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 11, 2014

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 also compare our approach to principal stratification another commonly used framework for the analysis of post-treatment but intermediate quantities. 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.

*For helpful comments and suggestions, we thank Dylan Small, Edward Kennedy, Stijn Vansteelandt, Eric Tchetgen Tchetgen, and Tepppei Yamamoto.

†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

‡Associate Professor, 211 Pond Lab, Penn State University, University Park, PA 16802, E-mail: ljk20@psu.edu, Phone: 814-863-1592

§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. Randomized experiments are a generally accepted tool for making causal inferences about the effectiveness of such interventions. In a randomized experiment, subjects are randomly assigned to either a treatment group that receives training or to a control group that does not. Randomization ensures that on average the treated and control groups have the same distribution on observed and unobserved pretreatment characteristics.

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. These skills included the preparation of job applications and resumes, how to successfully interview, how to contact potential employers, and how to use social networks to obtain job leads. 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 setbacks. These mental health 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.

Most 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) as one exception). While ITT effects are important, there 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 effective among those who actually attended the job search seminar. To estimate treatment effects among subjects who actually received treatment, we must utilize 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.

Post-treatment effect modification is an important but yet unstudied aspect of JOBS II. In designing the study, effect modification by pretreatment levels of depression was of particular concern. Prior research indicated that significant levels of depression would reduce the effectiveness of the intervention. 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 contemplate that re-employment failures—failed interviews, a lack of call backs—may also increase levels of the depressive symptoms that were thought to reduce the effectiveness of the intervention. 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. We also generate hypotheses about how future studies might identify this subpopulation or target the group for an additional intervention.

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 outline identifiability conditions for our estimand, and rewrite our estimand under the principal stratification (PS) framework, an alternative method for addressing causal inference problems in the presence of post-treatment variables (Frangakis and Rubin 2002), to better clarify interpretation. The PS framework in causal inference has been successfully used in the analysis of other job training applications with complications such as missing data and noncompliance (Zhang et al. 2009; Frumento et al. 2012). 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 the treated (Vansteelandt and Goetghebeur 2003). While G-estimation of GSMM parameters does not require further methodological development, this is the first application of GSMMs to a causal estimand that allows for post-treatment effect modification. A key aspect of our analysis will be exploring identifiability for this novel estimand.

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, and explores linkages to principal stratification. 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 units in the JOB 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. We use all units from the original sample with nonmissing values at baseline and at intermediate data collection time points. Summary statistics for baseline covariates are displayed in Table 1. The table shows means and absolute standardized differences in means (difference in means divided by the pooled standard deviation between groups). Randomization appears to have successfully balanced treated and control groups such that they have the same distribution on observed pretreatment characteristics.

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. 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 items from the Hopkins Symptom Checklist with scores ranging from 0 to 6. A score of 3 or greater on the depression index was considered to be a clinically significant indication of depression. We re-
Table 1: Univariate descriptive statistics for pretreatment covariates by treatment group.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean Treated</th>
<th>Mean Control</th>
<th>Std. Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education (years of schooling)</td>
<td>13.66</td>
<td>13.69</td>
<td>0.02</td>
</tr>
<tr>
<td>Income</td>
<td>7.40</td>
<td>7.20</td>
<td>0.05</td>
</tr>
<tr>
<td>Age</td>
<td>37.18</td>
<td>37.37</td>
<td>0.02</td>
</tr>
<tr>
<td>Depression Score</td>
<td>1.89</td>
<td>1.87</td>
<td>0.03</td>
</tr>
<tr>
<td>Economic Hardship Score</td>
<td>2.98</td>
<td>3.04</td>
<td>0.05</td>
</tr>
<tr>
<td>Female</td>
<td>0.60</td>
<td>0.54</td>
<td>0.11</td>
</tr>
<tr>
<td>Nonwhite</td>
<td>0.15</td>
<td>0.16</td>
<td>0.03</td>
</tr>
<tr>
<td>Professional</td>
<td>0.20</td>
<td>0.18</td>
<td>0.05</td>
</tr>
<tr>
<td>Management</td>
<td>0.18</td>
<td>0.17</td>
<td>0.03</td>
</tr>
<tr>
<td>Sales</td>
<td>0.05</td>
<td>0.08</td>
<td>0.12</td>
</tr>
<tr>
<td>Craftperson</td>
<td>0.10</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>Operatives</td>
<td>0.09</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>Laborer</td>
<td>0.11</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>Risk Score</td>
<td>1.22</td>
<td>1.22</td>
<td>0.01</td>
</tr>
<tr>
<td>N</td>
<td>997</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Std. Diff.= absolute standardized difference.

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 as 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. It is among this subpopulation that the intervention may be less effective.

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. We also outline the assumptions needed for identification of our estimand, since under both non-
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.

compliance and post-treatment effect modification, we condition on post-treatment quantities. Rosenbaum (1984) proved that conditioning on post-treatment covariates may result in biased estimates of the causal parameter without additional assumptions. This result necessitates careful elucidation of the assumptions needed for identification.

We characterize the estimand using potential outcomes and GSMMs. GSMMs 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 focus on a specific generalized structural mean model known as the double-logistic SMM (Vansteelandt and Goetghebeur 2003), which utilizes the logit link function to estimate the causal odds ratio with binary outcomes. We modify the primary GSMM estimand, which focuses on noncompliance to accommodate post-treatment effect modification. We examine the identifiability conditions for post-treatment effect modification in detail since the necessary identification assumptions under noncompliance are well-known. We conclude this section by re-writing our estimand under 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 \(X_i = (X_1, \ldots, X_k)\) are measured at baseline prior to randomization, including age, race, and baseline depression. We define post-treatment effect modifiers, \(S_i\), 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 order of observed variables is \(X_i, R_i, A_i, S_i, Y_i\).

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.

Next, we stipulate a set of assumptions need 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 implies that \( Y_{i,0} \) is equal to \( Y_{i,0,0} \), and we denote \( Y_{i,0} = Y_{i,0,0} \) (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 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

\[
Y_{i,0} \perp R_i | X_i
\]  

Further, we assume the “no-contamination” restriction, which is also justified by the design since control subjects could not access the intervention, so that \( Pr(A_i = 0 | R_i = 0) = 1 \) (Cuzick et al. 2007). Alternatively we could use either a monotonicity assumption (Angrist et al. 1996; Frangakis and Rubin 2002) or the assumption of “no effect modification by randomization” (Hernán and Robins 2006). However, the no-contamination restriction implies that both assumptions hold by design (Clarke and Windmeijer 2010, 2012). We also assume that \( R_i \) has a nonzero causal effect on \( A_i \).

In JOBS II, the primary outcome of interest is binary, so we focus on the logistic SMM (Robins et al. 1999), which models the log-causal-odds of employment vs. unemployment among subjects randomized to treatment as a linear function of exposure and covariates. Our causal estimand which allows the effect of exposure to vary across \( S_i \) in the logistic SMM is as follows:

\[
\logit \{ E[Y_i | S_i, A_i, R_i = 1, X_i] \} - \logit \{ E[Y_{i,0} | S_i, A_i, R_i = 1, X_i] \} = \eta_0'(A_i, S_i) \psi = f_1(A; \psi_{01}) + f_2(A, S_i; \psi_{02}),
\]  

(2)
where \( \psi \) represents the unknown causal parameter, with the truth denoted by \( \psi_0 \), and \( \eta_s(\cdot) \) is a function of covariates \( S_i \) and \( A_i \) with dimension equal to that of \( \psi \). We use the subscript ‘s’ to indicate that this a structural model, since it refers to counterfactual quantities instead of observed associations in the data. In Equation 2 the structural model is comprised of arbitrary known functions \( f_1(\cdot) \) and \( f_2(\cdot) \) up to an unknown \( p \)-dimensional parameter \( \psi_0 = (\psi_{01}, \psi_{02}) \), where \( p = \dim(\psi_{01}) + \dim(\psi_{02}) \). When \( \psi_0 = 0 \) or \( A_i = 0 \) we set \( f_1(\psi_{01}) = f_2(S_i; \psi_{02}) = 0 \) to indicate no effect of treatment, and for \( S_i = 0 \) or \( \psi_{02} = 0 \), we set \( f_2(S_i; \psi_{02}) = 0 \). We condition on \( X_i \) but by assumption we rule out structural models of the form \( \eta_s(A_i, X_i) = (A_i, A_iX_i)' \). This is discussed in further detail in section 3.2. The parameter \( \psi_0 \) characterizes the causal odds ratio

\[
\frac{Pr(Y_i = 1|S_i, A_i, R_i = 1, X_i)}{Pr(Y_i = 0|S_i, A_i, R_i = 1, X_i)} / \frac{Pr(Y_{i0} = 1|S_i, A_i, R_i = 1, X_i)}{Pr(Y_{i0} = 0|S_i, A_i, R_i = 1, X_i)}.
\]

(3)

Given the exclusion restriction, the contrast of interest compares \( Pr(Y_i = 0|A_i = 1, R_i = 1, X_i) \) to \( Pr(Y_{i0} = 1|A_i = 0, R_i = 1, X_i) \), which is the total effect of \( A_i \), self-selection into the job-training sessions, on the outcome re-employment status. We use the term ‘total’ effect to reflect that the effect of \( A_i \) may be mediated through \( S_i \), since under our structural model \( S_i \) is not manipulated.

As an example, consider the case when \( f_1(\psi_{01}) = \psi_{01}A_i \) and \( f_2(\psi_{02}) = \psi_{02}A_iS_i \). 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 \( S_i = 0 \). For \( \psi_{02} \neq 0 \) and \( S_i \neq 0 \), the causal odds ratio for subjects with \( S_i = s \) is captured by \( \exp(\psi_{01} + \psi_{02}s) \), which allows the effect of \( A_i \) to vary with the observed value \( s \).

Equation 2 is an example of a retrospective structural mean model, which is characterized by exposure effects defined conditional on variables observed subsequent to treatment. Unlike standard structural mean models (Robins 1994), which follow a decision-theoretic approach of defining causal effects as a function of covariates observed prior to treatment, retrospective structural mean models allow effects to be defined in subsets of the data not identifiable.
at baseline. This type of model was discussed in Joffe et al. (2007) as single potential stratification, in contrast with PS, which considers stratification on joint potential outcomes. We reserve the discussion of the connections between the estimand for (2) and PS estimands to subsection 3.3. 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 contrast the causal odds ratio with a more familiar one

\[
\frac{Pr(Y_i = 1|A_i, R_i = 1, X_i)}{Pr(Y_i = 0|A_i, R_i = 1, X_i)} / \frac{Pr(Y_{i0} = 1|A_i, R_i = 1, X_i)}{Pr(Y_{i0} = 0|A_i, R_i = 1, X_i)},
\]

which only allows effect modification by \(X_i\). Here, \(\eta(A_i, X_i) = A_i\) implies that the odds of being employed 20 or more hours per week are \(\exp(\psi_{01})\) times higher for subjects with covariates \(X_i\) who chose exposure \(A_i = 1\) than if they selected to not attend the training seminars. Structural models of this type admit the possibility that average causal effects are not constant over different strata of \(X_i\). The distinction between odds ratios (3) and (4) are made clearer in the following section.

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 (4), under a single binary pre-treatment effect modifier $X_i$,

$$\logit\{E(Y_i|A_i, R_i = 1, X_i)\} - \logit\{E(Y_{i,0}|A_i, R_i = 1, X_i)\} = \psi_{01}A_i + \psi_{02}A_iX_i$$  (5)

This model is nonparametrically identified under the assumptions in Section 3.1. The contrasts in Equation (5) are uniquely defined in terms of observable quantities given the equivalence between $E[Y_i|A_i = 0, R_i = 1, X_i]$ and $E[Y_{i,0}|A_i = 0, R_i = 1, X_i]$ under ignorability, the exclusion restriction, the consistency component of SUTVA, and the no-contamination restriction.

We contrast the model in Equation (5) with an example of of Equation (2) as given by:

$$\logit\{E(Y_i|S_i, A_i, R_i = 1, X_i)\} - \logit\{E(Y_{i,0}|S_i, A_i, R_i = 1, X_i)\} = \psi_{01}A_i + \psi_{02}A_iS_i$$  (6)

using a single binary potential post-treatment effect modifier $S_i$. Nonparametric identification for the above model does not hold since there are four possible nonzero effects defined by joint levels of $X_i$, and $S_i$, 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, $X_i$, and post-treatment covariate, $S_i$, is greater than the number of restrictions imposed by our assumptions (Vansteelandt and Goetghebeur 2004).

To achieve model-based identification of effect modification by $S_i$, we place a restriction on the $A_iX_i$ interactions in the causal model in Equation 2. To derive the model-based
identification conditions, we marginalize over $S_i$ and $A_i$:

$$\Delta_x = \sum_a \left\{ \sum_s \{ \text{logit}(E[Y_i|S_i, A_i, R_i = 1, X_i]) - \text{logit}(E[Y_{i,0}|S_i, A_i, R_i = 1, X_i]) \} \times Pr(S_i = s|A_i, R_i = 1, X_i) \right\} \times Pr(A_i = a|X_i, R_i = 1)$$

$$= \sum_a \left\{ \sum_s \pi_{sax}(\psi_{01}a + \psi_{02}as) = \pi_{1ax}(\psi_{01}a + \psi_{02}a) + \pi_{0ax}(\psi_{01}a) \right\} Pr(A_i = a|X_i, R_i = 1)$$

$$= 0 + p_{1x}(\psi_{01} + \pi_{11x}\psi_{02}),$$

where $p_{1x} = Pr(A_i = 1|X_i, R_i = 1)$ and $\pi_{11x} = Pr(S_i = 1|A_i = 1, X_i, R_i = 1)$. Equation (7) involves two equations in two unknown quantities, $\psi_{01}$ and $\psi_{02}$, and has a unique solution so long as $X_i$ predicts $p_{1x}$ or $\pi_{11x}$ and thus identifiability holds under the model. In sum, the no-interaction assumption allows for model-based identification. Without this assumption models (5) and (6) fit the data equally well (Vansteelandt and Goetghebeur 2004). 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). In many cases no-interaction assumptions are untestable, but in the context, here, the identifying assumption does have a testable implication. The analyst can omit post-treatment effect modification, and estimate the model with pretreatment effect modification. If the analyst finds considerable treatment heterogeneity, this provides some evidence against the no-interaction assumption. Moreover, the no-interaction assumption may be partially relaxed when a rich set of covariates is available. Next, we demonstrate this possibility.

Consider the case where we have two binary covariates $X_{i,1}$ and $X_{i,2}$, and we want to
estimate $\psi_0$ in the model

$$\logit\{E[Y_i|S_i, A_i, X_{i,1}, X_{i,2}, R_i = 1]\} - \logit\{E[Y_{i,0}|S_i, A_i, X_{i,1}, X_{i,2}, R_i = 1]\} = \psi_{01} + \psi_{02}A_iX_{i,1} + \psi_{03}A_iS_i.$$  (8)

This model is similar to (6) but now includes effect modification by a pretreatment covariate.

We re-write the ignorability assumption as

$$R_i \perp Y_{i,0}|X_{i,1}, X_{i,2}.$$  

In this model, nonparametric identification still does not hold for $\psi_0 = (\psi_{01}, \psi_{02}, \psi_{03})$ since there are more nonzero effects in subgroups defined by joint levels of $S_i$, $X_{i,1}$, $X_{i,2}$ 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 $\psi_0$ can be identified using the additional information provided by adding $X_{i,2}$ and its ignorability assumption. That is, the parameters in model (8) may be identified potentially assuming no interactions involving $X_{i,2}$, but not $X_{i,1}$. However, the parameters in the following model

$$\logit\{E[Y_i|S_i, A_i, R_i = 1, X_{i,1}, X_{i,2}]\} - \logit\{E[Y_{i,0}|S_i, A_i, R_i = 1, X_{i,1}, X_{i,2}]\} = \
\psi_{01}A_i + \psi_{02}A_iX_{i,1} + \psi_{03}A_iS_i + \psi_{04}A_iX_2 + \psi_{05}A_iX_{i,1}X_{i,2} \
+ \psi_{06}A_iS_iX_{i,1}X_{i,2} + \psi_{07}A_iS_iX_{i,1} + \psi_{08}A_iS_iX_{i,2}$$  (9)

cannot not be identified, nor can any other model with parameter $\text{dim}(\psi_0) \geq 4$. As multiple models with $\psi_0$ 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 interactions 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.

Although nonparametric identification fails for models with post-treatment effect modification, we believe they serve several useful purposes as a form of exploratory analysis. 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 were 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 could also be used as a form of sensitivity analysis in conjunction with other causal models. While the analysis of JOBS II does not lend itself to this form of application, [Keele (2014)] provides one example where post-treatment effect modification is used as sensitivity analysis. Moreover, one could also probe the no-interaction assumption using ancillary data. This would allow for a full exploration of pre-treatment effect modification using a separate data source. 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 argue that the model can still serve a number of useful purposes so long as the estimates are given an exploratory rather than causal interpretation.

3.3 Post-treatment Effect Modification within the Principal Stratification Framework

Next, we develop an alternative characterization of post-treatment effect modification using 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]
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.

Under the PS framework, we write the observed variables $A_i$ and $S_i$ as potential outcomes such that counterfactual values are denoted as $A_{i,r}$ and $S_{i,r}$ under $R_i = r$ as in (Joffe et al. 2007). In both cases, treatment effects are defined for subsets of the population using strata that are not observable at the time of treatment assignment. Based on this redefinition of $A_i$ and $S_i$ as $A_{i,r}$ and $S_{i,r}$, respectively, next we re-write our causal model. Under our stated assumptions, the following set of equalities hold:

$$Pr(Y_i = 1|S_i, A_i, R_i = 1, X_i) = Pr(Y_{i,r = 1} = 1|S_{i,r = 1}, A_{i,r = 1}, R_i = 1, X_i) = Pr(Y_{i,r = 1} = 1|S_{i,r = 1}, A_{i,r = 1}, X_i).$$

The consistency assumption justifies the first equality, while the second holds due to ignorability in (1). Similarly due to ignorability, we note that $Pr(Y_{i,0}|S_{i,r = 1}, A_{i,r = 1}, R_i = 1, X_i) = Pr(Y_{i,0}|S_{i,1}, A_{i,r = 1}, X_i)$. Under this notation, we write the causal model as:

$$\text{logit}\{Pr(Y_{i,1} = 1|S_{i,r = 1}, A_{i,r = 1}, X_i)\} - \text{logit}\{Pr(Y_{i,0} = 1|S_{i,r = 1}, A_{i,r = 1}, X_i)\} = \psi_{01}A_{i,1} + \psi_{02}A_{i,1}S_{i,1}$$

Under this causal model, we condition on the potential auxiliary variables $A_{i,1}$ and $S_{i,1}$ as opposed to conditioning on the observed values of $A_i$ and $S_i$.

Re-writing the causal model in this form, however, does little to characterize the role of the principal strata. 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 defiers which allows us to recover the other three 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. If $S_i$ is binary, there are four principal strata defined by the joint levels of $S_{i,1}$ and $S_{i,0}$. Following, Hsu and Small (2014) we characterize these four strata as: ‘always-high’ ($S_{i,1} = 1$ and $S_{i,0} = 1$), ‘never-high’ ($S_{i,1} = 0$ and $S_{i,0} = 0$), ‘treatment positively effected’ ($S_{i,1} = 1$ and $S_{i,0} = 0$), and ‘treatment negative effected’ ($S_{i,1} = 0$ and $S_{i,0} = 1$). If we have both noncompliance and post-treatment effect modification as in JOBS II, this implies that there are 16 principal strata since within each of the noncompliance principal strata, we have four effect modification principal strata. However, under the identifiability assumptions for noncompliance, we only consider the four effect modification principal strata within the complier strata. Moreover, since $S_{i,1}$ is of primary interest, the effect modification parameter $ψ_{02}$ describes a weighted average of the always-high and treatment positively effected strata.

In this context with a binary treatment, stratification on an observed post-treatment variable is equivalent to conditioning on a set of potential outcomes under randomization to treatment and so is a coarsening of the PS estimand: \( \logit \{ Pr(Y_{i,1} = 1 | S_{i,1}, S_{i,0}, A_{i,1}, A_{i,0}, X_i) \} - \logit \{ Pr(Y_{i,1,0} = 1 | S_{i,1}, S_{i,0}, A_{i,1}, A_{i,0}, X_i) \} \). Thus our estimand can be characterized as either one articulated under principal stratification or as stratification on an observed auxiliary variable (Joffe et al. 2007). However, easy definition of the principal strata depends on $S_i$ being
a binary covariate. If $S_i$ is continuous, as it is in the JOBS II application, elaboration of
the principal strata become cumbersome. As such, we retain the notation of $S_i$ as observed
auxiliary variable.

4 Estimation

For estimation, we use the method developed by Vansteelandt and Goetghebeur (2003) for
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 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

$$\text{logit}(E(Y_i|S_i, A_i, R_i = 1, X_i; \beta)) = \eta_a(S_i, A_i, X_i; \beta),$$

for a known function $\eta_a$ and unknown finite-dimensional parameter vector $\beta$, and predicted
mean treatment-free outcomes are constructed as

$$H_i(\psi) = \exp[\eta_a(S_i, A_i, X_i; \beta)] - \{f_1(A; \psi_1) + f_2(A, S_i; \psi_2)\}$$

for subjects randomized to treatment, where $f_1(A; \psi_1)$ and $f_2(A, S_i; \psi_2)$ are defined as in
model (6) and $S_i$ represents intermediate depression at either 6 weeks or 6 months. The
subscript $a$ distinguishes the association model from the structural, causal model. As in
generalized linear regression models, components of the linear predictor $\eta_a(S_i, A_i, R_i = 1, X_i; \beta)$
may take any functional form, including main effects, interactions, and higher order terms. For subjects randomized to control and hence unable to access treatment, the observed outcome \(Y_i\) equals the treatment free outcome \(Y_{i,0}\) following the consistency assumption. In the control arm, it is therefore unnecessary to estimate \(H_i(\psi)\) by removing the treatment effect from a model-predicted mean outcome; \(H_i(\psi)\) is simply set to \(H_i(\psi) = Y_i\).

The second stage of estimation then defines the estimating function

\[
U(S_i, A_i, R_i, X_i, H_i(\psi)) = d(X_i, R_i) [H_i(\psi) - q(X_i)],
\]

for \(d(X_i, R_i)\), a \(p\)-dimensional weight function defined such that its elements \(d_1(X_i, R_i), ... d_p(X_i, R_i)\) are non-collinear, and \(q(X_i)\), a function of baseline covariates. The causal parameter \(\psi\) is estimated as the solution to \(\sum_{i=1}^{n} U(S_i, A_i, R_i, X_i, H_i(\psi)) = 0\). By the randomization assumption, under the true \(\psi\) and \(\beta\), \(E\left[\sum_{i=1}^{n} U\{S_i, A_i, R_i, X_i, H_i(\psi)\}\right] = 0\) for \(U(S_i, X_i, H_i(\psi))\) as defined in (12). The chosen \(d(X_i, R_i)\) and \(q(X_i)\) 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 \(d(X_i, R_i)\) and guarantee that type I error is preserved. Term \(q(X_i)\) does not affect bias under correct or incorrect specification of the association model. Vansteelandt and Goetghebeur (2003) recommend the choice \(d(X_i, R_i) = d^*(X_i)\) with

\[
d^*(X_i) = E\left[\frac{\partial H_i(\psi)}{\partial \psi}\right|X_i],
\]

following semiparametric optimality arguments for their GSMM under known \(\beta\). Under the logit model \(\frac{\partial H_i(\psi)}{\partial \psi} = H_i(\psi)(1 - H_i(\psi))[R_i R_i S_i]\) where \(H_i(\psi)\) is evaluated at an initial estimate of \(\psi\), \(\hat{\psi}_b\). 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 uncertainty in 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 $X_{i1} \sim Bernoulli(p = 0.4)$ and $X_{i2} \sim Bernoulli(p = 0.7)$ were generated. Binary treatment $R_i$ was simulated following an unstratified randomization design, with $R_i \sim Bernoulli(p = 0.5)$. A compliance variable $A_i$ was generated following the model $\text{logit}(P(A_i = 1|X_{i1})) = \text{logit}(0.9) - 3X_{i1}$. For subjects with $R_i = 0$, we set $A_i = 0$ 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 $S_i$ was simulated from the model $\text{logit}(P(S_i = 1|R_i, X_i)) = \gamma_0 + \gamma_1X_{i1} + \gamma_2X_{i2} + \gamma_3A_i$, with $\gamma = (\text{logit}(0.2), 1.0, 1.0, \gamma_3)$, with $\gamma_3 = 0, 0.3, \text{ or } 1.2$ for $S_i$ not associated, weakly associated, or strongly associated with $A_i$. The data design contained binary covariates to ensure congeniality among various stages of conditional mean treatment-free potential outcome models. Each simulated dataset consisted of $n=5,000$ subjects.

Outcomes were generated following the strategy described in [Robins and Scharfstein (1999)]. The conditional mean of $Y_0$ given baseline covariates was defined as $\text{logit}(E[Y_{i,0}|X_{i1}, X_{i2}]) = \rho_0^X + \rho_1^X X_{i1} + \rho_2^X X_{i2} + \rho_3^X X_{i1}X_{i2}$, with $\rho^X = (\text{logit}(0.35), 0.8, -0.8, 1.5)$. Conditional mean treatment-free potential outcomes were then adjusted for $S_i$, by setting $E[Y_{i,0}|S_i, R_i, X_i] = \exp[\text{logit}(E[Y_{i,0}|X_{i1}, X_{i2}]) + \rho_1^S S_i + \rho_0^S (1 - S_i)]$, where $\rho_1^S = 0.4$ For given $\rho^X$, $\gamma$ and values of covariates $R, X_{i1}, X_{i2}$, $\rho_0^S$ was the solution to $E_S[E[Y_{i,0}|S_i, R_i, X_i]] = E[Y_{i,0}|X_i]$ for $j = 1, ..., 8$, indexing unique profiles of $R, X_{i1}, X_{i2}$. Under ignorable noncompliance we set
\[ E[Y_{i,0}|S_i, A_i, R_i, X_i] = E[Y_{i,0}|S_i, R_i, X_i] \] and simulated observed outcomes as \( Bernoulli(p_Y) \), where \( \logit(p_Y) = \logit(E[Y_{i,0}|S_i, A_i, R_i, X_i]) + \psi_1 R_i + \psi_2 A_i S_i \) for \( \psi=(0.5,-0.5) \). Under non-ignorable non-compliance \( E[Y_{i,0}|S_i, A_i, R_i, X_i] \) was generated by adjusting the conditional treatment-free mean \( E[Y_{i,0}|S_i, R_i, X_i] \) for compliance \( A_i \) and then \( S_i \) while maintaining congeniality across conditional means. Observed outcomes were then generated after incrementing mean treatment-free outcomes for the effect of observed compliance to treatment as for the ignorable noncompliance, also with \( \psi = (0.5, -0.5) \).

The application of the two-stage GSMM considered the association model fully saturated for \( S_i, A_i, X_{i1}, X_{i2} \). 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 \( R_i \) and \( R_i S_i \) with full saturation for \( S_i, X_{i1}, X_{i2} \); and model 2 was an as-treated approach, including \( A_i \) and \( A_i S_i \), also with full saturation for \( S_i, X_{i1}, X_{i2} \). All results were based on 1,000 replicated datasets.

Tables 2 and 4 contain detailed results from simulations across the several described scenarios with Table 2 featuring ignorable noncompliance and Table 3 demonstrating nonignorable non-compliance. Table 4 also features nonignorable noncompliance but differs from 3 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 under weak prediction of treatment for the intermediate variable. 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 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
Table 2: 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. $\rho^X=(\gamma_1,\gamma_2)$ characterizes the association between baseline covariates $X_i$, and $(\gamma_1,\gamma_2)$ characterize the association of $X_i$ and $S_i$. $\gamma_3$ is the coefficient of $A_i$ in the data-generating model for $S_i$. $\rho = 0$ indicates that $S_i$ does not predict $Y_{i,0}$. 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.

$$\psi_0 = (0.5, -0.5), \rho^X=(\text{logit}(0.35),0.8,0.8,1.5), (\gamma_0,\gamma_1,\gamma_2)=(\text{logit}(0.2),1,1)$$

<table>
<thead>
<tr>
<th>$\gamma_3$</th>
<th>$\rho = 0$</th>
<th>$\rho \neq 0$</th>
<th>$\rho \neq 0$</th>
<th>$\rho \neq 0$</th>
<th>$\rho \neq 0$</th>
<th>$\rho \neq 0$</th>
<th>$\rho \neq 0$</th>
<th>$\rho \neq 0$</th>
<th>$\rho \neq 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\psi}_1$</td>
<td>Estimate</td>
<td>% Bias</td>
<td>MCSD</td>
<td>Estimate</td>
<td>% Bias</td>
<td>MCSD</td>
<td>Estimate</td>
<td>% Bias</td>
</tr>
<tr>
<td>$\gamma_3 = 1.2$</td>
<td>GSMM</td>
<td>0.52</td>
<td>5.48</td>
<td>0.50</td>
<td>-0.47</td>
<td>-195.87</td>
<td>0.11</td>
<td>0.48</td>
<td>-195.63</td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>-0.37</td>
<td>-26.51</td>
<td>0.10</td>
<td>-0.37</td>
<td>-26.68</td>
<td>0.14</td>
<td>-0.37</td>
<td>-26.75</td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.50</td>
<td>-0.62</td>
<td>0.11</td>
<td>-0.49</td>
<td>-1.06</td>
<td>0.15</td>
<td>-0.49</td>
<td>-1.06</td>
</tr>
<tr>
<td>$\gamma_3 = 0.3$</td>
<td>GSMM</td>
<td>0.49</td>
<td>-1.26</td>
<td>0.32</td>
<td>-0.45</td>
<td>-10.10</td>
<td>0.65</td>
<td>-0.45</td>
<td>-10.06</td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.23</td>
<td>-53.53</td>
<td>0.09</td>
<td>-0.24</td>
<td>-52.68</td>
<td>0.14</td>
<td>-0.24</td>
<td>-52.68</td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.42</td>
<td>-16.36</td>
<td>0.10</td>
<td>-0.42</td>
<td>-15.04</td>
<td>0.15</td>
<td>-0.42</td>
<td>-15.04</td>
</tr>
<tr>
<td>$\gamma_3 = 0$</td>
<td>GSMM</td>
<td>0.49</td>
<td>-1.46</td>
<td>0.29</td>
<td>-0.43</td>
<td>-13.86</td>
<td>0.71</td>
<td>-0.43</td>
<td>-13.86</td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.34</td>
<td>-31.49</td>
<td>0.09</td>
<td>-0.35</td>
<td>-30.75</td>
<td>0.14</td>
<td>-0.35</td>
<td>-30.75</td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.50</td>
<td>0.65</td>
<td>0.10</td>
<td>-0.51</td>
<td>-1.89</td>
<td>0.15</td>
<td>-0.51</td>
<td>-1.89</td>
</tr>
</tbody>
</table>

22
Table 3: **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. $\rho^X =$ characterizes the association between baseline covariates $X_i$, and $(\gamma_1, \gamma_2)$ characterize the association of $X_i$ and $S_i$. $\gamma_3$ is the coefficient of $A_i$ in the data-generating model for $S_i$. $\rho = 0$ indicates that $S_i$ does not predict $Y_{i,0}$. 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.

$$\psi_0 = (0.5, -0.5), \rho^X = (\text{logit}(0.35), 0.8, 0.8, 1.5), (\gamma_0, \gamma_1, \gamma_2) = (\text{logit}(0.2), 1, 1)$$

<table>
<thead>
<tr>
<th>$\gamma_3$</th>
<th>$\rho$</th>
<th>Method</th>
<th>Estimate</th>
<th>% Bias</th>
<th>MCSD</th>
<th>Estimate</th>
<th>% Bias</th>
<th>MCSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2, $\rho \neq 0$</td>
<td>GSMM</td>
<td>0.52</td>
<td>3.61</td>
<td>0.40</td>
<td>-0.51</td>
<td>2.87</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>-0.19</td>
<td>-137.65</td>
<td>0.09</td>
<td>0.23</td>
<td>-145.79</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.07</td>
<td>-86.34</td>
<td>0.12</td>
<td>0.25</td>
<td>-149.03</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>0.3, $\rho \neq 0$</td>
<td>GSMM</td>
<td>0.49</td>
<td>-2.68</td>
<td>0.31</td>
<td>-0.43</td>
<td>-13.53</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.26</td>
<td>-47.24</td>
<td>0.09</td>
<td>-0.26</td>
<td>-48.42</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.68</td>
<td>36.30</td>
<td>0.10</td>
<td>-0.37</td>
<td>-25.40</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>0, $\rho \neq 0$</td>
<td>GSMM</td>
<td>0.49</td>
<td>-2.98</td>
<td>0.29</td>
<td>-0.41</td>
<td>-18.57</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.35</td>
<td>-29.69</td>
<td>0.08</td>
<td>-0.36</td>
<td>-27.70</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.78</td>
<td>55.88</td>
<td>0.10</td>
<td>-0.47</td>
<td>-5.83</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>1.2, $\rho = 0$</td>
<td>GSMM</td>
<td>0.52</td>
<td>3.18</td>
<td>0.51</td>
<td>-0.48</td>
<td>-3.62</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.23</td>
<td>-53.11</td>
<td>0.09</td>
<td>-0.19</td>
<td>-62.66</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.73</td>
<td>45.76</td>
<td>0.11</td>
<td>-0.38</td>
<td>-23.26</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>0.3, $\rho = 0$</td>
<td>GSMM</td>
<td>0.50</td>
<td>0.97</td>
<td>0.30</td>
<td>-0.45</td>
<td>-9.44</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.33</td>
<td>-33.46</td>
<td>0.09</td>
<td>-0.32</td>
<td>-36.15</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.75</td>
<td>49.14</td>
<td>0.10</td>
<td>-0.41</td>
<td>-18.87</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>0, $\rho = 0$</td>
<td>GSMM</td>
<td>0.51</td>
<td>1.46</td>
<td>0.27</td>
<td>-0.45</td>
<td>-9.83</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.35</td>
<td>-29.04</td>
<td>0.08</td>
<td>-0.36</td>
<td>-28.60</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.75</td>
<td>49.89</td>
<td>0.09</td>
<td>-0.41</td>
<td>-18.05</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>
Table 4: **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 and baseline covariates are weakly predictive of post-baseline effect modifier. $\rho \times (\gamma_1, \gamma_2)$ characterizes the association between baseline covariates $X_i$, and $(\gamma_0, \gamma_1, \gamma_2)$ characterize the association of $X_i$ and $S_i$. $\gamma_3$ is the coefficient of $A_i$ in the data-generating model for $S_i$. $\rho = 0$ indicates that $S_i$ does not predict $Y_{i,0}$. 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.

\[
\psi_0 = (0.5, -0.5), \; \rho \times (\logit(0.35), 0.8, 0.8, 1.5), \; (\gamma_0, \gamma_1, \gamma_2) = (\logit(0.2), 0.2, 0.4)
\]

<table>
<thead>
<tr>
<th>$\gamma_3$</th>
<th>$\rho$</th>
<th>Method</th>
<th>Estimate</th>
<th>% Bias</th>
<th>MCSD</th>
<th>Estimate</th>
<th>% Bias</th>
<th>MCSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2, $\neq 0$</td>
<td>GSMM</td>
<td>0.44</td>
<td>-12.84%</td>
<td>0.67</td>
<td>-0.27</td>
<td>-46.53%</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.11</td>
<td>-78.58%</td>
<td>0.08</td>
<td>-0.05</td>
<td>-90.08%</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.48</td>
<td>-4.01%</td>
<td>0.10</td>
<td>-0.24</td>
<td>-52.72%</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.3, $\neq 0$</td>
<td>GSMM</td>
<td>0.45</td>
<td>-9.74%</td>
<td>0.40</td>
<td>-0.18</td>
<td>-64.84%</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.26</td>
<td>-47.10%</td>
<td>0.07</td>
<td>-0.26</td>
<td>-48.97%</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.66</td>
<td>31.49%</td>
<td>0.09</td>
<td>-0.45</td>
<td>-10.78%</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0, $\neq 0$</td>
<td>GSMM</td>
<td>0.46</td>
<td>-8.25%</td>
<td>0.34</td>
<td>-0.10</td>
<td>-80.45%</td>
<td>1.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.31</td>
<td>-37.11%</td>
<td>0.07</td>
<td>-0.32</td>
<td>-35.35%</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.71</td>
<td>41.40%</td>
<td>0.08</td>
<td>-0.49</td>
<td>-1.24%</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>1.2, $= 0$</td>
<td>GSMM</td>
<td>0.41</td>
<td>-18.78%</td>
<td>0.60</td>
<td>-0.20</td>
<td>-59.99%</td>
<td>1.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.23</td>
<td>-53.79%</td>
<td>0.07</td>
<td>-0.16</td>
<td>-67.59%</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.68</td>
<td>35.20%</td>
<td>0.10</td>
<td>-0.41</td>
<td>-18.36%</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.3, $= 0$</td>
<td>GSMM</td>
<td>0.45</td>
<td>-9.90%</td>
<td>0.36</td>
<td>-0.13</td>
<td>-74.58%</td>
<td>1.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.30</td>
<td>-39.35%</td>
<td>0.07</td>
<td>-0.29</td>
<td>-42.56%</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.69</td>
<td>38.61%</td>
<td>0.08</td>
<td>-0.45</td>
<td>-10.57%</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>0, $= 0$</td>
<td>GSMM</td>
<td>0.46</td>
<td>-7.96%</td>
<td>0.32</td>
<td>-0.07</td>
<td>-86.15%</td>
<td>1.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ITT Log. Reg.</td>
<td>0.32</td>
<td>-36.41%</td>
<td>0.07</td>
<td>-0.32</td>
<td>-35.86%</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AT Log. Reg.</td>
<td>0.70</td>
<td>39.29%</td>
<td>0.08</td>
<td>-0.46</td>
<td>-8.55%</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>
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 4 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 allow the treatment effect estimates to vary as a function of post-treatment levels of depression. Given the necessarily strong assumption for identification, we next explore the plausibility of that assumption. Finally, we review how the statistical estimates from post-treatment effect modification might be used to design future interventions.

6.1 Model Based Estimates

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 allow 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. We also produced an estimate of the effect among compliers based on a plug-in approach proposed by [Palmer et al., 2008] to verify that this estimate is not dependent on the GSMM method. The estimate of the causal odds ratio from this method is 1.93 with a 95% confidence interval of (1.18, 3.15).

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. Richer specifications 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 5 contains estimates for the two causal parameters, $\psi_{01}$ and $\psi_{02}$, under two-stage G-estimation and logistic regression. The GSMM causal estimates (robust standard errors are in parenthesis with $p$-values appearing below) for the $\psi_{02}$ 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 $\psi_{02}$ 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 5 do not readily convey the conditional nature of the causal effect implied by the model. We next explore in more detail how post-treatment levels
Table 5: **Empirical Analysis.** Estimates are presented for the log-odds ratio causal effect parameter $\psi_1$ and post-treatment effect modification $\psi_2$ 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. P-values appear below.

<table>
<thead>
<tr>
<th></th>
<th>Depression at Six Weeks</th>
<th>Depression at Six Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\psi}_1$</td>
<td>$\hat{\psi}_2$</td>
</tr>
<tr>
<td>GSMM</td>
<td>1.45</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>(0.73)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.06</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.008)</td>
</tr>
<tr>
<td></td>
<td>0.775</td>
<td>0.617</td>
</tr>
</tbody>
</table>

of depression modify the effect of the JOBS 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 displays a clear trend. 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). The width of this interval is quite long, which reflects the fact that the support in the data is relatively low at that level of depression. 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 to the point that it appears to be completely without effect for those with higher levels of depression at six weeks.\(^1\) A test for a negative trend based on

\(^1\)The corresponding p-values, which are adjusted for multiple testing, are: 0.035, 0.017, 0.017, 0.585,
the nonparametric Theil-Sen slope estimate was also statistically significant at conventional levels ($p = 0.027$). Thus, while the estimate of the effect modification parameter does not meet the criterion for statistical significance, there does appear to be a significant negative trend in the effect of the treatment as a function of depression at six weeks.

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 minimum and the quartiles of the depression scale distribution.

6.2 Assessing Identification

Identifiability of the estimates requires that we assume that the effect of exposure to the training intervention did not vary across levels of baseline covariates. For example, this implies that the effect of the intervention among those taking treatment did not vary across depression measured at baseline. Next, we explore the plausibility of the necessary no-interaction assumption.

and 0.622. We corrected for multiple testing using the step-up false discovery rate controlling procedure of Benjamini and Hochberg (1995). The trend test doesn’t not account for estimation uncertainty in the point estimates.
As we noted in Section 3, the no-interaction assumption has a testable implication: pretreatment effect modification should be largely absent. The set of baseline covariates is rich enough that there are more possible sub-populations with varying treatment effects than could be fully specified. To that end, we focus specifically on whether treatment effects vary with depression as baseline. We estimated the GSMM to allow for effect heterogeneity by pretreatment levels of depression, while omitting the post-treatment interaction from the model. The exponentiated parameter for effect modification in this model is close to one (0.83) and has 95% confidence intervals that covers one. This is not surprising since the JOBS II study intentionally screened out subjects that were clinically depressed prior to randomization. This screening effectively increases the homogeneity of the study population making the no-interaction more plausible.

We then 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 5. We found that the original pattern of post-treatment effect modification held in the stratified samples. While the no-interaction assumption is untestable, these subsequent analyses lend credibility to our analysis.

6.3 Hypothesis Generation

As we noted above, models with post-treatment effect modification are useful for both designing future interventions and hypothesis generation. Here, we review how the empirical results from the analysis of the JOBS II may be used to design future interventions. The key insight our analysis reveals is that while there is little evidence of pre-treatment effect modification by depression, we do find evidence of post-treatment effect modification as the intervention was notably less successful among job seekers with higher depression levels. Given the fact that clinically depressed subjects were screened out, it would seem that some
post-intervention aspects of the job search process lead to higher levels of depression.

Given these results, analysts might target subjects with elevated levels of depression at an intermediate wave with additional training or counseling. As we noted earlier, job-search setbacks are one likely candidate for the increased levels of depression. A string of failed interviews might easily increase depression. The current study did not attempt to measure job search setbacks at later follow-up periods. In a future intervention, the survey could be reconstructed to measure activities like the number of interviews that did not lead to a job or other setbacks. If post-treatment levels of depression are uncorrelated with job search setbacks that would indicate that further training or counseling may be ineffective.

7 Discussion

We have used GSMMs to estimate causal effects that may be modified by post-treatment variables. Our work complements existing literature on noncompliance and mediation where conditioning occurs on post-treatment variables. One natural comparison is to casual 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). For example, in the original study it was thought that exposure to the training intervention had a direct effect on re-employment but also had an indirect effect through a measure of self-confidence. In contrast, we stipulate only a total effect of the treatment that is conditional on levels of $S_i$. The analysis here also differs from mediation in that the causal contrast involves only the effect of the treatment and does not attempt to parse out the separate pathways by which treatment influences the outcome or the role of the mediator in those pathways, whereas mediation analysis attempts to do those things. The mediation analysis is, in principle, more ambitious, and, thus, potentially less subject to
testing. To see this, consider a (structural nested) model for a mediation analysis in which
the effect of the treatment and the mediator are both modeled conditional on levels of the
treatment $A_i$ and the mediator $S_i$. By modeling the effect of treatment and the mediator
jointly conditional on the same things, we introduce even more parameters into our problem.
Under the assumption (often warranted in studies where the main intervention is randomly
assigned but the mediator is not) that initial randomization holds but sequential ignorability
does not, we have the same number of identifying restrictions. In such circumstances, the
models of joint effects involved in mediation analysis will be even less subject to testing than
our models of effect modification.

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 $H_i(\psi)$ 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 $\psi_0$
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 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, al-
ternatives 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 sub-
ject 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.
A second drawback is the increased variability of estimates determined through G-estimation versus standard logistic regression. This additional variability may be explained by a number of factors, including the semiparametric versus parametric nature of the estimator, as well as the two-stage nature of the estimation strategy. Variance of G-estimation estimates are reduced when baseline covariates contain strong predictors of intermediate variables. For studies using these types of analyses, it is therefore beneficial to record the baseline value of potential intermediate modifying variables and other baseline correlates of intermediates when intending to use this methodology. In this context, efficiency is secondary to bias, and G-estimation yields consistent estimates when intermediates are affected by treatment and predict the outcome, whereas logistic regression does not. Unless it is known that intermediate variables are not affected by treatment, bias considerations support the use of the logistic G-estimation despite its decreased efficiency compared to logistic regression.

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


