here we assume that the missing data mechanism only depends on observed mediation level rather than the potential mediator level for different intervention level, so we assume there is no unmeasured confounder between the mediator and the missing indicator. In the presence of missing data, we propose to use the following weighted GEE to estimate $\boldsymbol{\theta}$:

$$\sum_{i=1}^{n} \frac{R_i}{\pi(Z_i, \boldsymbol{X}_i, \boldsymbol{M}_i, \hat{\boldsymbol{\gamma}})} \mathbf{A}(Z_i, \boldsymbol{X}_i)^T (\tilde{Y}_i(\theta) - \tilde{g}(\boldsymbol{X}_i)) = \mathbf{0}. \quad (5)$$

4. Some Special Cases

In this section, we apply the general methodology to several specific and important applications. In each application, we give the specific identifiability assumption (i.e. assumption 6), and estimating equations with the best weight.

4.1 Multiple-level mediator

Now we consider the case that the mediator, M, has L+1 different levels (0-L), and there is a different effect of mediator on outcome under different levels. We assume the following model:

$$Y_{ij}^{z\boldsymbol{m}} = g(\boldsymbol{X}_{ij}) + \sum_{l=1}^{L} \theta_l I(m = l) + \theta_z z + \varepsilon_{ij}^{zm}. \tag{6}$$

Using the general approach, it is easy to get that the estimating equation with the optimal weight has the following form:

$$\sum_{i=1}^{n} (Z_i - E(Z)) W(\boldsymbol{X}_i)(Y_i - \sum_{l=1}^{L} \theta_l I(M_i = l) - \tilde{g}(\boldsymbol{X}_i)) = \mathbf{0},$$

where weight is

$$[1, P(M = 1|\boldsymbol{X}, Z = 1) - P(M = 1|\boldsymbol{X}, Z = 0),$$
$$\cdots, P(M = L|\boldsymbol{X}, Z = 1) - P(M = L|\boldsymbol{X}, Z = 0)].$$

To obtain $W(\boldsymbol{X})$, we just need to fit a multinomial logistic regression or some nonparametric regression to obtain the distribution of $f(M|Z, \boldsymbol{X})$. And the Assumption 6 now means that $Var(W(\boldsymbol{X})) > 0$ which indicates that there is interaction between the intervention and the covariates on $f(M|Z, \boldsymbol{X})$.

4.2 Multiple-component mediator

This is the case of our IMPACT data where the mediator M includes two components, the use of antidepressant drug and the PST-PC session. More generally, we can consider that $\boldsymbol{M}$ has L components noted as $M^{(1)}, \cdots, M^{(L)}$. If all components are discrete, then they can be modeled as one component with many levels, which represents different combinations of levels of $M^{(1)}, \cdots, M^{(L)}$. When at least one component is continuous, we can only directly derive the result from a general model. For an illustration purpose, we consider $L = 2$ and a model with one interaction. We write this model as follows:

$$Y_{ij}^{z\boldsymbol{m}} = g(\boldsymbol{X}_{ij}) + \theta_1 z + \theta_2 m^{(1)} + \theta_3 m^{(2)} + \theta_4 m^{(1)} \times m^{(2)} + \varepsilon_{ij}^{z\boldsymbol{m}}. \tag{7}$$

Using the general approach, it is easy to show that the estimating equation with the optimal weight has the following form:

$$\sum_{i=1}^{n} (Z_i - E(Z)) W(\boldsymbol{X}_i)(Y_i - \theta_1 Z_{ij} - \theta_2 M_{ij}^{(1)} - \theta_3 M_{ij}^{(2)} - \theta_4 M^{(1)} \times M^{(2)} - \tilde{g}(\boldsymbol{X}_i)) = \mathbf{0},$$

where

$$W(\boldsymbol{X}) = [1, E(M^{(1)}|\boldsymbol{X}, Z = 1) - E(M^{(1)}|\boldsymbol{X}, Z = 0), E(M^{(2)}|\boldsymbol{X}, Z = 1)$$
$$-E(M^{(2)}|\boldsymbol{X}, Z = 0), E(M^{(1)} \times M^{(2)}|\boldsymbol{X}, Z = 1) - E(M^{(1)} \times M^{(2)}|\boldsymbol{X}, Z = 0)].$$

The Assumption 6 which requires $Cov(W(\boldsymbol{X})) > 0$ now has the similar implication as in multiple level mediator case. It means there is some interactions between $\boldsymbol{X}$ and Z on the joint distribution of $M^{(1)}$ and $M^{(2)}$. To estimate the weight, we need to estimate

the joint conditional distribution $f(\boldsymbol{M}|\boldsymbol{X}, Z)$, saying by a multinomial regression model when all components are discrete. When the number of components in M is large, or there is some continuous components, fitting such a multinomial regression model may be difficult. In this case, we may use three models for $f(M^{(1)}|\boldsymbol{X}, Z)$, $f(M^{(2)}|\boldsymbol{X}, Z)$, and $f(M^{(1)} \times M^{(2)}|\boldsymbol{X}, Z)$ to estimate the weight. Although these models might be inconsistent to each other and thus cause efficiency lose, our estimator for $\boldsymbol{\theta}$ is still consistent and asymptotically normal distributed.

4.3 Multiple level intervention

This is the case of our drink problem data where we have three levels of intervention: BASICS, DBT-BASICS and control. More generally, we consider a multi-arm intervention, Z, which can take value $0, 1, \cdots, or L$. And the model can be written as follows:

$$Y_{ij}^{z\boldsymbol{m}} = g(\boldsymbol{X}_{ij}) + \sum_{l=1}^{L} \theta_l I(z = l) + \theta_{L+1} m + \varepsilon^{z\boldsymbol{m}}. \tag{8}$$

Using the general approach, it is easy to show that the estimating equation with the optimal weight have the following form:

$$\sum_{i=1}^{n} \mathbf{a}(Z, \boldsymbol{X})(Y_i - \sum_{l=1}^{L} \theta_l I(Z_i = l) - \theta_{L+1} M_i - \tilde{g}(\boldsymbol{X}_i)) = \mathbf{0},$$

Following the general result, we can easily obtain that the optimal weight, $\mathbf{a}(Z, \boldsymbol{X})$, should have the following form:

$$(I(Z = 0) - P(Z = 0|\boldsymbol{X}), \cdots,$$
$$I(Z = L) - P(Z = L|\boldsymbol{X}), E(M|\boldsymbol{X}, Z) - \sum_{l=0}^{L} E(M|\boldsymbol{X}, Z = l) P(Z = l|\boldsymbol{X})).$$

To estimate these weights, we need to fit an regression model for $f(M|\boldsymbol{X}, Z, \boldsymbol{\alpha})$ and a multinomial regression model for $P(Z|\boldsymbol{X})$. For a randomized trial, we know that the intervention is randomized rather than just ignorable conditional on baseline covariates, so we can use $E(\tilde{Z}_{(l)}))$ instead of $E(\tilde{Z}_{(l)}|X))$. Since there is $L + 1$ weights here and if all $P(Z = l|\boldsymbol{X})$ are in linear form of $\boldsymbol{X}$, then we need that covariates X modify the effect of Z on M at all level of Z. We need these interaction terms has an positive definite covariance matrix. If we assume a linear model $E(M|\boldsymbol{X}, Z, \boldsymbol{\alpha})$, then a necessary condition is that we need at least L covariate to make the model possibly identifiable.

5. Simulation

In this section, we conducted several simulations studies to evaluate the relative performance of our proposed methods with the commonly used OLS approach in the finite sample sizes for four different settings. These methods are compared under both the settings sequential ignorability assumption for the treatment and without the assumption. In the simulation results, we refer our model as a robust method.

5.1 Multi-level Mediator Case

In the first simulation setting, we considered a four-level mediator model. We want to see how our method performs, compared with the OLS. We simulated a four valued M,

a vector of observed covariates $\boldsymbol{X}$ with two components, X_1 and X_2, and an unobserved confounder U under the following model:

$$\begin{aligned}
&\log \frac{P(M=1|Z=1,\boldsymbol{X})}{P(M=3|Z=1,\boldsymbol{X})} = \alpha_1 X_1 + \alpha_{u1} U, \\
&\log \frac{P(M=2|Z=1,\boldsymbol{X})}{P(M=3|Z=1,\boldsymbol{X})} = \alpha_2 X_2 + \alpha_{u2} U, \\
&P(M=0|Z=0,\boldsymbol{X}) = 1, \\
&P(M=0|Z=1,\boldsymbol{X}) = 0, \\
\textit{and} \quad &Y^{zm} = \theta_1 I(m=1) + \theta_2 I(m=2) + \theta_3 I(m=3) + \beta_x X + \beta_u U + N(0,\sigma^2),
\end{aligned}$$

where $X \sim N(0,1), U \sim U(0,1)$, and $\theta_1 = 0.5, \theta_2 = 1, \theta_3 = 1.5, \beta_x = 1, \beta_u = 2$, $\alpha_1 = -5, \alpha_2 = 5, \alpha_{u1} = 3, \alpha_{u2} = -3, \sigma = 0.1$. Since the M is 0 for $Z = 0$ and nonzero for $Z = 1$, we exclude the term for Z in the model of Y to make the model identifiable. With a sample size $n = 400$, and 10000 simulations. We also show simulation result where β_u is set to be 0 and hence the sequential ignorability assumption hold. All results are given in table 1. From the result, we notice that in this case, OLS is a little bit more efficient.

5.2 Intervention has Multiple Level Case

In this section, we study the performance of our estimator in a simulation setting similar to the drink problem data where the intervention, Z, is multi-level and the mediator M is continuous. The model we used to generate data can be written as:

$$Y_i^{zm} = \beta X_i + \theta_1 I(z=1) + \theta_2 I(z=2) + \theta_3 m + \varepsilon. \tag{9}$$

$$\begin{aligned}
&E(M|Z,X,U) \\
= \quad &\gamma_{z1} I(Z=1) + \gamma_{z2} I(Z=2) + \gamma_x X + \gamma_{z1x} I(Z=1) X + \gamma_{z2x} I(Z=1) X^2 + \gamma_u U,
\end{aligned}$$

X is standard normally distributed. Here U is an unmeasured confounder with the standard uniform distribution, which is independent of X and a non-zero coefficient associated with U can cause the violation of the assumption of sequential ignorability, which is necessary for the validity of OLS. And ε follows a normal distribution with mean zero and standard error 0.1, and the working model for $g(X)$ is $\beta_0 + \beta_1 X$.

In the simulation, we chose the sample size to be 400 and ran 10000 simulations, parameters were set as below:

$\beta = 1, \theta_3 = 1, \theta_1 = 1, \theta_2 = 1, \gamma_{z1} = 0.2, \gamma_{z2} = -0.2, \gamma_x = 0, \gamma_{z1x} = -0.5, \gamma_{z2x} = 0.5, \gamma_u = 0.1$. And the variance of the residual for both model of Y and model of M was set to be 0.1.

From the simulation result, we conclude that our method has better coverage rate closer to the nominal level and smaller bias than the OLS method, but tend to have larger variance which cause the MSE is similar with that from OLS. However, as discussed before, this deficit can be improve when sample size increase and the contribution of bias to MSE increase. This simulation results are similar to that given in Ten Have's paper for his original RPM. So it is probably that the larger variance is a feature of this kind of robust method unless we can find suitable covariates X to make the weight far from collinear. However, the mean square error of the new method is not always better than OLS, especially in the case with small sample size.

Variable	Method	Bias	MSE	95% CR
Multi-level mediator with sequential ignorability				
θ_1	Robust	-0.0005	0.0063	94.7
θ_1	OLS	-0.0006	0.0038	94.9
θ_2	Robust	-0.0011	0.0128	97.2
θ_2	OLS	-0.0010	0.0059	95.1
θ_3	Robust	-0.0028	0.3278	95.1
θ_3	OLS	-0.0008	0.0168	93.8
Multi-level mediator without sequential ignorability				
θ_1	Robust	0.0022	0.0147	95.0
θ_1	OLS	0.0965	0.0181	83.2
θ_2	Robust	0.0014	0.0290	94.8
θ_2	OLS	-0.1458	0.0346	75.9
θ_3	Robust	0.0021	0.7064	98.0
θ_3	OLS	0.0357	0.0389	94.4
Multi-level intervention without sequential ignorability				
θ_1	Robust	-0.0004	0.0015	95.5
θ_1	OLS	-0.0075	0.0015	95.0
θ_2	Robust	-0.0023	0.0021	94.8
θ_2	OLS	0.0045	0.0020	96.0
θ_3	Robust	-0.0003	0.0014	95.7
θ_3	OLS	0.0349	0.0014	29.5
Multi-level intervention with missing data				
θ_{z1}	Robust	-0.0002	0.0016	94.5
θ_{z1}	OLS	-0.0039	0.0046	94.7
θ_{z2}	Robust	-0.0004	0.0023	94.4
θ_{z2}	OLS	0.0017	0.0059	94.9
θ_m	Robust	-0.0001	0.0015	94.9
θ_m	OLS	0.0349	0.0017	61.4

Table 1: Simulation Results

5.3 Missing data Case

Now, we use the same model as in simulation for three-level z but introduce missing mechanism for the outcome Y which is specified by model:

$$logitE(D|Z, M, X) = \eta_0 + \eta_{z1}I(Z = 1) + \eta_{z2}I(Z = 2) + \eta_m M + \eta_x X$$

where D is the indicator whether certain outcome is observed. And the true parameters are set to be $\eta_0 = -1, \eta_{z1} = 1, \eta_{z2} = 1, \eta_m = 0.2, \eta_x = 0.5$. From the result, we notice that the OLS gives estimator with larger bias, MSE and lower CR while our model gives unbiased estimator with good coverage rate.

From the results, we notice that the robust method has smaller bias, smaller MSE and give correct coverage rate. So we conclude by weighting estimating equation, our robust estimator works well when the missing is at random,

6. Data Analysis

In this section, we applied our developed new method to our two real data problems to see whether the results are different from those using OLS.

Variable	Method	Estimates	95% CI
DBT-BASICS	Robust	2.88	(0.03,5.73)
DBT-BASICS	OLS	2.45	(-0.86,5.76)
BASICS	Robust	3.68	(0.89,6.46)
BASICS	OLS	2.63	(-0.34,5.60)
One month DERS	Robust	0.13	(0.004,0.26)
One month DERS	OLS	0.23	(0.13,0.32)

Table 2: Results for College Student Drinking Data

6.1 Interventions for College Student Drinking and Comorbid Mental Health Problems

We applied our method to College Student Drinking data. Our intervention is three leveled variable (0: Control, 1: DBT-BASICS, 2: BASICS) and the mediator is one month difficulties in emotion regulation scale (DERS) while the outcome is one month Beck depression inventory (BDI). The covariates included in the analysis are baseline BAI and BDI. The result for the estimates as well as their confidence intervals are shown in Table 2. Although the parameter estimates from the new method and the OLS method yield the same directions, their magnitude are different. The variance estimates for the two methods also similar. Also, we can see that the direct effects of two intervention level are both significant better than control using our method, but this direct effect is not significant if we use the OLS method. This suggests that using the OLS method, we might make a wrong conclusion due to potential violation of the sequential ignorability assumption.

6.2 Improving Mood-Promoting Access to Collaborative Treatment for Late Life Depression(IMPACT)

In this data set, we considered the care manager as the intervention and the use of antidepressant medication and PST-PC as two components of a mediator. The outcome used in this analysis is the continuous-scale depression score. As the data set contain many covariates, it provides us a good opportunity to obtain good weight which make the covariance matrix in assumption 6 has large eigenvalue. The results are shown in Table 3. The covariates used in the model to construct the weight function are the organization type, age, baseline SCL score, the total number of health therapy in the past 3 months before baseline, the total number of anti-depressants taken before baseline and the total number of months of previous anti-depressants taken before baseline. And it shows that the direct effect of the care manager intervention is nearly 0 while the drug has a strong treatment effect, however it is interesting that mental health therapy has side effect which means taking PST-PC session will reduce the effect of drug taking. The working correlation for people in different clinic center was set to be independence.

From the result, we notice that the direct effect of having the care manager is significantly different from 0 when using the OLS method. However, it is no longer significant when using our method. Also, we notice that the difference in the point estimates from these methods is large. Also, the variance of our estimator is larger than that from OLS estimator, which indicates the covariance matrix in Assumption 6 is approximate singular since we know the variance will goes to infinity when the covariance matrix given in Assumption 6 is singular.

Variable	Method	Estimates	95% CI
Care Manager	Robust	-0.36	(-1.01, 0.28)
Care Manager	OLS	-0.28	(-0.34, -0.22)
antidepressant medication	Robust	-0.11	(-2.34, 0.20)
antidepressant medication	OLS	-0.06	(-0.16, 0.04)
PST-PC	Robust	0.35	(-0.62, 1.33)
PST-PC	OLS	-0.001	(-0.08, 0.08)
Interaction	Robust	1.76	(-1.22, 4.75)
Interaction	OLS	0.08	(-0.01, 0.21)

Table 3: Results for IMPACT data covariates selected by AIC

7. Discussion

In this paper, we have proposed a general approach for mediation analysis, which has the advantage of dealing with missing data and correlated data as well as allowing for flexible types of mediators and interventions. From the simulation results, we have shown that our model yields an unbiased estimate for the causal parameters and the correct coverage rates. However, the data analysis shows that the variance of our estimate could still be large. This may be due to the misspecified model for $g(\boldsymbol{X})$, inefficiency estimation of $\mathbf{a}(\boldsymbol{X}, Z)$, non-constant variance, or not including enough covariates to make Assumption 6 hold. Because the property of positive definite is needed for the covariance matrix given in Assumption 6 to make the parameter identifiable and if the covariance matrix is near to singular, the variances of our estimators might be large. In this case, we need to include more covariates that not modify the direct and mediator effects to make Assumption 6 hold to reduce the variances. But this is not always easy either for the reason that we did not have very important covariates or because the included covariates may be modeled incorrectly for their interaction with intervention and mediator. Since the identification of the model rely on the non-collinearity for $\mathbf{a}(Z, X)$, we need to include enough X to satisfy that.

In conclusion, this paper has introduced a new model as an alternative to the traditional method for mediation analysis with more flexible type of mediator and interventions. The proposed method can deal with the correlated data and the missing data. The key advantage of our model is that its estimators are unbiased, even in the presence of unmeasured confounding. Our model has a weaker assumption than the original RPM which can also handle unmeasured confounding issues. Other two ways to deal with the unmeasured confounding include (1) to perform sensitivity analysis, and (2) to use the principal stratification framework. For the first way, Imai et al. (2010) proposed sensitivity analysis, in which the effect of a hypothesized unmeasured confounder on the indirect effect can be examined. There are also other different kind of sensitivity analysis such as VanderWeele (2010) and Imai, Keele and Yamamoto (2010). The second way is the use of the principal stratification (PS) framework, in which the confounders no longer exist after the models is stratified on the principle strata. However, the identification of PS model often needs to specify the residual distributions and solve some mixture model, such as in Gallop et al. (2009). This restricts the use of the PS model to mediators with only a few levels, e.g. binary. Also, the definition of parameters for direct and indirect effects in the PS model is slightly different from the traditional method and is in fact associative and dissociative effect.

REFERENCES

Baron, R. M. and Kenny, D. A. (1986), "The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations," *J Pers Soc Psychol*, 51, 1173–1182.

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joeff, M. and TenHave, T. R. (2009), "Mediation analysis with principal stratification," *Stat in Medicine*, 28, 1108–1130.

Imai, K., Keele, L. and Yamamoto, T. (2010), "Identification, inference, and sensitivity analysis for causal mediation analysis," *Statistical Science*, 25, 51–71.

Imai, K., Keele, L. and Tingley, D. (2010), "A general approach to causal mediation analysis," *Psychol Methods*, 15, 309–334.

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

VanderWeele, T. J. (2010), "Bias formulas for sensitivity analysis for direct and indirect effects," *Epidemiology*, 21, 540–551.

VanderWeele, T. J. (2011), "Principal stratification: uses and limitations.," *International Journal of Biostatistics*, 7, 1–14.