Generalized Structural Mean Models for Evaluating Depression as a Post-treatment Effect Modifier of a Jobs Training Intervention*

Alisa Stephens† Luke Keele‡ Marshall Joffe§

July 7, 2016

Abstract

In randomized controlled trials, the evaluation of an overall treatment effect is often followed by effect modification or subgroup analyses, where the possibility of a different magnitude or direction of effect for varying values of a covariate is explored. While studies of effect modification are typically restricted to pretreatment covariates, longitudinal experimental designs permit the examination of treatment effect modification by intermediate outcomes, where intermediates are measured after treatment but before the final outcome. We present a novel application of generalized structural mean models (GSMMs) for simultaneously assessing effect modification by post-treatment covariates and accounting for noncompliance to assigned treatment status. The proposed approach may also be used to identify post-treatment effect modifiers in the absence of noncompliance. The methods are evaluated using a simulation study that demonstrates that our approach retains consistent estimation of effect modification by intermediate variables that are affected by treatment and also predict outcomes. We illustrate the method using a randomized trial designed to promote re-employment through teaching skills to enhance self-esteem and inoculate job seekers against setbacks in the job search process. Our analysis provides some evidence that the intervention was much less successful among subjects that displayed higher levels of depression at intermediate post-treatment waves of the study. We also compare the assumptions of our approach and principal stratification as alternatives to account for differences in effects by intermediate variables.

*For helpful comments and suggestions, we thank Stijn Vansteelandt, Eric Tchetgen Tchetgen, Teppei Yamamoto, the Associate Editor and the reviewers.

†Department of Epidemiology and Biostatistics, University of Pennsylvania Perelman School of Medicine 624 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, E-mail: alisaste@mail.med.upenn.edu

‡304 Old North, 37th & O St, NW, Georgetown University, Washington, D.C., E-mail: lk681@georgetown.edu

§Department of Epidemiology and Biostatistics, University of Pennsylvania Perelman School of Medicine 602 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104

1 The JOBS II Study and Effect Modification

Evaluation is an important aspect of policy interventions such as job-training programs. Here, we evaluate the JOBS II Intervention Project developed at the University of Michigan and designed to enhance the reemployment prospects of unemployed workers (Vinokur et al. 1995). The intervention aimed to teach unemployed workers skills related to searching for employment such as the preparation of job applications and resumes and how to successfully interview. An additional focus of the intervention, however, was on the mental health aspects of the job search process. This component of the training included activities to enhance self-esteem, increase a sense of self-control, and cope with set-backs. These skills were taught to help job-seekers maintain motivation and persist in the job-search process.

Of the sampled workers, the researchers randomly assigned 1249 to the job search seminar (treatment) and 552 to the control condition, which consisted of a short pamphlet on job search strategies. Workers assigned to the treatment condition attended a 20-hour job search seminar over one week. Follow-up interviews were conducted 6 weeks, 6 months, and 2 years after the intervention. We focus on whether the intervention increased re-employment. Unlike the original analysis, we also examine how covariates measured post-treatment might be used to better evaluate the effectiveness of JOBS II. We conduct two different analyses based on post-treatment covariates.

Many previous analyses have focused on intention-to-treat (ITT) effects of participation in JOBS II (Vinokur et al. 1995; Vinokur and Schul 1997; Imai et al. 2010a; Jo 2008) (though see Jo and Vinokur (2011); Little and Yau (1998); Mattei et al. (2013) exceptions). While ITT effects are important, there are other relevant causal quantities when there is noncompliance. In JOBS II only 61% of those assigned to the intervention actually attended the training seminars, while those assigned to control could not access the treatment. It is therefore relevant to focus on whether the intervention was efficacious among those who actually attended the job search seminar, which requires conditioning on post-treatment

information (Robins 1994; Angrist et al. 1996).

In addition to accounting for noncompliance, we also evaluate post-treatment effect modification, an understudied use of post-treatment covariates in the analysis of randomized trials. In a randomized study of treatments, effects may be heterogenous, observed as an interaction between a treatment and an effect modifying covariate such that the average treatment effect varies across values of the covariate. For example, we can consider treatment effects that vary by covariate-defined subpopulations such as sex or race. While analyses with effect modification by a pretreatment covariate are relatively common, it is also possible for effect modification to occur as a function of a post-treatment covariate. In many randomized studies, data on post-treatment or intermediate covariates, defined as variables measured post-intervention but prior to the study endpoint, are often collected. For example in JOBS II, after treatment, intermediate measures were collected at time intervals such as six weeks and six months after treatment, whereas the final outcome measures were collected two years later. In such designs, we may suspect that the treatment effect may vary across levels of a covariate measured after the treatment but before the final outcome.

There are several reasons to consider the possibility of effect modification by a post-treatment variable. First, post-treatment effect modification can be used for intermediate decision making if the trial is ongoing. Analysts could use the model to identify subgroups for whom the treatment is particularly ineffective and a new intervention might be implemented. Second, results from an analysis of this form could also be used as a method for hypothesis-generation and the design of future interventions. Third, the model might be combined with other identification strategies to show a consistent pattern of associations in support of a causal hypothesis. Keele (2014) provides one example where post-treatment effect modification is used as an alternative identification strategy to instrumental variables. In that example, similar conclusions from alternative identification strategies are used to bolster a single causal hypothesis.

Post-treatment effect modification is an important but yet unstudied aspect in JOBS II.

In designing the study, effect modification by pretreatment levels of depression was of particular concern. As a result 520 unemployed workers were excluded from the overall sample of eligible subjects prior to randomization since they displayed a clinically significant level of depression (Vinokur et al. 1995). This exclusion allowed the researchers to apply the intervention to the subpopulation in which it would be most effective. While job loss is known to induce depression, the original study did not consider that re-employment failures—failed interviews, a lack of call backs—may also increase levels of depressive symptoms. If re-employment failures elevated levels of depression after the intervention, the effectiveness of the treatment for this subpopulation may be reduced. We use a model of post-treatment effect modification to estimate whether post-treatment levels of depression reduced the effectiveness of the treatment.

In our analysis, we adopt the framework of potential outcomes to define causal effects based on comparisons of potential outcomes on a common set of units (Rubin 1974, 1978). Our primary estimand is the causal odds ratio among the treated within subgroups defined by post-treatment levels of depression. We show how our estimand may be characterized as an example of single potential outcome stratification under the principal stratification (PS) framework considering depression levels among those assigned to treatment (Joffe et al. 2007). A key contribution of our analysis is exploring and outlining identifiability conditions for this estimand. We use generalized structural mean models (GSMMs) for binary outcomes and a modified G-estimation procedure to estimate the post-treatment effect modification of the causal odds ratio among compliers (Vansteelandt and Goetghebeur 2003). While additive SMMs have been applied to post-treatment effect modification (Dunn and Bentall 2007), we adapt them to allow for estimation in the odds-ratio scale.

Our paper has the following structure. Section 2 provides basic descriptive statistics and some preliminary analyses. Section 3 outlines our notation, describes our causal estimand, states identifiability conditions. In Section 4, we detail the estimation procedure. Section 5 evaluates the properties of the two-parameter logistic GSMM for post-treatment effect

modification through a simulation study. Section 6 presents estimates of post-treatment effect modification of causal effects in JOBS II. Section 7 includes discussion and concluding remarks.

2 Descriptive Summaries and Preliminary Analyses

For all subjects in the JOBS II study, researchers collected covariates prior to treatment assignment. Baseline covariates include education, income, sex, age, occupation, race, risk for failure, level of economic hardship, and a measure of depressive symptoms. The primary outcome of interest is a binary indicator for whether subjects were employed 20 or more hours per week at the two year follow-up period. We use all units from the original sample with nonmissing values at baseline and at intermediate data collection time points. The intent-to-treat (ITT) analysis reveals that the odds of success in the treatment arm as compared to the control arm is 1.49 with a 95% confidence interval $(1.10, 2.01)$. This implies that for those assigned to treatment the odds of re-employment were 49% higher. In ignoring noncompliance, the ITT estimate speaks to program effectiveness but not efficacy. We address the question of efficacy below.

Next, we examine whether levels of depression appeared to be elevated at post-treatment follow-up periods. Depressive symptoms were measured with a scale of 11 items from the Hopkins Symptom Checklist with scores ranging from 0 to 6.0. A score of 3.00 or greater on the depression index was considered to be a clinically significant indication of depression. As we noted above, subjects with scores of 3 or more were excluded from the JOBS II study prior to randomization. We re-scaled the depression scale to range from 0 to 100, which aids interpretation. On this scale, subjects with a score of 50 or higher were removed from the study. Figure 1 contains box plots of depression scores at baseline and the two follow-up periods. The measure of depression in the plot excludes all subjects who were removed due to a high level of depressive symptoms at baseline. While the median level of depression decreases at the follow-up periods, for some subjects, levels of depression are elevated well

above the 50 point threshold which indicates a clinically significant level of depression in the post-treatment periods. Here, we examine whether the intervention was less effective among subjects with higher levels of post-treatment depression.

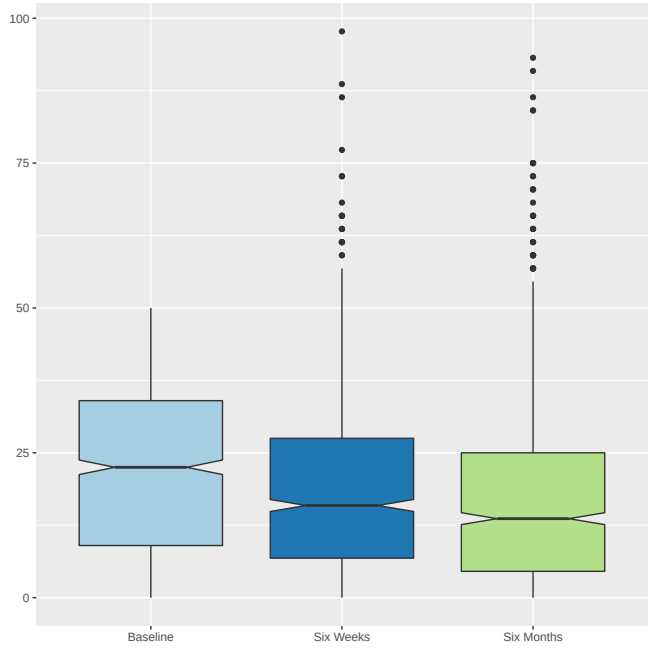

Figure 1: Levels of depression at baseline, the six week, and six month follow up. Scores above 50 are considered to be a clinically significant indication of depression.

3 Estimand and Identification Conditions in the Analysis of JOBS II

Next, we describe the causal estimand of interest in the analysis of the JOBS II trial using potential outcomes structural mean models (SMMs). SMMs were developed for the analysis of randomized trials with noncompliance (Robins 1994), but provide a general structure for estimating the effect of post-randomization exposures (Vansteelandt and Goetghebeur 2004; Vansteelandt 2010). We also outline the assumptions needed for identification of our estimand, since under both noncompliance and post-treatment effect modification, we condition on post-treatment quantities. Rosenbaum (1984) demonstrates that conditioning on post-treatment covariates may result in biased estimates of the causal parameter. We

examine the identifiability conditions for post-treatment effect modification in detail since identification assumptions under noncompliance are well-known. We conclude this section with a detailed discussion of our estimand within the PS framework (Frangakis and Rubin 2002).

3.1 Causal Estimand and Initial Assumptions

In JOBS II, subjects ($i = 1, \ldots, n$) are randomly assigned to either treatment ($R_i = 1$) or control ($R_i = 0$). Covariates $\mathbf{X}_i = (X_1, \ldots, X_k)$ are measured at baseline prior to randomization, including age, race, and baseline depression. We define post-treatment effect modifiers, $\mathbf{S}_i$, as the set of intermediate covariates observed after treatment but prior to the outcome, also possibly multivariate. In our application, we focus on a single potential post-treatment effect modifier: level of depressive symptoms which we denote by S_i. Further, to reflect the choice by subjects to comply with their treatment assignment, we denote actual exposure to the treatment by A_i. In JOBS II, those assigned to control did not have access to the training sessions in the intervention condition. Therefore when $R_i = 0$ then $A_i = 0$. Conversely, if $R_i = 1$, then $A_i = 1$ if subject i complies with treatment assignment and attends the job training seminars, and if $A_i = 0$, the subject does not comply with treatment assignment and does not attend. The observed response, denoted by Y_i and which follows self-selected exposure A_i, is an indicator of employment of 20 hours or more per week. The temporal order of observed variables is $\mathbf{X}_i, R_i, A_i, S_i, Y_i$.

One common way to define causal effects is in terms of counterfactual or potential outcomes (Neyman 1923; Rubin 1978; Holland 1986). Under the potential outcomes framework, $Y_{ir,a}$, is the potential outcome when $R_i = r$ and $A_i = a$. To estimate the causal effect of the JOBS II intervention on employment, we would like to compare how each subject would respond under treatment to his or her response under the control condition. The potential outcome under no treatment is $Y_{ir,0}$, and it indicates a treatment-free response that would have been observed if, perhaps counter to fact, subject i had not received treatment. We

further define $A_{i,r}$ and $S_{i,r,A_{i,r}}$ as potential outcomes of treatment exposure and intermediate depression under treatment assignment $R_i = r$.

Next, we stipulate a set of assumptions for identifiability of the effect of A_i. First, we assume that the SUTVA holds (Rubin 1986) which has the two following components: 1) there are no hidden forms of treatment, which implies that for unit i under $R_i = r$ and $A_i = a$, we assume that $Y_{ir,a} = Y_i$ and 2) a subject's potential outcome is not affected by other subjects' exposures. The first component of SUTVA is often referred to as the consistency assumption in the epidemiological literature. Schwartz et al. (2012) contains a discussion of the relationship between these forms of assumptions.

We assume the exclusion restriction holds which states that $Y_{i1,0}$ is equal to $Y_{i0,0}$, and we denote $Y_{i0} = Y_{i1,0} = Y_{i0,0}$ (Angrist et al. 1996). The exclusion restriction implies that being invited to the job training seminar has no direct effect on the odds of re-employment two years later. It seems plausible that being invited to a training seminar itself does little to change the probability of re-employment other than through actual participation in the seminar. We assume that the treatment assignment is ignorable, defined as independence between the treatment assignment and potential outcomes under no treatment. In JOBS II, this assumption is justified by the randomized design. Formally, this restriction can be expressed as

$$Y_{i0} \perp R_i | \mathbf{X}_i \tag{1}$$

Further, we assume the "no-contamination" restriction, defined as the absence of off-protocol use of the intervention among controls, such that $P(A_i = 0 | R_i = 0) = 1$ (Cuzick et al. 2007). This assumption is justified by the design since control subjects could not access the intervention. The no contamination restriction implies the no current treatment interaction assumption for SMMs, which states that the causal effect of the intervention is identical among subjects who actually took it, regardless of assigned treatment (Hernán and Robins 2006), since it reduces the number of causal effects to one among those assigned to the intervention. The causal effect among those assigned to control can never be estimated,

so the underlying data-generating mechanism remains agnostic about the existence of the interaction under unrestricted designs (Clarke and Windmeijer 2010; Robins 1994). We also assume that R_i has a nonzero causal effect on A_i. These assumptions will identify the effect of A_i on Y_i and are equivalent to the usual instrumental variables assumptions (Angrist et al. 1996). We consider further assumptions to identify effect modification later.

To ease exposition, we denote two changes to the notation. First, we drop the index i, since throughout we assume all quantities are indexed by individual participants in the JOBS II trial. Second we denote $A_{r=1}$ as A_1. That is, A_1 represents the potential compliance status under fixed treatment assignment $R = 1$. We also simplify S_{1,A_1}, the potential level of depression for a potential compliance status with fixed treatment assignment $R = 1$, to S_{A_1} and define Y_{A_1} similarly. In our application, we focus on treatment effect modification by S_{A_1}. Specifically, we consider the following form of effect modification by S_{A_1} where if for some s not equal to s'

$$E[Y_{A_1} - Y_0 | S_{A_1} = s, R = 1] \neq E[Y_1 - Y_0 | S_{A_1} = s', R = 1]. \tag{2}$$

In JOBS II, the primary outcome of interest is binary, so we focus on effect modification by S_{A_1} the logistic SMM (Robins et al. 1999), which models the log-causal-odds of employment vs. unemployment among subjects randomized to treatment as a function of exposure and covariates. Under the logistic SMM, our estimand is the causal odds ratio

$$\exp(\psi) = \frac{P(Y = 1 | S_{A_1}, A_1, R = 1, \mathbf{X})}{P(Y = 0 | S_{A_1}, A_1, R = 1, \mathbf{X})} \Big/ \frac{P(Y_0 = 1 | S_{A_1}, A_1, R = 1, \mathbf{X})}{P(Y_0 = 0 | S_{A_1}, A_1, R = 1, \mathbf{X})}. \tag{3}$$

which we allow to vary across S_{A_1}. Under a set of appropriate identification conditions, we may obtain estimates of the estimand using the following model

$$\begin{aligned} \text{logit}\{P(Y_1 = 1 | S_{A_1}, A_1, R = 1, \mathbf{X})\} - \text{logit}\{P(Y_0 = 1 | S_{A_1}, A_1, R = 1, \mathbf{X})\} = \\ \eta_s'(A_1, S_{A_1})\boldsymbol{\psi}_0 = f_1(A_1; \psi_{01}) + f_2(A_1, S_{A_1}; \psi_{02}), \end{aligned} \tag{4}$$

where $\boldsymbol{\psi}$ represents the unknown causal parameter, with the truth denoted by $\boldsymbol{\psi}_0$, and $\eta_s'(\cdot)$ is a function of potential outcomes S_{A_1} and A_1 with dimension equal to that of $\boldsymbol{\psi}$. We use the subscript 's' to indicate that this is a structural model, since it refers to counterfactual quantities instead of observed associations in the data. In Equation (4) the structural model is comprised of arbitrary known functions $f_1(\cdot)$ and $f_2(\cdot)$ up to an unknown p-dimensional parameter $\boldsymbol{\psi}_0 = (\psi_{01}, \psi_{02})$, where $p = dim(\psi_{01}) + dim(\psi_{02})$. When $\boldsymbol{\psi}_0 = \mathbf{0}$ or $A_1 = 0$ we assume $f_1(\psi_{01}) = f_2(S_1; \psi_{02})$=0 to indicate no effect of treatment, and for S_{A_1} or $\psi_{02} = 0$, we assume $f_2(S_i; \psi_{02}) = 0$. We condition on $\mathbf{X}_i$ but by assumption we rule out structural models of the form $\eta_s'(A_1, \mathbf{X}_i) = (A_1, A_1\mathbf{X}_i)'$. This is discussed in further detail in section 3.2.

As an example, consider the case when $f_1(\psi_{01}) = \psi_{01}A_1$ and $f_2(\psi_{02}) = \psi_{02}A_1S_{A_1}$. For $\psi_{02} = 0$, the causal odds ratio is quantified by $\exp(\psi_{01})$ for all subjects, implying no effect heterogeneity. When $\psi_{02} \neq 0$, $\exp(\psi_{01})$ is the causal odds ratio among the subset of subjects with $S_{A_1} = 0$, and when $\psi_{02} \neq 0$, the causal odds ratio for subjects with $S_{A_1} = s$ is captured by $\exp(\psi_{01} + \psi_{02}s)$, which allows the effect of A_1 to vary with the counterfactual S_{A_1}.

Joffe et al. (2007) show that models like (4) were discussed as models with a single potential outcome stratification, in contrast with a principal stratification approach, which considers a stratification on joint potential outcomes under $r = 0, 1$. We show in section 3.2 that model (4) is equivalent to the model

$$\text{logit}\{E[Y_i|S_i, A_i, R_i = 1, \mathbf{X}_i]\} - \text{logit}\{E[Y_{i0}|S_i, A_i, R_i = 1, \mathbf{X}_i]\} = \eta_s'(A_i, S_i)\boldsymbol{\psi}_0 = f_1(A_i; \psi_{01}) + f_2(A_i, S_i; \psi_{02}), \quad (5)$$

which stratifies on the observed auxiliary variable S_i (Joffe et al. 2007), and falls under the class of Retrospective Structural Mean Models (RSMMs). RSMMs are characterized by exposure effects defined conditional on variables observed subsequent to treatment, in contrast with standard structural mean models (Robins 1994), which define causal effects

as a function of covariates observed prior to treatment. Vansteelandt (2010) also considered retrospective models for assessing mediation when outcomes are binary and modeled using a logit link.

Alternatively, we could use a linear SMM, which models mean differences linearly in exposure and covariates under an identity link. For positive outcomes, we might apply the log link to estimate the causal risk ratio. When mean outcomes are close to 1, either marginally or conditionally within subgroups, modeling binary outcomes using the identity or log link may result in predicted mean outcomes that are out of range, which can cause nonconvergence or falsely reported convergence in estimation routines. The logistic SMM allows for general binary outcomes that may be common or rare.

We might compare the causal odds ratio in Equation (3) to a more familiar one

$$\frac{P(Y=1|A_1,R_i=1,\mathbf{X})}{P(Y_i=0|A_1,R=1,\mathbf{X})}\Big/\frac{P(Y_0=1|A_1,R=1,\mathbf{X})}{P(Y_0=0|A_1,R=1,\mathbf{X})}, \tag{6}$$

which only allows effect modification by $\mathbf{X}$. For this estimand, the structural model would be rewritten as $\eta_s'(A_1,\mathbf{X})\boldsymbol{\psi}$. Structural models of this type admit the possibility that causal odds ratio is not constant over different strata of $\mathbf{X}$. The distinction between Equations (3) and (6) are made clearer in the following section where we consider questions of identifiability.

Finally, we could also consider an alternative form of effect modification

$$E[Y_{A_1}-Y_0|S_{A_1}-S_{A_0}=s,R=1]\neq E[Y_1-Y_0|S_{A_1}-S_{A_0}=s',R=1] \tag{7}$$

This second form of effect modification allows for the effect of the intervention to vary by $S_{A_1}-S_{A_0}$ instead of S_{A_1}. Under this second form of effect modification, we would seek to answer whether the intervention is more or less effective among those whose intermediate levels of depression are affected by exposure to the intervention. Under the form of effect modification in 2, we are attempting to understand whether the intervention was less effective with increasing depression levels under treatment.

Under a simplified setting without noncompliance, effect modification of the type in (2) may be captured by the model

$$E[Y_1 - Y_0|S_1, S_0, \mathbf{X}, R = 1] = \gamma_1 R + \gamma_2 R S_1,$$

whereas effect modification of the type in (7) may be modeled by

$$E[Y_1 - Y_0|S_1, S_0, \mathbf{X}, R = 1] = \gamma_1 R + \gamma_2 R(S_1 - S_0).$$

These models are nested in the following more general model

$$E[Y_1 - Y_0|S_1, S_0, \mathbf{X}, R = 1] = \gamma_1 R + \gamma_2 R S_1 + \gamma_3 R S_0, \tag{8}$$

which suggests that a test for the appropriateness of other model may be conducted by evaluating the hypothesis that γ_3=0. In applications where the intermediate variable S is only defined among the treated, these models are equivalent. Dunn and Bentall (2007) consider effect modification of this type where the intervention is assignment to therapy and the effect modifier is attachment to the therapist. Since controls patients do not receive therapy, they cannot form an attachment to a therapist. See (Follmann 2006) for an additional example of this type in vaccine trials.

When the effect modification variable occurs in both treatment arms and varies both under intervention and control, as it does in the JOBS II application, choosing between these two models will depend on subject matter knowledge. We argue that effect modification of the form in (2) is more relevant when interest focuses on the level of the effect modifier rather than the difference in the effect modifier caused by the intervention. As we noted above, the eligibility criteria for JOBS II excluded otherwise eligible subjects that had high levels of depression at baseline because the intervention was less likely to be effective among them. Given this, we focus on the levels of depression achieved under treatment as the effect

modifier. Moreover, model (8) can only be identified under additional parametric modeling assumptions beyond those we use for identification.

3.2 Identifiability Under Post-treatment Effect Modification and Noncompliance

We next consider the identifiability of SMMs with post-treatment effect modification, since identification of treatment effects for those who complied with the JOBS II holds given the assumptions stated thus far. We address identifiability under a theorem presented by Vansteelandt and Goetghebeur (2004) in the context of Strong Structural Mean Models. First, we consider the following model that parameterizes the odds ratio (6), under a single binary pre-treatment effect modifier X_i,

$$\text{logit}\{E(Y|A_1, R=1, X)\} - \text{logit}\{E(Y_0|A_1, R=1, X)\} = \psi_{01}A_1 + \psi_{02}A_1X \tag{9}$$

This model is nonparametrically identified in the sense of Robins (1997) under the assumptions in Section 3.1. The log odds ratio in Equation (9) is uniquely defined in terms of observable quantities given the ignorability assumption, the consistency component of SUTVA, the no-contamination assumption, and equivalence between $E[Y|A_1 = 0, R = 1, X]$ and $E[Y_0|A_1, R = 1, X]$ under the exclusion restriction.

We contrast the model in Equation (9) with an example of Equation (4) as given by:

$$\text{logit}\{E(Y|S_{A_1}, A_1, R=1, X_i)\} - \text{logit}\{E(Y_0|S_{A_1}, A_1, R=1, X_i)\} = \psi_{01}A_1 + \psi_{02}A_1S_{A_1} \tag{10}$$

using a single binary potential post-treatment effect modifier S_{A_1}. Nonparametric identification for the above model does not hold since there are four possible nonzero effects defined by joint levels of X, and S_{A_1}, but the exchangeability assumption allows identification of only two parameters. Generally, nonparametric identifiability does not hold for models of

this form, since the number of strata jointly defined by the baseline covariate, X, and potential intermediate, S_{A_1}, is greater than the number of restrictions imposed by our stated assumptions (Vansteelandt and Goetghebeur 2004).

To achieve model-based identification of effect modification by S_{A_1}, we must place a restriction on the number of A_1X interactions in the structural model for the effect of treatment on outcomes. To derive the model-based identifiability conditions, we consider the marginal treatment effect Δ_x, marginalizing over S_{A_1} and A_1X. Examination of the marginal treatment effect reveals how identification depends on observed-data constraints.

$$\begin{aligned}
\Delta_x &= \sum_a \left\{ \sum_s \{\text{logit}(E[Y|S_{A_1}, A_1, R=1, X]) - \text{logit}(E[Y_0|S_{A_1}, A_1, R=1, X])\} \times \right. \qquad (11) \\
&\quad \left. P(S_{A_1} = s|A_1, R=1, X) \right\} \times P(A_1 = a|X, R=1) \\
&= \sum_a \left\{ \sum_s \pi_{sax}(\psi_{01}a + \psi_{02}as) = \pi_{1ax}(\psi_{01}a + \psi_{02}a) + \pi_{0ax}(\psi_{01}a) \right\} P(A_1 = a|X, R=1) \\
&= 0 + p_{1x}(\psi_{01} + \pi_{11x}\psi_{02}),
\end{aligned}$$

where $p_{1x} \equiv P(A_1 = 1|X, R=1)$ and $\pi_{sax} \equiv P(S_{A_1} = s|A_1 = a, X = x, R=1)$. Equation (11) involves two equations in two unknown quantities, ψ_{01} and ψ_{02}, and has a unique solution so long as X_i predicts p_{1x} or π_{11x} and thus identifiability holds under the model. In sum, model-based identification holds if we restrict the number of interactions between treatment assignment and baseline covariates on the outcome in the structural model. Without this assumption, the model in Equation (10) and the following model

$$\text{logit}\{E(Y|S_{A_1}, A_1, R=1, X_i)\} - \text{logit}\{E(Y_0|S_{A_1}, A_1, R=1, X)\} = \psi_{01}A_1 + \psi_{02}A_1X$$

may fit the data equally well (Vansteelandt and Goetghebeur 2004). The essence of the identifiability problem is that since S_{A_1} is unobserved we are forced to increase the number

of observed-data constraints to identify the model, which is accomplished by enforcing the ignorability criterion within baseline covariate-defined subgroups. In short, we must assume that the complexity of the structural model occurs at a rate slower than the increase in observed-data constraints. Specifically, the treatment effects implied by interactions in the structural model must not outnumber the constraints imposed by (1). A set of assumptions that would allow for nonparametric identification of post-treatment effect modifiers has not yet been identified.

No-interaction assumptions are often used for identification of causal effects. No-interaction assumptions have been invoked with instrumental variable analysis (Hernán and Robins 2006), in the estimation of direct and indirect effects (Robins and Greenland 1992; Ten Have et al. 2007; Vansteelandt 2010), and for other causal analyses (Vansteelandt and Goetghebeur 2004). Under some modeling configurations, we can partially relax this no-interaction assumption as we demonstrate next.

Consider the case where we have two binary covariates X_1 and X_2, and we want to estimate $\boldsymbol{\psi}_0$ in the model

$$\text{logit}\{E[Y|S_{A_1}, A_1, X_1, X_2, R=1]\} - \text{logit}\{E[Y_0|S_{A_1}, A_1, X_1, X_2, R=1]\} = \psi_{01} + \psi_{02}A_1X_1 + \psi_{03}A_1S_{A_1}. \quad (12)$$

This model is similar to (10) but now includes effect modification by a pretreatment covariate. We re-write the ignorability assumption as

$$R_i \perp Y_0|X_1, X_2.$$

In this model, nonparametric identification still does not hold for $\boldsymbol{\psi}_0 = (\psi_{01}, \psi_{02}, \psi_{03})$ since there are more nonzero effects in subgroups defined by joint levels of of S_{A_1}, X_1, X_2 than identifying restrictions. Under our parametric model, however, the same argument as above can be applied to establish model-based identifiability. The 3–dimensional parameter $\boldsymbol{\psi}_0$ can

be identified using the additional information provided by adding X_2 and its ignorability assumption. That is, the parameters in model (12) may be identified potentially assuming no interactions involving X_2, but not X_1. However, the parameters in the following model

$$\begin{aligned}\text{logit}\{E[Y|S_{A_1}, A_1, R=1, X_1, X_2]\} - \text{logit}\{E[Y_0|S_{A_1}, A_1, R=1, X_1, X_2]\} = \\ \psi_{01}A_1 + \psi_{02}A_1X_1 + \psi_{03}A_1S_{A_1} + \psi_{04}A_1X_2 + \psi_{05}A_1X_1X_2 \\ + \psi_{06}AS_{A_1}X_1X_2 + \psi_{07}A_1S_{A_1}X_1 + \psi_{08}A_1S_{A_1}X_2 \quad (13)\end{aligned}$$

cannot not be identified, nor can any other model with parameter $dim(\boldsymbol{\psi}_0) \geq 4$. As in the case of the bivariate parameter above, multiple models with $\boldsymbol{\psi}_0$ of the same dimension may fit the data equally well. Therefore, the no-interaction assumption is required for some, but not all of the possible baseline covariate-treatment interactions on outcomes and will depend on the richness of the available data. This result suggests a focus on a limited number of pre-treatment effect modifiers that are considered to be the most relevant. In the JOBS II data, for example, we would want to focus on baseline depression as the critical pretreatment effect modifier given its substantive relevance. However, for pre-treatment covariates such as age or gender, we would enforce the no-interaction assumption. Finally, measures can be taken to reduce the likelihood of a violation of the no-interaction assumption by limiting heterogeneity in selected subjects, as was done in JOBS II. In general, we believe the model is still useful so long as the estimates are given an exploratory rather than confirmatory interpretation.

Given the model-based identification for effect modification by S_{A_1}, it remains to be shown that the RHS of model (4) may be expressed in terms of observed data. Under our stated assumptions, the following set of equalities hold:

$$\begin{aligned}P(Y_{A_1}=1|S_{A_1}, A_1, \mathbf{X}) = P(Y_{A_1}=1|S_{A_1}, A_1, R=1, \mathbf{X}) \\ = P(Y=1|S, A, R=1, \mathbf{X}).\end{aligned}$$

The ignorability assumption justifies the first equality (1), while the second holds due to consistency. Similarly due to ignorability, we note that $P(Y_0|S_{A_1}, A_1, R = 1, \mathbf{X}) = P(Y_0|S_{A_1}, A_1, \mathbf{X})$. Under this equivalence, we would write the causal model in (4) as:

$$\begin{aligned}\text{logit}\{E[Y|S, A, R = 1, \mathbf{X}]\} - \text{logit}\{E[Y_0|S, A, R = 1, \mathbf{X}]\} &= \eta_s'(A, S)\boldsymbol{\psi}_0 \\ &= f_1(A; \psi_{01}) + f_2(A, S; \psi_{02}), \quad (14)\end{aligned}$$

where we condition on the observed values A and S as opposed to conditioning on the potential outcomes A_1 and S_{A_1}.

3.3 Post-treatment Effect Modification within the Principal Stratification Framework

Next, we further examine our structural model for post-treatment effect modification within the framework of principal stratification (Frangakis and Rubin 2002). Principal stratification is a popular approach for thinking about certain classes of causal effects, particularly when analysts condition on post-treatment quantities. A principal stratification with respect to a post-treatment variable is a partition of units into latent classes defined by the joint potential values of that post-treatment variable under each of the treatments being compared (Mealli and Mattei 2012). The PS framework often provides useful insights into causal estimands based on post-treatment variables, and we use it to clarify the estimands of interest. Both noncompliance and post-treatment effect modification have been written in the PS framework as separate concepts. Here, we consider them jointly. We should note in advance that in our example the estimands are equivalent under the SMM and PS frameworks and the identification assumptions are identical.

To fully characterize our estimand under the principal stratification approach, we consider the cases of noncompliance and post-treatment effect modification separately. Under noncompliance, our estimand is identical to the principal stratification estimand in that

there are four principal strata of always-takers, never-takers, defiers, and compliers (Angrist et al. 1996; Frangakis and Rubin 2002). In the PS framework, defiers are ruled out via the monotonicity assumption. Here, the no-contamination restriction that we adopt is a strong form of the usual monotonicity assumption and thus serves an equivalent role (Clarke and Windmeijer 2010). That is, the no-contamination restriction rules out the presence of both defiers and always-takers which allows us to identify the other two strata in the observed data. Under the PS framework, to identify causal effects, we must also assume the exclusion restriction holds, but we have already stipulated the exclusion restriction under our stated assumptions. Under noncompliance, the PS estimand is often referred to as the local average treatment effect (LATE) or the complier average causal effect (CACE). The SMM estimand is also a local estimand under the no-contamination restriction (Clarke and Windmeijer 2012).

Next, we characterize post-treatment effect modification using the PS framework. For the moment, we ignore compliance, and thus we denote potential levels of depression as S_r. If S is binary, as we have defined it, there are four principal strata defined by the joint levels of S_1 and S_0. Following, Hsu and Small (2014) we characterize these four basic strata as: 'always-high' ($S_1 = 1$ and $S_0 = 1$), 'never-high' ($S_1 = 0$ and $S_0 = 0$), 'treatment positively affected' ($S_1 = 1$ and $S_0 = 0$), and 'treatment negatively affected' ($S_1 = 0$ and $S_0 = 1$). Our estimand considers patient classes defined by levels of S_1. Thus to compare our estimand to that under PS, we consider the union principal strata defined by $S_1 = 1$ and $S_1 = 0$, taking the union of the 'always-high' and 'treatment positively affected' and separately the 'never-high' and 'treatment negatively affected' strata. If we have both noncompliance and post-treatment effect modification as in JOBS II, this implies that there are generally eight principal strata since within each of the four noncompliance principal strata, we have two effect modification union principal strata. However, under the no contamination restriction, we assume that effects are only identifiable for the two effect modification principal strata within the complier strata.

4 Estimation

We use the Vansteelandt and Goetghebeur (2003) method for the estimation of causal effects under generalized structural mean models with binary outcomes using the logit link. This estimation strategy was developed as a solution to Robins (1999), which showed that the causal odds ratio could not be estimated using the same G-estimation procedure as used for identity and log links in the presence of high dimensional covariates. To facilitate the definition of mean treatment-free outcomes used in this modified version of G-estimation, the first stage of a two stage model is an association model among subjects randomized to the job search seminar treatment. A detailed argument motivating the need for the association model is described in Vansteelandt and Goetghebeur (2003) and largely stems from the noncollapsibility of the logit link.

The first stage model is defined as

$$\text{logit}(E(Y|S, A, R = 1, \mathbf{X}; \beta)) = \eta_a(S, A, \mathbf{X}; \beta), \tag{15}$$

for a known function η_a and unknown finite-dimensional parameter vector β, and predicted mean treatment-free outcomes are constructed as

$$H(\boldsymbol{\psi}) = \text{expit}[\eta_a(S, A, \mathbf{X}; \beta)] - [f_1(A; \psi_1) + f_2(A, S; \psi_2)] \tag{16}$$

for subjects randomized to treatment, where $f_1(A; \psi_1)$ and $f_2(A, S; \psi_2)$ are defined as in model (10) and S represents intermediate depression at either 6 weeks or 6 months. The subscript a distinguishes the association model from the structural, causal model. The functional form for $\eta_a(S, A, \mathbf{X}; \beta)$ is generally linear in the parameters but may include main effects, interactions, and higher order terms for the variables. For subjects randomized to control and hence unable to access treatment, the observed outcome Y equals the treatment free outcome Y_0 following the consistency assumption. In the control arm, it is therefore

unnecessary to estimate $H(\boldsymbol{\psi})$ by removing the treatment effect from a model-predicted mean outcome; $H(\boldsymbol{\psi})$ is simply set to $H(\boldsymbol{\psi}) = Y$.

The second stage of estimation then defines the estimating function

$$U(S, A, R, \mathbf{X}, H(\boldsymbol{\psi})) = d(\mathbf{X}, R)\left[H(\boldsymbol{\psi}) - q(X)\right], \tag{17}$$

for $d(\mathbf{X}, R)$, a $p-$dimensional weight function defined such that its elements $d_1(\mathbf{X}, R), \ldots d_p(\mathbf{X}, R)$ are non-collinear, and $q(\mathbf{X})$, a function of baseline covariates. The causal parameter $\boldsymbol{\psi}$ is estimated as the solution to $\sum_{i=1}^{n} U(S, A, R, \mathbf{X}, H(\boldsymbol{\psi})) = \mathbf{0}$. By the randomization assumption, under the true ψ and β, $E\left[\sum_{i=1}^{n} U\{S, A, R, \mathbf{X}, H(\boldsymbol{\psi})\}\right] = \mathbf{0}$ for $U(S, A, R, \mathbf{X}, H(\boldsymbol{\psi}))$ as defined in (17). The chosen $d(\mathbf{X}, R)$ and $q(\mathbf{X})$ affect efficiency but not bias in the resulting estimate $\hat{\psi}$ when the association model is correctly specified. Under a misspecified association model, the robust estimating equations referenced above are constructed through strategic selection of $d(\mathbf{X}, R)$ and guarantee that type I error is preserved. Term $q(\mathbf{X})$ does not affect bias under correct or incorrect specification of the association model. Vansteelandt and Goetghebeur (2003) recommend the choice $d(\mathbf{X}, R) = \frac{d^*(\mathbf{X})(-1)^R}{R*P(R=1|\mathbf{X})+(1-R)(1-P(R=1|\mathbf{X}))}$, with $d^*(\mathbf{X}) = E\left[\frac{\partial H(\boldsymbol{\psi})}{\partial \psi}|\mathbf{X}\right]$, following semiparametric optimality arguments for their GSMM under known β. Under the logit model $\frac{\partial H(\boldsymbol{\psi})}{\partial \psi} = H(\boldsymbol{\psi})(1 - H(\boldsymbol{\psi}))[R/RS]$ where $H(\boldsymbol{\psi})$ is evaluated at an initial estimate of ψ, $\hat{\psi}_b$.

Under the null hypothesis $\psi_0 = 0$, one may construct locally robust estimating equations that yield consistent inference under misspecified association models. For variance estimation the sandwich variance estimator that jointly considers estimation of components of the structural and association models is used, thus taking into account the estimation of the association model parameters (Vansteelandt and Goetghebeur 2003).

5 Simulation Study

A simulation study was conducted to evaluate the proposed estimator for assessing effect modification by post-treatment variables while also accounting for noncompliance. A second set of simulations which we show in the Appendix displays the results of a simulation study for using this approach to evaluate post-treatment effect modification under full compliance. These additional simulations also explore the impact of misspecification of the association model.

For each subject, independent baseline covariates $X_1 \sim Bernoulli(p = 0.4)$ and $X_2 \sim Bernoulli(p = 0.7)$ were generated. Binary treatment R was simulated following an unstratified randomization design, with $R \sim Bernoulli(p = 0.5)$. A compliance variable A was generated following the model $\text{logit}(P(A = 1|X_1)) = \text{logit}(0.9) - 3X_1$. For subjects with $R = 0$, we set A=0 following the setting where subjects randomized to control are unable to access active treatment. Under this design, compliance was approximately 66% among those randomized to treatment. The post-treatment variable S was simulated from the model $\text{logit}(P(S = 1|A, R, \mathbf{X})) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 A$, with $\gamma = (\text{logit}(0.2), 1.0, 1.0, \gamma_3)$, with γ_3 =0, 0.3, or 1.2 for S_i not associated, weakly associated, or strongly associated with A. The data design contained binary covariates to ensure compatibility among various stages of conditional mean treatment-free potential outcome models. Each simulated dataset consisted of n=5,000 subjects.

Outcomes were generated under a likelihood consistent with the Retrospective Structural Mean Model (14), which conditions on observed posttreament data and was shown to be equivalent to model (4) that conditions on the potential intermediates using a modification of the strategy described in Robins and Scharfstein (1999). The conditional mean of Y_0 given baseline covariates was defined as $\text{logit}(E[Y_0|X_1, X_2]) = \rho_0^X + \rho_1^X X_1 + \rho_2^X X_2 + \rho_3^X X_1 X_2$, with $\rho^X = (\text{logit}(0.35), 0.8, -0.8, 1.5)$. Conditional mean treatment-free potential outcomes were then adjusted for S, by setting $E[Y_0|S, R, \mathbf{X}] = \text{expit}[\text{logit}(E[Y_0|X_1, X_2]) + \rho_1^S S + \rho_{0_j}^S (1 - S)]$,

where $\rho_1^S = 0.4$, and for given ρ^X, γ and values of covariates R, X_1, X_2, $\rho_{0_j}^S$ was the solution to $E_S[E[Y_0|S, R, \mathbf{X}_i]] = E[Y_0|\mathbf{X}_i]$ for $j = 1, ..., 8$, indexing unique profiles of R, X_1, X_2. Under ignorable noncompliance we set $E[Y_0|S_i, A_i, R_i, \mathbf{X}_i] = E[Y_{i,0}|S_i, R_i, \mathbf{X}_i]$ and simulated observed outcomes as $Bernoulli(p_Y)$, where $\text{logit}(p_y) = \text{logit}(E[Y_0|S, A, R, \mathbf{X}]) + \psi_1 A + \psi_2 AS$ for $\boldsymbol{\psi_0}$=(0.5,-0.5). Under non-ignorable non-compliance $E[Y_0|S, A, R, \mathbf{X}]$ was generated by adjusting the conditional treatment-free mean $E[Y_0|R, \mathbf{X}]$ for compliance A and then S while maintaining compatibility across conditional means. Observed outcomes were then generated after incrementing mean treatment-free outcomes for the effect of observed compliance to treatment, also with $\psi = (0.5, -0.5)$.

The application of the two-stage GSMM considered the association model fully saturated for S, A, X_1, X_2. The GSMM estimates were compared to estimates from two logistic regression models: model 1 was similar to an intent-to-treat analysis with the addition of post-intervention covariates and contained terms R and RS with full saturation for S, X_1, X_2; and model 2 was an as-treated approach, including A and AS, also with full saturation for S, X_1, X_2. All results were based on 1,000 replicated datasets.

Tables 1 and 2 contain detailed results from simulations across the several described scenarios with Table 1 featuring ignorable noncompliance and Table 2 demonstrating nonignorable non-compliance. Table 3 also features nonignorable noncompliance but differs from 2 in the weak relationship between baseline covariates X_1,X_2, and S_i. The intent-to-treat style analysis using logistic regression demonstrated substantial bias of at least 28% in all settings, and up to nearly 200% bias when treatment strongly predicted the intermediate variable. When the intermediate was not affected by treatment or associated with the outcome, estimates were generally attenuated compared to the true value for both ignorable and non-ignorable noncompliance. Moderate bias (9-18%) was observed for the GSMM when treatment was a weak predictor of the intermediate variable (see Table 2). Additional simulations showed that this bias is removed by considering larger sample sizes or by increasing the predictiveness of baseline covariates for the intermediate variable. Results of

Table 1: **Simulation Study Results.** Mean estimates, percent bias, and Monte Carlo standard deviations of the modified G-estimation and standard logistic regression when ignorable noncompliance is present. ρ^X=(characterizes the association between baseline covariates $\mathbf{X}$, and (γ_1, γ_2) characterize the association of $\mathbf{X}$ and S. γ_3 is the coefficient of A in the data-generating model for S. $\rho = 0$ indicates that S does not predict Y_0. The first row in each parameter configuration corresponds to the G-estimation; the second row reports estimates from the intent-to-treat logistic regression; the third row is the as-treated logistic regression.

$\boldsymbol{\psi}_0 = (0.5, -0.5)$, ρ^X=(logit(0.35),0.8,0.8,1.5), $(\gamma_0, \gamma_1, \gamma_2)$=(logit(0.2),1,1)							
		$\hat{\psi}_1$			$\hat{\psi}_2$		
		Estimate	% Bias	MCSD	Estimate	% Bias	MCSD
$\gamma_3 = 1.2, \rho \neq 0$	GSMM	0.53	5.67	0.43	-0.53	5.03	0.55
	ITT Log. Reg.	-0.48	-195.87	0.11	0.48	-195.63	0.14
	AT Log. Reg.	-0.20	-140.21	0.12	0.20	-139.97	0.15
$\gamma_3 = 0.3, \rho \neq 0$	GSMM	0.49	-1.26	0.32	-0.45	-10.10	0.65
	ITT Log. Reg.	0.23	-53.53	0.09	-0.24	-52.68	0.14
	AT Log. Reg.	0.42	-16.36	0.10	-0.42	-15.04	0.15
$\gamma_3 = 0, \rho \neq 0$	GSMM	0.49	-1.46	0.29	-0.43	-13.86	0.71
	ITT Log. Reg.	0.34	-31.49	0.09	-0.35	-30.75	0.14
	AT Log. Reg.	0.50	0.65	0.10	-0.51	1.89	0.15
$\gamma_3 = 1.2, \rho = 0$	GSMM	0.52	4.82	0.50	-0.49	-1.22	0.79
	ITT Log. Reg.	0.37	-26.51	0.10	-0.37	-26.68	0.14
	AT Log. Reg.	0.50	-0.62	0.11	-0.49	-1.06	0.15
$\gamma_3 = 0.3, \rho = 0$	GSMM	0.52	3.04	0.30	-0.48	-4.24	0.71
	ITT Log. Reg.	0.36	-28.22	0.09	-0.36	-28.09	0.13
	AT Log. Reg.	0.50	0.39	0.10	-0.50	0.80	0.15
$\gamma_3 = 0, \rho = 0$	GSMM	0.52	3.52	0.27	-0.48	-3.49	0.74
	ITT Log. Reg.	0.36	-28.93	0.09	-0.36	-28.72	0.13
	AT Log. Reg.	0.50	0.18	0.09	-0.50	0.58	0.15

Table 2: **Simulation Study Results.** Mean estimates, percent bias, and Monte Carlo standard deviations of the modified G-estimation and standard logistic regression when non-ignorable noncompliance is present. ρ^X characterizes the association between baseline covariates $\mathbf{X}$, and (γ_1, γ_2) characterize the association of $\mathbf{X}$ and S. γ_3 is the coefficient of A in the data-generating model for S. $\rho = 0$ indicates that S does not predict Y_0. The first row in each parameter configuration corresponds to the G-estimation; the second row reports estimates from the intent-to-treat logistic regression; the third row is the as-treated logistic regression.

$\boldsymbol{\psi}_0 = (0.5, -0.5)$, ρ^X=(logit(0.35),0.8,0.8,1.5), $(\gamma_0, \gamma_1, \gamma_2)$=(logit(0.2),1,1)							
		$\hat{\psi}_1$			$\hat{\psi}_2$		
		Estimate	% Bias	MCSD	Estimate	% Bias	MCSD
$\gamma_3 = 1.2, \rho \neq 0$	GSMM	0.52	3.61	0.40	-0.51	2.87	0.52
	ITT Log. Reg.	-0.19	-137.65	0.09	0.23	-145.79	0.13
	AT Log. Reg.	0.07	-86.34	0.12	0.25	-149.03	0.15
$\gamma_3 = 0.3, \rho \neq 0$	GSMM	0.49	-2.68	0.31	-0.43	-13.53	0.66
	ITT Log. Reg.	0.26	-47.24	0.09	-0.26	-48.42	0.13
	AT Log. Reg.	0.68	36.30	0.10	-0.37	-25.40	0.15
$\gamma_3 = 0, \rho \neq 0$	GSMM	0.49	-2.98	0.29	-0.41	-18.57	0.73
	ITT Log. Reg.	0.35	-29.69	0.08	-0.36	-27.70	0.13
	AT Log. Reg.	0.78	55.88	0.10	-0.47	-5.83	0.15
$\gamma_3 = 1.2, \rho = 0$	GSMM	0.52	3.18	0.51	-0.48	-3.62	0.80
	ITT Log. Reg.	0.23	-53.11	0.09	-0.19	-62.66	0.13
	AT Log. Reg.	0.73	45.76	0.11	-0.38	-23.26	0.15
$\gamma_3 = 0.3, \rho = 0$	GSMM	0.50	0.97	0.30	-0.45	-9.44	0.73
	ITT Log. Reg.	0.33	-33.46	0.09	-0.32	-36.15	0.13
	AT Log. Reg.	0.75	49.14	0.10	-0.41	-18.87	0.15
$\gamma_3 = 0, \rho = 0$	GSMM	0.51	1.46	0.27	-0.45	-9.83	0.77
	ITT Log. Reg.	0.35	-29.04	0.08	-0.36	-28.60	0.13
	AT Log. Reg.	0.75	49.89	0.09	-0.41	-18.05	0.15

Table 3: **Simulation Study Results.** Mean estimates, percent bias, and Monte Carlo standard deviations of the modified G-estimation and standard logistic regression when nonignorable noncompliance is present and baseline covariates are weakly predictive of post-baseline effect modifier. ρ^X characterizes the association between baseline covariates $\mathbf{X}$, and (γ_1, γ_2) characterize the association of $\mathbf{X}_i$ and S. γ_3 is the coefficient of A in the data-generating model for S. $\rho = 0$ indicates that S does not predict Y_0. The first row in each parameter configuration corresponds to the G-estimation; the second row reports estimates from the intent-to-treat logistic regression; the third row is the as-treated logistic regression.

$\boldsymbol{\psi}_0 = (0.5, -0.5)$, ρ^X=(logit(0.35),0.8,0.8,1.5), $(\gamma_0, \gamma_1, \gamma_2)$=(logit(0.2),0.2,0.4)							
		$\hat{\psi}_1$			$\hat{\psi}_2$		
		Estimate	% Bias	MCSD	Estimate	% Bias	MCSD
$\gamma_3 = 1.2, \rho \neq 0$	GSMM	0.44	-12.84%	0.67	-0.27	-46.53%	1.25
	ITT Log. Reg.	0.11	-78.58%	0.08	-0.05	-90.08%	0.14
	AT Log. Reg.	0.48	-4.01%	0.10	-0.24	-52.72%	0.16
$\gamma_3 = 0.3, \rho \neq 0$	GSMM	0.45	-9.74%	0.40	-0.18	-64.84%	1.30
	ITT Log. Reg.	0.26	-47.10%	0.07	-0.26	-48.97%	0.14
	AT Log. Reg.	0.66	31.49%	0.09	-0.45	-10.78%	0.16
$\gamma_3 = 0, \rho \neq 0$	GSMM	0.46	-8.25%	0.34	-0.10	-80.45%	1.44
	ITT Log. Reg.	0.31	-37.11%	0.07	-0.32	-35.35%	0.15
	AT Log. Reg.	0.71	41.40%	0.08	-0.49	-1.24%	0.17
$\gamma_3 = 1.2, \rho = 0$	GSMM	0.41	-18.78%	0.60	-0.20	-59.99%	1.21
	ITT Log. Reg.	0.23	-53.79%	0.07	-0.16	-67.59%	0.14
	AT Log. Reg.	0.68	35.20%	0.10	-0.41	-18.36%	0.16
$\gamma_3 = 0.3, \rho = 0$	GSMM	0.45	-9.90%	0.36	-0.13	-74.58%	1.28
	ITT Log. Reg.	0.30	-39.35%	0.07	-0.29	-42.56%	0.15
	AT Log. Reg.	0.69	38.61%	0.08	-0.45	-10.57%	0.17
$\gamma_3 = 0, \rho = 0$	GSMM	0.46	-7.96%	0.32	-0.07	-86.15%	1.42
	ITT Log. Reg.	0.32	-36.41%	0.07	-0.32	-35.86%	0.15
	AT Log. Reg.	0.70	39.29%	0.08	-0.46	-8.55%	0.17

these additional simulations are shown in the appendix. The as-treated analytic approach was consistent and more efficient than the GSMM under ignorable non-compliance when the intermediate was not affected by treatment or not associated with the outcome but exhibited substantial bias under all scenarios considering non-ignorable non-compliance. The final set of simulation results shown in Table 3 shows that the modified G-estimation can behave poorly when there are no strong predictors of the intermediate covariate among baseline covariates. In this scenario the main effect estimate $\hat{\psi}_1$ was biased by $8\% - 19\%$, and $\hat{\psi}_2$ was even more biased at $46\% - 86\%$.

6 Post-treatment Effect Modification in JOBS II

In this section, we analyze the data from JOBS II. We first present the results based on the double logistic GSMM, which allows the treatment effect estimates to vary as a function of intermediate depression levels under treatment. We restrict the analysis to the subset of the subjects for which depression levels and the re-employment outcome are fully observed at all follow up periods. We condition on a large set of pretreatment covariates that were measured in the JOBS II study. We use pretreatment covariates to specify the association model, which models the observed outcomes. The pretreatment covariates include binary indicators for seven categories of occupation type, sex, marital status, whether the subject was nonwhite, years of education, income, age, a measure of financial strain, and depression at baseline.

We begin with an analysis that accounts for noncompliance, but does not adjust for post-treatment effect modification. An analysis based on the double-logistic GSMM shows that the odds ratio of success for participating in the job training seminars versus not participating is 1.83 with a corresponding 95% confidence interval (1.17, 2.87). This estimate implies that the odds of being employed are 83% higher among those who attend the JOBS II training seminars.

In the JOBS II study, depression levels were measured six weeks and six months after

subjects completed the training sessions which comprised the intervention. We conduct separate analyses for the two intermediate follow-up periods. In the first analysis, the causal effect of being exposed to the treatment is potentially modified by depression levels at six weeks, and in the second analysis the effect of the intervention is potentially modified by depression levels at six months. The two separate analyses allow us to understand whether the magnitude of effect modification varies over time. We found that model convergence was somewhat sensitive to specification of the association model. In particular, we found that when we failed to condition on depressive symptoms at baseline estimates either became so large as to signal a lack of convergence or convergence failed outright. This was consistent with our simulation study that showed poor behavior with weak baseline correlates of potential post-treatment modifiers. Specifications that condition on a larger set of baseline covariates also did little to aid precision of the model estimates. We compare the GSMM estimates to estimates from logistic regression. We use the same covariates in the specification of the logistic regression model.

Table 4 contains estimates for the two causal parameters, ψ_{01} and ψ_{02}, under two-stage G-estimation and logistic regression. The GSMM causal estimates (robust standard errors are in parenthesis) for the ψ_{02} parameter: -0.05 (0.05) at the six week follow-up and -0.01 (0.04) at the six month follow-up. Estimates of the ψ_{02} parameter from logistic regression are much smaller in comparison: 0.004 (0.008) at six weeks and -0.007 (0.007) at six months. The bias we observe in the simulations when logistic regression is applied appears to be present in this application as well.

The parameter estimates in Table 4 do not readily convey the dependence of the effect of job training on the intermediate depression modifier, since the parameter estimates cannot fully convey how the treatment effect may vary across levels of depression. Specifically, conditional effects may be bound away from zero for some values of the effect modifier, even if the interaction effect is itself statistically insignificant (Franzese and Kam 2009). We next explore in more detail how post-treatment levels of depression modify the effect of the JOBS

Table 4: **Empirical Analysis.** Estimates are presented for the log-odds ratio causal effect parameter ψ_1 and post-treatment effect modification ψ_2 under 2 different approaches: (i) two-stage G-estimation; and (ii) logistic regression. All method condition on the same set of pretreatment covariates. Standard errors are in parentheses.

	Depression at Six Weeks		Depression at Six Months	
	$\hat{\psi}_1$	$\hat{\psi}_2$	$\hat{\psi}_1$	$\hat{\psi}_2$
GSMM	1.45	-0.05	0.73	-0.01
	(0.73)	(0.05)	(0.67)	(0.04)
Logistic Regression	0.06	0.004	0.26	-0.007
	(0.21)	(0.008)	(0.21)	(0.007)

II intervention. Here, we use the measure of depressive symptoms from the six week follow-up with the parameter estimates from GSMM. We calculate the causal odds ratio and an associated 95% confidence interval for the intervention conditional on levels of the depression scale. We plot the pattern of effect modification for quartiles of 6-week depression in Figure 2, which shows that for some values of depression the confidence intervals for the treatment effect are bound away from zero.

In the plot, as depression scores rise the causal odds ratio decreases. In the sample, approximately ten percent of subjects recorded no depressive symptoms. The estimated causal odds ratio for these subjects is 4.29 with an associated 95% confidence interval of (1.05, 16.77). Next we calculate the causal odds ratio for subjects with a score of seven on the depression scale, which represents the 25th percentile. The causal odds ratio is 2.97 with a corresponding 95% confidence interval (1.28, 6.89). When depressive symptoms increase to a score of 16, the median of the depression scale, the causal odds ratio decreases further to 1.90 with 95% confidence interval (1.14, 3.18). The magnitude of the treatment effect is further reduced such that it is not statistically significant for those with higher levels of depression at six weeks. We next used stratification to partially relax the no-interaction assumption. That is, we stratified the sample by baseline depression and re-estimated the model with post-treatment effect modification within the strata. We used the median score

of pretreatment depression to stratify the sample into high and low depression subsamples. Within each of these strata, we fit a GSMM with a specification identical to Table 4. We found that the original pattern of post-treatment effect modification held in the stratified samples.

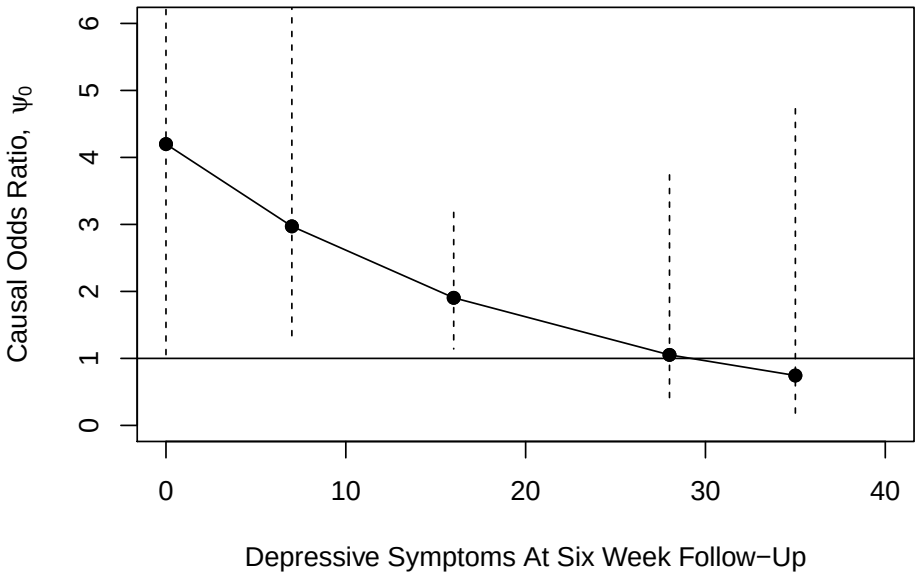

Figure 2: Causal Odds Ratio Effect Modified by Depression Levels at Six Week Follow-Up. The dotted lines represent 95% confidence intervals. Point estimates calculated at the minimum, 25th percentile, median, 75th percentile, and 90th percentile on the depression scale distribution.

7 Discussion

We have used GSMMs to estimate causal effects that may be modified by potential intermediates and shown that our models can be equivalently expressed in terms of effect modification by observed posttreatment variables. Our work complements existing literature on noncompliance and mediation where conditioning also occurs on post-treatment variables. One natural comparison is to causal mediation analysis. It would appear that the analysis we have proposed differs substantially from the purpose of a causal mediation analysis. In mediation, the goal is to decompose a treatment effect into direct and indirect components

(Imai et al. 2010b). The indirect treatment effect is an effect mediated by a third variable which transmits the treatment effect to the outcome. Mediation effects were of key interest in other analyses (Vinokur and Schul 1997; Imai et al. 2010a). In contrast, we stipulate only a total effect of the treatment that is conditional on levels of S. More specifically, we are only interested in how S alters effects, but we do not focus on any effect S has on Y. However, the identification conditions we invoke can, in another form, be used to identify mediation effects Small (2011).

Our analysis has focused on a binary treatment. For treatments with more than two levels, the analysis may be extended by fitting a separate association model for each level of treatment, and defining $H(\psi)$ for each subject by subtracting off the parametrized effect of the subjects' observed treatment according to the proposed structural model. Restrictions on the effects of various treatment levels may be enforced through the parametrization of ψ in the blip down function from each treatment arm.

One weakness of this approach is its dependence on the specification of the association model. When the associated model is nonsaturated, it can be incompatible or uncongenial to the logistic SMM (Robins and Rotnitzky 2004). Vansteelandt et al. (2011) argue that the biases from uncongenial estimators are small compared with other assumption failures. Moreover, alternatives are computationally demanding. Robust weights may be used to provide valid testing in the absence of treatment effects, but estimation of treatment effects may be subject to bias under the alternative. Moreover, in data analysis, nonconvergence was observed when baseline depression, a covariate that was highly predictive of the intermediate variable 6-week or 6-month depression, was omitted from the auxiliary model. The implication of this for practitioners is that model fitting of the association model should be completed carefully, with careful attention to functional form and the potential presence of interaction. Additional methodology to enhance robustness is one potential area for further research.

Identification of post-treatment effect modification may also be a useful tool in the de-

velopment of "adaptive treatment strategies." Under an adaptive treatment strategy, the treatment level and type are adjusted according to individual level characteristics (Murphy 2005; Lavori and Dawson 2000; Almirall et al. 2012; Robins 2004; Murphy 2003; Collins et al. 2007). The design of adaptive treatment strategies requires choosing tailoring variables, variables that are used to decide how to adapt the treatment to specific individuals. Post-treatment effect modification provides one method for identification of tailoring variables. If the effect of a treatment varies across levels of a post-randomization variable, this would suggest that this covariate may be a good tailoring variable. Thus models where post-treatment covariates are allowed to modify causal effect estimates could be used for further actions within a study or to tailor clinical decision-making.

References

Almirall, D., Compton, S. N., Gunlicks-Stoessel, M., Duan, N., and Murphy, S. A. (2012), "Designing a Pilot Sequential Multiple Assignment Randomized Trial for Developing an Adaptive Treatment Strategy," *Statistics in Medicine*, 31, 1887–1902.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Clarke, P. S. and Windmeijer, F. (2010), "Identification of Causal Effects on Binary Outcomes Using Structural Mean Models," *Biostatistics*, 11, 756–770.

— (2012), "Instrumental Variable Estimators for Binary Outcomes," *Journal of the American Statistical Association*, 107, 1638–1652.

Collins, L. M., Murphy, S. A., and Strecher, V. (2007), "The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions," *American journal of preventive medicine*, 32, S112–S118.

Cuzick, J., Sasieni, P., Myles, J., and Tyrer, J. (2007), "Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination," *Journal of The Royal Statistical Society, Series B*, 69, 565–588.

Dunn, G. and Bentall, R. (2007), "Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments)," *Statistics in Medicine*, 26, 4719–4745.

Follmann, D. (2006), "Augmented designs to assess immune response in vaccine trials," *Biometrics*, 62, 1161–1169.

Frangakis, C. A. and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Franzese, R. and Kam, C. (2009), *Modeling and interpreting interactive hypotheses in regression analysis*, University of Michigan Press.

Hernán, M. A. and Robins, J. M. (2006), "Instruments for Causal Inference: An Epidemiologists Dream," *Epidemiology*, 17, 360–372.

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

Hsu, J. Y. and Small, D. S. (2014), "Discussion on "Dynamic Treatment Regimes: Technical Challenges and Applications," *Electronic Journal of Statistics*, Forthcoming.

Imai, K., Keele, L., and Tingley, D. (2010a), "A General Approach to Causal Mediation Analysis," *Psychological Methods*, 15, 309–334.

Imai, K., Keele, L., and Yamamoto, T. (2010b), "Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Jo, B. (2008), "Causal Inference in Randomized Experiments with Mediational Processes," *Psychological Methods*, 13, 314–336.

Jo, B. and Vinokur, A. (2011), "Sensitivity Analysis and Bounding of Causal Effects with Alternative Identifying Assumptions," *Journal of Educational and Behavioral Statistics*, 36, 415–440.

Joffe, M. M., Small, D. S., and Hsu, C.-Y. (2007), "Defining and Estimating Intervention Effects for Groups that will Develop an Auxiliary Outcome," *Statistical Science*, 22, 74–97.

Keele, L. J. (2014), "Conditioning on Posttreatment Quantities with Structural Mean Models," Unpublished Manuscript.

Lavori, P. and Dawson, R. (2000), "A Design for Testing Clinical Strategies: Biased Adaptive Within-Subject Randomization," *Journal of The Royal Statistical Society Series A*, 163, 29–38.

Little, R. J. and Yau, L. H. (1998), "Statistical Techniques for Analyzing Data From Prevention Trials: Treatment of No-Shows Using Rubin's Causal Model," *Psychological Methods*, 3, 147–159.

Mattei, A., Li, F., Mealli, F., et al. (2013), "Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program," *The Annals of Applied Statistics*, 7, 2336–2360.

Mealli, F. and Mattei, A. (2012), "A Refreshing Account of Principal Stratification," *The International Journal of Biostatistics*, 8, 1–37.

Murphy, S. A. (2003), "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 331–355.

Murphy, S. M. (2005), "An Experimental Design for the Development of Adaptive Treatment Strategies," *Statistics in Medicine*, 24, 1455–1618.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5, 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Robins, J. and Greenland, S. (1992), "Identifiability and Exchangeability For Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Robins, J. M. (1994), "Correcting for Non-compliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics-Theory and Methods*, 23, 2379–2412.

— (1997), "Non-response Models for the Analaysis of Non-monotone Non-Ignorable Missing Data," *Statistics in Medicine*, 16, 21–37.

— (1999), "Marginal structural models versus structural nested models as tools for causal inference," in *Statistical Methods in Epidemiology: The Environment and Clinical Trials*, eds. E., H. and Berry, D., New York, NY: Springer-Verlag, p. 95134.

— (2004), "Optimal structural nested models for optimal sequential decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics*, eds. Lin, D. and Hagerty, P., New York: Springer-Verlag, pp. 189–326.

Robins, J. M. and Rotnitzky, A. (2004), "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., Rotnitzky, A., and Scharfstein, D. (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in *Statistical Models in Epidemiology: The Environment and Clinical Trials.*, eds. Halloran, E. and Berry, D., Springer, pp. 1–92.

Robins, J. M. A. R. and Scharfstein, D. (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in *Statistical Models in Epidemiology: The Environment and Clinical Trials.*, eds. Halloran, M. E. and Berry, D., Springer-Verlag, vol. 116, pp. 1–92.

Rosenbaum, P. R. (1984), "The Consequences of Adjusting For a Concomitant Variable That Has Been Affected By The Treatment," *Journal of The Royal Statistical Society Series A*, 147, 656–666.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 6, 688–701.

— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, 6, 34–58.

— (1986), "Which Ifs Have Causal Answers," *Journal of the American Statistical Association*, 81, 961–962.

Schwartz, S., Gatto, N. M., and Campbell, U. B. (2012), "Extending the sufficient component cause model to describe the Stable Unit Treatment Value Assumption (SUTVA)," *Epidemiologic Perspectives & Innovations*, 9, 1–11.

Small, D. S. (2011), "Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables," *Journal of Statistical Research*, 46, 91–103.

Ten Have, T. R., Joffe, 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.

Vansteelandt, S. (2010), "Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models," *Biometrika*, 97, 921–934.

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. (2011), "On Instrumental Variables Estimaton of Causal Odds Ratios," *Statistical Science*, 26, 403–422.

Vansteelandt, S. and Goetghebeur, E. (2003), "Causal Inference with Generalized Structural Mean Models," *Journal of the Royal Statistical Society, Series B*, 65, 817–835.

— (2004), "Using Potential Outcomes as Predictors of Treatment Activity Via Strong Structural Mean Models," *Statistica Sinica*, 14, 907–925.

Vinokur, A., Price, R., and Schul, Y. (1995), "Impact of the JOBS intervention on unemployed workers varying in risk for depression," *American Journal of Community Psychology*, 23, 39–74.

Vinokur, A. and Schul, Y. (1997), "Mastery and Inoculation Against Setbacks as Active Ingredients in the JOBS Intervention for the Unemployed," *Journal of Consulting and Clinical Psychology*, 65, 867–877.