Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects

Trang Quynh Nguyen, Cyrus Ebnesajjad, Stephen R. Cole, Elizabeth A. Stuart

October 25, 2016

Abstract

In the presence of treatment effect heterogeneity, the average treatment effect (ATE) in a randomized controlled trial (RCT) may differ from the average effect of the same treatment if applied to a target population of interest. If all treatment effect moderators are observed in the RCT and in a dataset representing the target population, we can obtain an estimate for the target population ATE by adjusting for the difference in the distribution of the moderators between the two samples. This paper considers sensitivity analyses for two situations: (1) where we cannot adjust for a specific moderator V observed in the RCT because we do not observe it in the target population; and (2) where we are concerned that the treatment effect may be moderated by factors not observed even in the RCT, which we represent as a composite moderator U. In both situations, the outcome is not observed in the target population. For situation (1), we offer three sensitivity analysis methods based on (i) an outcome model, (ii) full weighting adjustment, and (iii) partial weighting combined with an outcome model. For situation (2), we offer two sensitivity analyses based on (iv) a bias formula and (v) partial weighting combined with a bias formula. We apply methods (i) and (iii) to an example where the interest is to generalize from a smoking cessation RCT conducted with participants of alcohol/illicit drug use treatment programs to the target population of people who seek treatment for alcohol/illicit drug use in the US who are also cigarette smokers. In this case a treatment effect moderator is observed in the RCT but not in the target population dataset.

Key words: sensitivity analysis, generalization, treatment effect heterogeneity, unobserved moderator, unobserved effect modifier

1   Introduction

Randomized controlled trials (RCTs) can be used to obtain unbiased estimates of the effect of the intervention of interest in the sample used in the trial, resulting in high *internal validity*. However, standard RCTs are not necessarily informative regarding the effects an intervention would have in a target population that may be somewhat different from the RCT sample; in other words, the RCT may have limited *external validity* or *generalizability*. Potential challenges in drawing inferences for populations of policy or decision-making relevance are becoming an increasing concern, as researchers aim to make their research results as relevant as possible.

As shown by Weisberg et al. [2009], Cole and Stuart [2010] and Olsen et al. [2013], results from RCTs may not directly carry over to populations if there are treatment effect moderators whose distribution differs between the RCT sample and the target population. Methods for assessing [Greenhouse et al., 2008, Stuart et al., 2011, 2015] and enhancing [Cole and Stuart, 2010, Tipton, 2013, Kern et al., 2016] generalizability have been proposed. The latter includes approaches that reweight the RCT sample so that it resembles the target population with respect to the observed covariates and plausible moderators [Cole and Stuart, 2010, Kern et al., 2016] or predict treatment effects for target population members based on an outcome model that captures effect heterogeneity [Kern et al., 2016]. However, those methods only adjust for observed characteristics. In practice, once a dataset is identified as representing the target population, it is often found that the number of variables measured consistently between this dataset and the RCT is small [Stuart,

Bradshaw, and Leaf, 2015, Stuart and Rhodes, in press]. In many cases, researchers and policymakers may be worried about unobserved differences between the RCT sample and the target population and how much they influence the conclusions regarding population effects.

This paper presents a set of approaches for assessing the sensitivity of population effect estimates to unobserved moderators, to be used when generalizing from a RCT to a target population. These sensitivity analyses are analogous to methods that assess sensitivity to an unobserved confounder in observational studies [such as Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin, and Wynder, 1959, Rosenbaum and Rubin, 1983a, Rosenbaum, 1987, Gastwirth, Krieger, and Rosenbaum, 1998, Greenland, 1996, Schneeweiss, 2006, Arah, Chiba, and Greenland, 2008, Vanderweele and Arah, 2011, Ding and VanderWeele, 2014, 2016]. They address two situations: (1) when a specific treatment effect moderator is observed in the RCT but is not measured in the target population; and (2) when researchers are concerned about possible effect moderation by factors that are not observed even in the RCT.

The data application in this paper involves generalizing the effect of a smoking cessation intervention from a RCT conducted with participants in alcohol/illicit drug use treatment programs [Reid et al., 2008] to the target population of people who seek treatment for alcohol/illicit drug use in the US who are also cigarette smokers. This RCT is one of the substance use treatment RCTs funded by the US National Institute on Drug Abuse (NIDA); these are deposited in a repository maintained by NIDA's Clinical Trials Network with the purpose to facilitate the use of evidence from RCTs to generate knowledge that informs the provision of treatment services to people with substance use disorders in the US.

With a subset of these RCTs (not including the one in our current application), Susukida and colleagues found significant differences in certain characteristics between the RCT samples and samples they identified as representing relevant target populations [Susukida et al., 2016], and for some interventions, a substantial difference between the average treatment effects (ATEs) for the target population and for the RCT sample due to treatment effect heterogeneity associated with such characteristics (work under review). Such work considers only variables measured in both each RCT and the corresponding target population dataset. With the proposed sensitivity analysis methods, we are able to take one step further, exploring treatment effect moderators among all baseline variables measured in the RCT and conducting sensitivity analysis when finding that one moderator (baseline cigarette addiction score) is not observed in the target population dataset (here drawn from the National Survey on Drug Use and Health, or NSDUH).

The paper is structured as follows: Section 2 describes two methods for obtaining estimates for target population treatment effects when the moderators are observed in both the RCT and the target population dataset; these are the basis of the sensitivity analyses we propose. Section 3 presents sensitivity analysis methods for settings where a moderator is observed in the RCT but not in the target population. Section 4 addresses sensitivity analyses for effect moderation that is not even observed in the RCT. Section 5 reports on the data application. Section 6 concludes with a discussion.

2 Two methods for generalization when the moderators are observed in both the RCT and in a dataset representing the target population

This section formalizes the goal of inference, desbribes notation, and reviews two methods for generalizing treatment effect estimates from an RCT to a target population; these methods form the basis for the sensitivity analyses described below.

Consider a RCT in which participants are randomly assigned to active treatment ($T = 1$) and control ($T = 0$) conditions, and their outcomes (Y) are observed. In this sample, we also observe pre-treatment covariates, including covariates Z that interact with treatment in influencing the outcome, and covariates X that influence the outcome but do not interact with treatment. Z and X are generally multivariate, but we use univariate notation to simplify presentation.

Suppose we have data from a second sample, one that represents the target population. Let S represent sample membership, with $S = 1$ if a person is in the RCT and $S = 0$ if a person is in the target population

sample. Here we assume that the two samples are disjoint. (For the case where the RCT sample is a subset of the target population sample, the methods are slightly modified [Cole and Stuart, 2010], which we comment on in the Discussion section.) In this section, we consider the situation where we also observe the treatment effect moderators Z in the target population dataset. All through this paper we assume that the outcome is not observed in the target population.

Let Y^t denote the potential outcome if under treatment condition $t, t \in \{0, 1\}$. For each RCT participant, we observe one of the two potential outcomes Y^1, Y^0. For those in the target population sample, we observe neither. We are interested in the average treatment effects (ATEs) both for the RCT sample and for the target population, which we refer to respectively as the *Sample Average Treatment Effect* (SATE) and the *Target Average Treatment Effect* (TATE). These are defined as the average of the individual additive treatment effects over the RCT sample and over the target population:

$$\text{SATE} \equiv \mathbb{E}[Y^1 - Y^0 | S = 1] = \mathbb{E}[Y^1 | S = 1] - \mathbb{E}[Y^0 | S = 1], \tag{1}$$

$$\text{TATE} \equiv \mathbb{E}[Y^1 - Y^0 | S = 0] = \mathbb{E}[Y^1 | S = 0] - \mathbb{E}[Y^0 | S = 0]. \tag{2}$$

Estimation of SATE is straightforward. For simplicity, consider simple randomization, with all RCT participants having the same probability of being assigned treatment.[1] An unbiased estimate of SATE can be obtained by taking the difference in mean outcome between the treated and control groups, or regressing outcome on treatment adjusting for pre-treatment covariates. Estimation of TATE, on the other hand, requires adjustment for treatment effect moderators whose distribution differs between the RCT sample and the target population [Olsen et al., 2013, Cole and Stuart, 2010]. The methods for estimating TATE described below assume *conditional sample ignorability for treatment effects* [Kern et al., 2016]: being in the RCT or in the target population sample does not carry any information about treatment effect once we condition on the moderators Z.

2.1 Outcome-model-based TATE estimation

We assume an additive model for the potential outcomes. With i indexing the individual, the model is

$$\mathbb{E}[Y_i^t] = \beta_0 + f_{zt}(Z_i, t) + f_{xz}(X_i, Z_i), \quad t = 0, 1, \tag{3}$$

where f_{zt}, f_{xz} are functions of the corresponding variables. For simplicity, we consider the special form $f_{zt}(Z_i, t) = \beta_t t + \beta_{zt} Z_i t$, which is perhaps the one most commonly used in practice. (This form assumes constant moderation effect, as β_{zt} does not depend on the level of Z.) The simplified model is

$$\mathbb{E}[Y_i^t] = \beta_0 + \beta_t t + \beta_{zt} Z_i t + f_{xz}(X_i, Z_i), \quad t = 0, 1. \tag{4}$$

The form of $f_{xz}(X_i, Z_i)$ is not of interest here, but a common practice is to use $\beta_x X_i + \beta_z Z_i$ and perhaps add some complexity such as quadratic or interaction terms.

The treatment effect for individual i is

$$\mathbb{E}[Y_i^1] - \mathbb{E}[Y_i^0] = \beta_t + \beta_{zt} Z_i \tag{5}$$

which means

$$\text{SATE} = \beta_t + \beta_{zt} \mathbb{E}[Z | S = 1], \tag{6}$$

$$\text{TATE} = \beta_t + \beta_{zt} \mathbb{E}[Z | S = 0]. \tag{7}$$

The difference between SATE and TATE, $\beta_{zt} \{\mathbb{E}[Z | S = 1] - \mathbb{E}[Z | S = 0]\}$, is the bias if we generalize the effect estimated in the RCT directly to the target population without adjusting for differences in Z. The magnitude of this bias depends on the moderation effect (β_{zt}) and the difference between the means of the

---

[1]If the RCT design is complex and treatment probabilities vary across individuals, a minor variation that incorporates inverse-probability-of-treatment weights can be used.

moderator in the two samples ($\{\mathbb{E}[Z|S=1] - \mathbb{E}[Z|S=0]\}$). If either of these is zero, SATE is equivalent to TATE.

When Z is observed in both samples, an estimate for TATE can be obtained using eq. 7, with $\mathbb{E}[Z|S=0]$ estimated using the target population dataset, and with β_t and β_{zt} estimated by fitting to the RCT data an outcome model with interaction terms.[2] While eq. 7 does not involve X, the accuracy and precision of the estimates of β_t and β_{zt} require a good estimate of the outcome model. Not only do we need to capture all Z variables, all X variables (or at least all X variables that are correlated with, or interact with, Z variables) should be included and correctly modeled.

Note that we have invoked the *conditional sample ignorability for treatment effects* assumption when equating $\{\beta_t, \beta_{zt}\}$ between equations 6 and 7. This assumption is violated if we do not observe all the moderators that are differentially distributed between the two samples. It is also violated if the range of Z in the target population includes segments not covered by the RCT, a violation of the *positivity* assumption [Rosenbaum and Rubin, 1983b]; using eq. 7 in this case would result in extrapolation beyond the support of the data. Positivity is often not a problem with a binary Z, but for a continuous Z, care needs to be taken to check overlap, and judgment needs to be made about whether extrapolation to any uncovered areas is reasonable.

2.2   Weighting-based TATE estimation

The idea of this method is to reweight the RCT sample so that it resembles the target population with respect to the distribution of the treatment effect moderators (Z) and then to use this weighted RCT sample to estimate TATE.

The weighting procedure involves first stacking the RCT and target population datasets and fitting a model predicting sample membership. The set of predictors in this model needs to include all the moderators (Z variables); outcome predictors that do not moderate treatment effect (X variables) do not need to be included. To determine which pre-treatment covariates are moderators requires a prior step of detecting them through modeling the outcome. There may be times when it is hard to know whether a variable is a moderator (e.g., its interaction term with treatment has a substantial but statistically non-significant coefficient), in which case it is preferable to treat it as a moderator and include it in the sample membership model. For the same reason (or to avoid having to model the outcome), one may also include a broader set of variables in this model, regardless of whether they may be moderators (Z) or not (X).

The fitted sample membership model is used to compute the predicted odds of being in the target population sample for the RCT participants. These odds are then used to reweight the RCT sample. As a result, the weighted RCT sample better resembles the target population with respect to the distribution of the variables used in the sample membership model. This strategy of weighting the RCT sample to the target population sample has been described by Kern et al. [2016] and Cole and Stuart [2010];[3] here we emphasize the distinction between moderators and other covariates, as the purpose of the weighting is to adjust for the diffential distribution of the moderators. Whether the weighting succeeds in doing this should be checked.

The weighted RCT sample is used to estimate an average treatment effect, which is taken as the estimated TATE. A simple estimator for TATE is the difference between the weighted means of the outcome in the RCT's treated and control groups. Another option is to fit a weighted regression model that controls for Z and X variables (but not their interaction terms with T), and estimate TATE with the coefficient of T.

---

[2]An alternative is to estimate the outcome model, predict treatment effects for the individuals in the target population dataset [Kern et al., 2016] using eq. 5, and average them. This strategy does not require the constant moderation effect assumption. However, it does not provide for a straightforward sensitivity analysis for an unobserved moderator.

[3]The weights are the same in Kern et al. [2016], but slightly different in Cole and Stuart [2010] because in the latter case the RCT sample was a subsample of the target population dataset. We will return to this in the Discussion section.

3 Sensitivity analysis for a moderator that is observed in the RCT but not in the target population sample

We continue using Z to denote moderators observed in both samples, and use V to denote a moderator observed in the RCT but not in the target population sample. (We hereafter refer to the current case as the V case, to differentiate it with the case to be addressed in Section 4.) In this case, although TATE cannot be estimated in a way that adjusts for all of the moderators, we can conduct sensitivity analysis to assess how TATE estimates would change based on what we assume about the distribution of V in the target population. Here we present several sensitivity analysis methods, and report on two simulation studies that compare some of these methods to one another.

3.1 Three sensivity analysis strategies

The methods described below are based on an outcome model, full weighting adjustment, and partial weighting combined with an outcome model.

3.1.1 Outcome-model-based sensitivity analysis

We rewrite the potential outcomes model, separating Z and V:

$$\mathbb{E}[Y_i^t] = \beta_0 + \beta_t t + \beta_{zt} Z_i t + \beta_{vt} V_i t + f_{xzv}(X_i, Z_i, V_i). \tag{8}$$

For simplicity, this model makes an additional assumption (compared to the model in eq. 4) that there is no three-way interaction of the treatment with both Z and V. Based on this model, the formula for TATE is

$$\text{TATE} = \beta_t + \beta_{zt}\mathbb{E}[Z|S=0] + \beta_{vt}\mathbb{E}[V|S=0], \tag{9}$$

where $\beta_t, \beta_{zt}, \beta_{vt}, \mathbb{E}[Z|S=0]$ can be estimated from data, whereas $\mathbb{E}[V|S=0]$ cannot. We will refer to the latter as an 'unknown' parameter, which is a slight abuse of terminology because the true values of all these parameters, $\beta_t, \beta_{zt}, \beta_{vt}, \mathbb{E}[Z|S=0]$ and $\mathbb{E}[V|S=0]$, are not known. By 'unknown' here, we mean that one cannot learn about this parameter from data, while one can learn about the other parameters from data.

The simple formula in eq. 9 results from the no three-way interaction assumption. Without such assumption, the potential outcomes model would have an additional term, $\beta_{zvt}Z_iV_it$, and the formula for TATE would include $\beta_{zvt}\mathbb{E}[ZV|S=0]$, with the unknown $\mathbb{E}[ZV|S=0]$ being more complex to consider than simply $\mathbb{E}[V|S=0]$.

To conduct the sensitivity analysis, first we need to estimate the estimable quantities. $\mathbb{E}[Z|S=0]$ is estimated using target population data. Assuming sample ignorability for treatment effects conditional on Z, V, we estimate $\beta_t, \beta_{zt}, \beta_{vt}$ using the RCT data; this involves estimating the outcome model with interaction terms (eq. 8) in the same manner as discussed in section 2.1, and extracting the estimated values and variance-covariance matrix of $\beta_t, \beta_{zt}, \beta_{vt}$.

We then specify a plausible range for the unknown $\mathbb{E}[V|S=0]$ (mean V in the target population). In doing this, it is important to check if the range of V being considered has good overlap with the RCT sample.

A range for the TATE point estimate is computed by plugging the point estimates of $\beta_t, \beta_{zt}, \beta_{vt}, \mathbb{E}[Z|S=0]$ and the specified range of $\mathbb{E}[V|S=0]$ into eq. 9.

A confidence band to accompany this TATE range can be obtained. For each value of $\mathbb{E}[V|S=0]$ in the specified range, the variance-covariance matrix of the estimated $\beta_t, \beta_{zt}, \beta_{vt}$ can be used to obtain a confidence interval for TATE. If the uncertainty in the estimated $\mathbb{E}[Z|S=0]$ is non-negligible, it can be incorporated by using the confidence limits of $\mathbb{E}[Z|S=0]$ (rather than its point estimate) in the construction of such confidence intervals.

3.1.2 Weighting-based sensitivity analysis

Ideally, had V been available from both samples, we would be able to estimate TATE using RCT data, weighting the individuals by their odds of being in the target population sample conditional on Z, V (as

described in section [2.2]). While such weights cannot be estimated when V is not observed in the target population, they can be reexpressed, using Bayes' rule, as

$$
\begin{aligned}
W_i &= \frac{\mathrm{P}(S=0|Z_i,V_i)}{\mathrm{P}(S=1|Z_i,V_i)} \\
&= \frac{\mathrm{P}(S=0,Z=Z_i,V=V_i)/\mathrm{P}(Z=Z_i,V=V_i)}{\mathrm{P}(S=1,Z=Z_i,V=V_i)/\mathrm{P}(Z=Z_i,V=V_i)} \\
&= \frac{\mathrm{P}(S=0,Z=Z_i,V=V_i)}{\mathrm{P}(S=1,Z=Z_i,V=V_i)} \\
&= \frac{\mathrm{P}(Z=Z_i)\mathrm{P}(S=0|Z_i)\mathrm{P}(V=V_i|S=0,Z_i)}{\mathrm{P}(Z=Z_i)\mathrm{P}(S=1|Z_i)\mathrm{P}(V=V_i|S=1,Z_i)} \\
&= \frac{\mathrm{P}(S=0|Z_i)}{\mathrm{P}(S=1|Z_i)} \cdot \frac{\mathrm{P}(V=V_i|S=0,Z_i)}{\mathrm{P}(V=V_i|S=1,Z_i)}.
\end{aligned}
\tag{10}
$$

Each weight is thus a product of two components: (1) the odds of being in the target population sample conditional on Z but not V, and (2) a ratio of the probability density/mass of $V=V_i$ in the $Z=Z_i$ stratum comparing the target population sample to the RCT sample.[4]

In this formula of the weights (eq. [10]), the first component can readily be estimated from data; the denominator of the second component can also be estimated. The numerator of the second component, $\mathrm{P}(V=V_i|S=0,Z_i)$, is unknown. This suggests that a sensitivity analysis can be conducted by specifying a plausible range for the unknown distribution of V given Z in the target population, $\mathrm{P}(V|S=0,Z)$, and for each distribution in this range, constructing weights and estimating TATE using the reweighted RCT sample. TATE can be estimated using either the difference in weighted mean outcome between the treated and control conditions, or using regression of the outcome on treatment and covariates.

The challenge is how to estimate $\mathrm{P}(V|S=1,Z)$ and how to specify plausible ranges for $\mathrm{P}(V|S=0,Z)$. Both these tasks are complicated and results are prone to misspecification bias when V or Z or both are of any form but binary. We hereby limit the consideration of this method to the case where V and Z are binary. With one binary Z and one binary V, there are only four unique weights:

$$
\begin{aligned}
W_{i|V_i=1,Z_i=1} &= \frac{\mathrm{P}(S=0|Z=1)}{\mathrm{P}(S=1|Z=1)} \cdot \frac{\mathrm{P}(V=1|S=0,Z=1)}{\mathrm{P}(V=1|S=1,Z=1)}, \\
W_{i|V_i=0,Z_i=1} &= \frac{\mathrm{P}(S=0|Z=1)}{\mathrm{P}(S=1|Z=1)} \cdot \frac{1-\mathrm{P}(V=1|S=0,Z=1)}{1-\mathrm{P}(V=1|S=1,Z=1)}, \\
W_{i|V_i=1,Z_i=0} &= \frac{\mathrm{P}(S=0|Z=0)}{\mathrm{P}(S=1|Z=0)} \cdot \frac{\mathrm{P}(V=1|S=0,Z=0)}{\mathrm{P}(V=1|S=1,Z=0)}, \\
W_{i|V_i=0,Z_i=0} &= \frac{\mathrm{P}(S=0|Z=0)}{\mathrm{P}(S=1|Z=0)} \cdot \frac{1-\mathrm{P}(V=1|S=0,Z=0)}{1-\mathrm{P}(V=1|S=1,Z=0)}.
\end{aligned}
$$

The denominators in the second component of these weights are easily estimated. For the numerators, we need to specify ranges for two probabilities: $\mathrm{P}(V=1|S=0,Z=1)$ and $\mathrm{P}(V=1|S=0,Z=0)$, the prevalence of $V=1$ in the target population given $Z=1$ and $Z=0$.

3.1.3 Weighted-outcome-model-based sensitivity analysis

While the full weighting strategy is hard to implement in the context of sensitivity analysis, a partial weighting version combined with an outcome model lends itself well to sensitivity analysis. The idea is to use weighting to adjust for known differences between the two samples (here differences in the distribution of Z) and then to use an outcome model to do sensitivity analysis on unknown quantities (V in the target population). First, we weight the RCT sample by the individuals' odds of being in the target population

---

[4]This ratio is of the same form as a ratio used elsewhere in weighting to control confounding in causal mediation analysis [Hong, 2010].

sample conditional on Z only, $\frac{P(S=0|Z_i)}{P(S=1|Z_i)}$. We then use this weighted RCT sample to estimate the outcome model (of the form in eq. 8). We use the estimates of $\beta_t, \beta_{zt}, \beta_{vt}$ and their variance-covariance matrix from this model as inputs for estimating TATE in the same manner as in the non-weighted *outcome-model-based* method (based on TATE formula eq. 9).

The difference between this method and the first method is that the weighting makes the distribution of Z in the RCT sample more similar to that in the target population, and thereby helps adjust for the discrepancy in average treatment effect due to effect moderation by Z. This may be helpful in the case the Z part of the outcome model is misspecified, which we investigate in a simulation study reported in section 3.2.

3.2 Simulation study comparing the outcome-model-based and weighted-outcome-model-based sensitivity analyses

We investigate how well these two methods perform relative to each other in recovering the true TATE, in situations where the outcome model is correctly or incorrectly specified. When the outcome model is correctly specified, we expect that both methods are unbiased. When the Z part of the outcome model is misspecified, we expect that the *weighted-outcome-model-based* method is less biased. When the V part of the outcome model is misspecified, we expect that the same method helps reduce bias due to this misspecification if Z and V are positively correlated and influence treatment effect in the same direction.

3.2.1 Data generation

We consider situations with one X, one Z and one V. X is a standard normal random variable. Z and V are first generated as multivariate normal with correlations ranging from 0 to $\pm.5$, and then each is either kept in continuous form or dichotomized. When either Z or V is binary, its prevalence is .25 in the RCT sample and .5 in the target population. When either Z or V is continuous, it has mean 0 in the RCT and .5 in the target population, and variance 1 in both.

In the RCT, T is randomly assigned to 0 and 1 with probability 0.5. With regards to the outcome, for the continuous Z and V combination, we use a base model with Z and V as moderators, plus six other models, each with one additional moderator from among Z^2, V^2 or ZV, whose moderation effect is either positive or negative.

$$
\begin{array}{ll}
\text{A.} & Y = X + T + Z + V + ZT + VT + \epsilon_Y \\
\text{B1.} & Y = X + T + Z + V + ZT + VT + Z^2T + \epsilon_Y \\
\text{B2.} & Y = X + T + Z + V + ZT + VT - Z^2T + \epsilon_Y \\
\text{C1.} & Y = X + T + Z + V + ZT + VT + V^2T + \epsilon_Y \quad , \quad \epsilon_Y \sim \mathrm{N}(0,4) \\
\text{C2.} & Y = X + T + Z + V + ZT + VT - V^2T + \epsilon_Y \\
\text{D1.} & Y = X + T + Z + V + ZT + VT + ZVT + \epsilon_Y \\
\text{D2.} & Y = X + T + Z + V + ZT + VT - ZVT + \epsilon_Y
\end{array}
$$

For the continuous Z and binary V combination, we use models A, B1-2 and D1-2. For the binary Z and continuous V combination, we use A, C1-2 and D1-2. For the binary Z and V combination, we use A and D1-2.

For each scenario (combining Z and V types and outcome model), 100,000 $n=400$ RCT samples and $n=5000$ target population samples are generated.

3.2.2 Methods implementation

Outcome models used in the sensitivity analyses For scenarios with the true outcome model A, we implement the *outcome-model-based* and *weighted-outcome-model-based* sensitivity analyses using the correctly specified outcome model. For the other scenarios, we implement these methods using the correct model as well as the misspecified model leaving out the third moderator (Z^2, V^2 or ZV). We choose to

consider this misspecified model because it is simple and perhaps most often used. In practice, detection of moderation effects using regression is often an exploratory analysis trying out interaction terms of different covariates with treatment. More complex interaction terms are less often considered, and even if they are, the power to detect them is limited.

Weighting details With the *weighted-outcome-model-based* method, the weighting is partial, adjusting for Z but not V. We compute the weights are based on a logistic sample membership model with Z as the predictor. With a continuous Z, to allow flexible modeling of sample membership probability, we use natural splines with nine knots.

3.2.3   Simulation results

The main results to report concern the bias or lack thereof of the two methods under investigation when using correct or misspecified outcome models. Since the true ATE in a particular target population sample may differ from the true TATE as set by the simulation parameters, we use the true ATE for a particular target population sample as the goal of inference for each simulation iteration. This helps avoid noise due to sampling variability and focuses on the bias itself.

Figure 1 presents the bias of these two methods using correct and misspecified outcome models. Across all scenarios, when the correct outcome model is used, both methods are unbiased. When a misspecified model is used, both methods are biased. When the model is correctly specified with respect to V but misspecified with respect to Z (in scenarios with Z^2 as the third moderator – column 2 in the Figure), as expected, the weighted-outcome-model-based method, which uses weighting to adjust for Z, is less biased than the outcome-model-based method.

When the model is correctly specified with respect to Z but misspecified with respect to V (scenarios with V^2 as the third moderator – column 3), the two methods are similarly biased if Z and V are uncorrelated. When Z and V are positively correlated, the weighted-outcome-model-based method becomes less biased, because the weighting adjustment for Z provides some adjustment for V. When Z and V are negatively correlated, the contrary is true, with the weighted-outcome-model-based method being more biased.

When the model is misspecified with respect to both Z and V (scenarios with ZV as the third moderator – column 4), both methods are biased. For most of the range of Z-V correlation considered, the weighted-outcome-model-based method is less biased. At a certain point when the correlation is high enough the side-effect adjustment for V that results from weighting adjustment for Z pulls the TATE estimate to zero bias and then past zero to bias of opposite sign.

These results confirm our hypotheses that (i) when the outcome model is misspecified with respect to Z, the weighted-outcome-model-based method is less biased than the outcome-model-based method; and that (ii) when the outcome model is misspecified with respect to V, and Z and V are positively correlated and influence treatment effect in the same direction, the weighted-outcome-model-based method tends to be less biased.

4   Sensitivity analysis for effect moderation that is completely unobserved

In this situation, there may be known moderators (Z) that we observe in both the RCT and the target population sample. We are concerned, however, that there may be additional effect moderation by factors that are unobserved even in the RCT that may be differentially distributed between the RCT sample and the target population. While in the previous situation we deal with a specific variable V (specific because it is observed in the RCT), here we think of a generic variable denoted by U that is unknown and unobserved. At this point in our conceptualization, U can potentially be an actual characteristic that is unobserved, or it can be a quantity that combines multiple effect moderation factors. Our purpose is to determine whether sensitivity analyses for the V case can be adapted for the current case (which we refer to as the U case).

Figure 1: Bias of *outcome-model-based* and *weighted-outcome-model-based* sensitivity analyses using correct and misspecified outcome models.
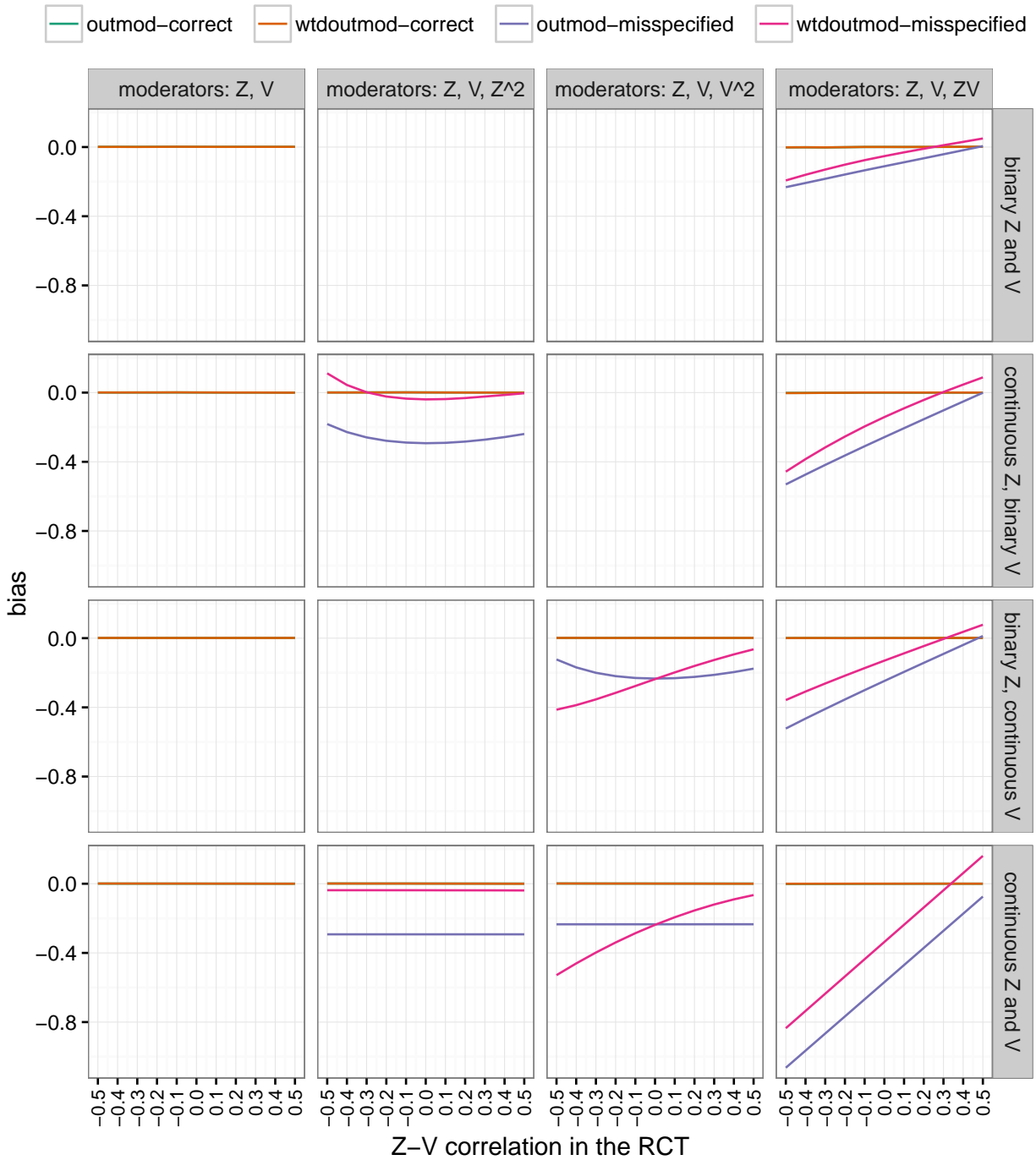
Notes: 'outmod' = outcome-model-based; 'wtdoutmod' = weighted-outcome-model-based. For both methods, the same outcome models are used. In scenarios with only two moderators (Z and V), only the correct model is used. In scenarios with a third moderator (Z^2, V^2 or ZV), the correct model and the misspecified model that excludes the third moderator are used. In all plots, the green curve (the outcome-model-based method using the correct model) lies underneath the brown curve (the weighted-outcome-model-based method also with the correct model) and is thus not visible. In these scenarios the moderation effects of Z^2, V^2 or ZV are positive. Results from scenarios where they are negative are mirror images of these plots across the horizontal zero line, i.e., the sign of bias is flipped.

The full-weighting method requires U to be observed in the RCT, so cannot be used here. The outcome-model-based methods, with a formula for TATE that includes the β_t term (like the formula in eq. 9, except replacing V with U), also cannot be used because the estimation of β_t depends on the variables interacting with treatment, including U. However, a small modification results in methods that work for a special definition of U.

4.1 Two sensitivity analyses for the U case

4.1.1 Bias-formula-based sensitivity analysis

Assume the linear potential outcomes model

$$\mathbb{E}[Y_i^t] = \beta_0 + \beta_t t + \beta_{zt} Z_i t + \beta_{ut} U_i t + f_{xzu}(X_i, Z_i, U_i) \tag{11}$$

similar to the model for the V case, also with no three-way interaction with treatment. With the assumption of sample ignorability for treatment effects conditional on Z, U, we have both

$$\text{SATE} = \beta_t + \beta_{zt}\mathbb{E}[Z|S=0] + \beta_{ut}\mathbb{E}[U|S=1], \text{ and} \tag{12}$$

$$\text{TATE} = \beta_t + \beta_{zt}\mathbb{E}[Z|S=0] + \beta_{ut}\mathbb{E}[U|S=0]. \tag{13}$$

Note that this assumption requires U to capture all effect moderating forces other than Z, thus narrowing the definition of U. Eq. 12 and eq. 13 imply that

$$\text{TATE} = \text{SATE} + \beta_{zt}\{\mathbb{E}[Z|S=0] - \mathbb{E}[Z|S=1]\} +$$
$$+ \beta_{ut}\{\mathbb{E}[U|S=0] - \mathbb{E}[U|S=1]\}. \tag{14}$$

On the right hand-side of eq. 14, SATE can be estimated unbiasedly as the difference in mean outcome between the two treatment conditions in the RCT. To use eq. 14 for sensitivity analysis, we need an unbiased estimate of β_{zt}. Like in the V case, the model used to estimate β_{zt} needs to include all X variables that are correlated, or interact, with Z. Leaving U out of the model, however, generally leads to bias in the estimated β_{zt} (as well as other coefficients). The only situation where omitting U would not bias β_{zt} is when U is independent of Z. This requires further refining the definition of U to a quantity that combines all the unobserved moderating factors after 'regressing out' Z. We call this variable the *remaining composite moderator after accounting for Z*, and denote it by $U_{(z)}$.[5]

With $U_{(z)}$ so defined, we can estimate β_{zt} and use

$$\text{TATE} = \text{SATE} + \beta_{zt}\{\mathbb{E}[Z|S=0] - \mathbb{E}[Z|S=1]\} +$$
$$+ \beta_{ut}\{\mathbb{E}[U_{(z)}|S=0] - \mathbb{E}[U_{(z)}|S=1]\} \tag{15}$$

for sensitivity analysis. By varying $\beta_{ut}\{\mathbb{E}[U_{(z)}|S=0] - \mathbb{E}[U_{(z)}|S=1]\}$, we get a range for the point estimate of TATE. We will address how to specify ranges for such an unknown quantity after discussing the *weighting-plus-bias-formula-based* method.

4.1.2 Weighting-plus-bias-formula-based sensitivity analysis

With this approach, we have the option of weighting the RCT sample to adjust for Z and conducting a bias-formula-based sensitivity analysis for a U that is independent of Z (the remaining composite moderator after accounting for Z). Yet it is plausible that X variables may carry some (even if limited) information about unobserved moderators—they may be correlated with unobserved moderators but the correlations are small so X do not appear to be moderators themselves. We therefore propose adjusting for both X and Z through weighting and then conducting sensitivity analysis for a U independent of X, Z (the *remaining composite moderator after accounting for X, Z*). We denote this variable by $U_{(xz)}$.

---

[5]This consideration of $U_{(z)}$ independent of all observed moderators Z parallels the convention of evaluating treatment effect sensitiveness to an unobserved confounder independent of observed confounders [Rosenbaum and Rubin, 1983a].

We weight the individuals in the RCT by their odds of being in the target population sample conditional on X, Z. The weighted RCT sample now better resembles the target population sample with respect to the distribution of X, Z. On the other hand, it resembles the unweighted RCT sample with respect to the distribution of U, because U is independent of X, Z. We call the ATE estimated from this weighted sample the *X-and-Z-adjusted* ATE (xzATE). Based on the potential outcomes model,

$$\text{xzATE} = \beta_t + \beta_{zt}\mathbb{E}[Z|S=1, \text{xz-wtd}] + \beta_{ut}\mathbb{E}[U_{(xz)}|S=1] \tag{16}$$
$$\approx \beta_t + \beta_{zt}\mathbb{E}[Z|S=0] + \beta_{ut}\mathbb{E}[U_{(xz)}|S=1] \tag{17}$$

(where 'xz-wtd' stands for 'weighted to adjust for X, Z'). This means

$$\text{TATE} = \text{xzATE} + \beta_{zt}\{\mathbb{E}[Z|S=0] - \mathbb{E}[Z|S=1, \text{xz-wtd}]\}+$$
$$+ \beta_{ut}\{\mathbb{E}[U_{(xz)}|S=0] - \mathbb{E}[U_{(xz)}|S=1]\}. \tag{18}$$

and

$$\text{TATE} \approx \text{xzATE} + \beta_{ut}\{\mathbb{E}[U_{(xz)}|S=0] - \mathbb{E}[U_{(xz)}|S=1]\}. \tag{19}$$

An unbiased estimate for xzATE is the difference between the weighted means of the outcome in the two treatment conditions in the RCT. If the weighting succeeds in equating the means of Z between the RCT and target population datasets, eq. 19 can be used for sensitivity analysis. If the weighting reduces the distance between these means but not to zero, eq. 18 can be used. If eq. 18 is used, we get a range for TATE point estimates corresponding to the plausible range specified for the unknown $\beta_{ut}\{\mathbb{E}[U_{(xz)}|S=0] - \mathbb{E}[U_{(xz)}|S=1]\}$. If eq. 19 is used, in addition to the point estimate range, we also get confidence limits for TATE.

Plausible range specification for sensitivity parameters Both of the bias-formula-based methods require specifying some plausible range for the unknown $\beta_{ut}\{\mathbb{E}[U|S=0] - \mathbb{E}[U|S=1]\}$ where U is either $U_{(z)}$ or $U_{(xz)}$. This quantity can be considered the combination of two sensitivity parameters: one representing the moderation effect (β_{ut}) and the other representing the association between U and sample membership (the difference in mean U between the two samples, $\mathbb{E}[U|S=0] - \mathbb{E}[U|S=1]$). As the remaining composite moderator (combining potentially multiple moderating factors), the most appropriate form for U to take is perhaps the form of a continuous variable. We propose using a standardized metric here, so the difference in mean U between the two samples is in standard deviation units, and β_{ut} is the change in treatment effect associated with one standard deviation difference in U.

Alternative conceptualization of U as a natural variable The definition of U as a composite variable—representing the remaining effect moderation factors after accounting for Z or for X, Z—requires some degree of abstraction away from real world quantities. It may be common, however, for scientists to think in more concrete terms, asking whether there may exist an unobserved natural variable (as opposed to a composite variable) that moderates treatment effect and that is differentially distributed between the target population and the RCT sample. It is important to note that this is a special-case interpretation of U, and that it requires that (1) this unobserved natural variable is the only unobserved moderator, and that it is either (2a) independent of X, Z (if using the weighting-plus-bias-formula-based method and weighting to adjust for X, Z), or (2b) independent of Z (if using the bias-formula-based method or if using the weighting-plus-bias-formula-based method but weighting to adjust for Z only). In this special case where U is a natural variable, it can be of any form, e.g., continuous, dichotomous, polytomous, etc.

5   Real data application

We consider a smoking cessation RCT for drug and/or alcohol-dependent adults [Reid et al., 2008], known as CTN9 in NIDA's Clinical Trial Network's repository of substance use treatment RCTs. Participants (n=225) were adult cigarette smokers who at baseline smoked at least 10 cigarettes per day, recruited from among people who attended outpatient community-based treatment programs for opiate, cocaine and alcohol

dependence. They were randomly assigned in a 2:1 ratio to receive either smoking cessation treatment or no such treatment. Smoking cessation treatment consisted of one week of group counseling before the target quit date and eight weeks of group counseling plus transdermal nicotine patch treatment (21 mg per day for weeks 1 to 6 and 14 mg per day for weeks 7 and 8) after the target quit date. We retain 200 participants in analysis (including 65 treated and 135 controls), excluding 18 with no outcome data, and then an additional seven who were either in a controlled environment, or Asian, Pacific Islanders, and Native Americans, since generalizing from such small numbers would be inadvisable, given the plausibility of these categories as effect moderators.

Given NIDA's interest in using NIDA-supported RCTs to generate evidence relevant to practice, we define the target population to be adults in the US who seek treatment for alcohol/substance use disorders who also smoke at least 10 cigarettes per day. To represent this target population, we use a subset of the 2014 National Survey on Drug Use and Health (NSDUH), a representative sample of the US population excluding homeless persons outside shelters, active duty personnel, and those in controlled environments. Out of 55,271 NSDUH respondents, 2,751 were adults (aged 18 and older) who reported having ever sought treatment for substance abuse, excluding Asians, Pacific Islanders, and Native Americans. Of those, 934 reported smoking at least 10 cigarettes a day on average, comprising our target population sample.

Table 1 summarizes baseline characteristics of the RCT sample (including demographics, education, employment, baseline smoking, cigarette addiction severity score, number of past quit attempts, years smoking, reasons for quitting, and primary substance of abuse – see Reid et al. [2008] for detailed description) and the same variables (if available) from the target population sample. The two samples differ in all the characteristics observed in both: the RCT sample has larger proportions of Hispanic, African-American and female participants, is older, and smokes more on average.

Reid and colleagues analyzed the RCT data using a longitudinal model with the daily numbers of cigarettes smoked (collected once a week during active treatment and at three and sixteen weeks after treatment) as repeated outcome measures. They found a significant reduction in the number of cigarettes smoked per day in the treatment group. In our analysis, we use the mean number of cigarettes smoked per day over the eight weeks after the target quit date as the outcome variable. This is justifiable since after the target quit date, the number of cigarettes smoked by the treatment group declined and stayed at about the same level throughout the end of treatment.

To get an estimate of SATE, we fit a linear model for the outcome that includes treatment condition and all the baseline covariates in Table 1. Consistent with findings by Reid and colleagues, we find a significant average decrease of ten cigarettes a day as a result of the treatment. We then explore treatment effect heterogeneity, considering models with the same variables (as main effects) plus covariate-treatment interactions. For model selection, we use stepwise regression with forward selection and backward elimination, minimizing the Akaike information criterion. The model selected, presented in Table 2, includes interaction terms of treatment with African-American race cateogry, baseline number of cigarettes per day, and with baseline cigarette addiction severity. Specifically, African-American participants and participants who smoked a larger number of cigarettes per day at baseline experienced a larger reduction, and those with higher baseline addiction score experienced a smaller reduction, in cigarettes smoked per day.

A key problem with trying to genearlize the RCT results to the target population is that the baseline cigarette addiction severity score is a treatment effect moderator, but its target population distribution is unknown, because this variable is not available from the NSDUH. The methods described in this paper can be used to generate TATE estimates accounting for possible distributions of cigarette addiction severity in the target population. In the RCT, the mean of this variable is 4.05 (on a 1 to 5 scale). We assume that the mean cigarette addiction severity score in the target population is somewhere between 3 and 5, and let this sensitivity parameter vary over this range. Applying the outcome-model-based method described in section 3.1.1, we have TATE ranging from -9.87, 95% CI $(-12.51, -7.23)$ (corresponding to mean baseline addiction score 3) to -5.13, 95% CI $(-7.98, -2.28)$ (corresponding to mean baseline addiction score 5).

For the weighted-outcome-model-based method, the first step is to weight the RCT sample to achieve balance in the moderators observed in both samples: race and baseline number of cigarettes per day. Exploring two methods of estimating the conditional probability of being in the RCT sample (logistic regression

Table 1: Baseline characteristics of the RCT (n=200) and target population (n=934) samples

	RCT		Target population	
Demographics				
Male gender: number (%)	105	(52.5)	587	(62.8)
Race/ethnicity: number (%)				
White	73	(36.5)	764	(81.8)
African-American	51	(25.5)	67	(7.2)
Hispanic	61	(30.5)	58	(6.2)
Multiple	15	(7.5)	45	(4.8)
Age in years: mean (SD)	42.3	(9.6)	36.9	(12.3)
Years of education: mean (SD)	11.5	(2.1)		
Employment: number (%)				
Full-time	49	(24.5)		
Part-time or student	25	(12.5)		
Retired or unemployed	126	(63.0)		
Primary substance abuse				
Primary substance of abuse: number (%)				
Opiates	113	(56.5)		
Cocaine	39	(19.5)		
Alcohol/other	48	(24.0)		
Severity of primary substance abuse: mean (SD)	0.76	(1.04)		
Cigarette smoking and addiction				
Daily number of cigarettes: mean (SD)	21.2	(11.3)	17.6	(8.4)
Number of smoking years: mean (SD)	26.4	(9.9)		
Number of quit attempts: mean (SD)	5.2	(12.6)		
Urine cotinine: mean (SD)	1209	(667)		
Addiction severity score: mean (SD)	4.05	(0.78)		
Withdrawal scale: mean (SD)	1.68	(0.98)		

SD = standard deviation.

Table 2: Treatment effects from the outcome model with interaction terms

Variable	Coefficient	95% confidence interval
Treatment	−6.15	(−14.37, −2.06)
Treatment X African-American	−4.09	(−8.09, −0.09)
Treatment X baseline daily number of cigarettes	−0.60	(−0.76, −0.43)
Treatment X baseline cigarette addiction severity	2.37	(0.33, 4.41)

This model also includes all covariates in Table 1 as main effects.

and generalized boosted models) and two sets of variables (moderators only and moderators plus other co-variates), we find that the weights generated from the logistic sample membership based on the combination of moderators and covariates observed in both samples (race, baseline cigarettes per day, age and gender) result in the best balance, for the moderators as well as the other covariates. Specifically, these weights reduced the standardized mean differences (between the target population sample and the RCT sample) for all race categories, age and gender to under 0.05; and reduced that for baseline daily number of cigarettes from 0.43 before weighting to 0.15 after weighting.

Using these weights, we fit the same outcome model with interactions to the weighted RCT sample; results are presented in Table 3. The coefficients of the interaction terms are similar to those from the unweighted model in Table 2, but their confidence intervals are wider due to the weighting. Applying the weighted-

Table 3: Treatment effects from the weighted outcome model with interaction terms

Variable	Coefficient	95% confidence interval
Treatment	-3.70	$(-17.13, 9.72)$
Treatment X African-American	-4.49	$(-8.11, -0.87)$
Treatment X baseline daily number of cigarettes	-0.55	$(-0.82, -0.27)$
Treatment X baseline cigarette addiction severity	1.77	$(-1.77, 5.30)$

This model was fit to RCT data weighted to adjust for the differential distribution of race, gender, age and baseline daily number of cigarettes between the two samples. The model includes all covariates in Table 1 as main effects.

Figure 2: Results of the *outcome-model-based* and *weighted-outcome-model-based* sensitivity analyses with the data application.
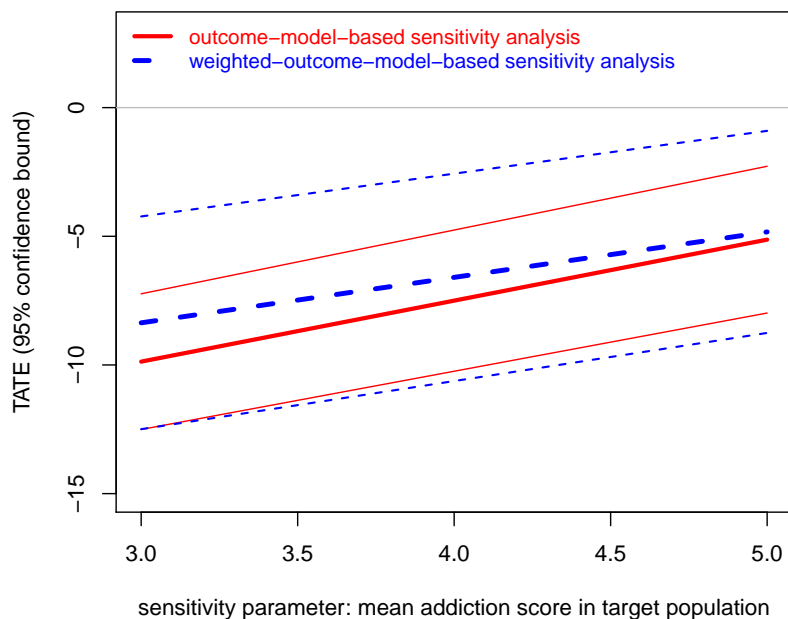
outcome-model-based sensitivity analysis, we have TATE ranging from -8.36, 95% CI $(-12.50, -4.23)$ (corresponding to mean baseline addiction score 3) to -4.83, 95% CI $(-8.76, -0.90)$ (corresponding to mean baseline addiction score 5).

Figure 2 includes the ranges of TATE with confidence bounds from both the sensitivity analyses presented here. These ranges are below zero, suggesting that this smoking cessation intervention, if applied to the target population, would result in smoking reduction. These results are generally consistent with the RCT findings, but give us more confidence in what the effects would be among the target population of cigarette smokers among people seeking alcohol/drug use treatment in the US.

6   Discussion

This paper presents approaches for assessing the sensitivity of the estimated treatment effect to an unobserved treatment effect moderator when generalizing from a RCT to a target population. The paper addresses two distinct situations. The first situation arises naturally as researchers find from analyzing RCT data that

there are variables that moderate treatment effect, but some of those variables (V) are not available from the data they have for the target population. For this case, we offer an outcome-model-based method, a method fully based on weighting, and a weighted-outcome-model-based method which combines elements of the first two. The two methods based on the outcome model are relatively straightforward and only requires specifying a plausible range for the mean of V in the target population; the weighting-based method is complicated and thus is recommended only for the case with binary moderators.

The second situation is when researchers are concerned that there may be treatment effect heterogeneity that is not detected given the observed variables in the RCT, and ask what that implies about TATE. Here we consider a composite variable U that represents the remaining effect moderation factors after accounting for the observed moderators (and possibly other covariates). For this case, we offer a bias-formula-based method and a weighting-plus-bias-formula-based method. With both methods, we vary two parameters, one representing U's association with being in the RCT, the other representing its moderation effect, and determine how TATE estimates change as a function of these parameters.

In this paper, we consider the RCT and the target population samples as disjoint sets. The proposed methods, however, are easily adapted to the situation where the RCT sample is a subset of the target population sample. In that case $S = 1$ denotes being in the RCT, but all individuals (with $S = 1$ or $S = 0$) are in the target population sample. All quantities regarding the target population are not conditioned on $S = 0$. The weighting procedures involve modeling S using the target population sample (which includes RCT subjects), and weighting RCT subjects by the inverse of their probability of being in the RCT.

There are several directions for future extension of these sensitivity analyses. First, these methods apply when we do not have outcome data for the target population. There are, however, situations where the outcome under control or a combination of outcome under control and treatment for different individuals is observed in the target population. Currently, methods exist that use target population outcome data under control, but only for generalization where all data, including moderating variables, are observed [Kern et al., 2016]. The proposed methods could be adapted to incorporate target population outcome data.

Second, the proposed methods that use weighting are based on a specific method of adjusting treatment effect estimates for the differential distribution of a moderator—adjustment by weighting. Another adjustment strategy is to fit a flexible model of the outcome as a function of treatment, covariates and interaction terms, and either impute potential outcomes for individuals in the target population as the basis to estimate TATE, or estimate treatment effects for covariate strata and average these estimates using the target population covariate distribution [see Kern, Stuart, Hill, and Green, 2016, for example]. Future work should investigate ways to extend the sensitivity analysis procedures we present here to that approach.

Third, the proposed methods do not cover the case where researchers are concerned about a specific variable (e.g., parents' education attainment) that is known or suspected (based on prior evidence) to moderate treatment effect, but is not measured by the RCT. Methods for U do not apply as the moderator in this case is a specific variable, not the remaining composite moderator. This situation is related to our first situation with a specific moderator V, except that this variable is not measured in the RCT. Future investigation is needed to extend our first set of proposed methods to this situation, perhaps using a combination of additional assumptions and external information about this variable.

Fourth, the proposed outcome-model- and bias-formula-based methods assume a linear model for the potential outcomes. In the case of a binary outcome, for example, this means assuming treatment affects the outcome on the risk difference scale. Recent work by Ding and VanderWeele [2015] shows that for a binary outcome, effects measured on the odds ratio scale tend to be less heterogenous than on the risk difference (and also risk ratio) scale. One of our next steps is to adapt these sensitivity analysis methods to effect scales that are less heterogeneous for specific outcomes, for example odds ratio for a binary outcome and rate ratio for a count outcome.

To conclude, in this paper we have presented methods for assessing sensitivity of results regarding generalizability of treatment effects to effect heterogeneity on unobserved characteristics. These methods are helpful to researchers and policy makers who are interested in the effect of a treatment for a certain population, given the common situation that not all effect moderators are measured both in the RCT and in the target population.

References

Onyebuchi a Arah, Yasutaka Chiba, and Sander Greenland. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. Annals of epidemiology, 18(8):637–46, 2008. doi: 10.1016/j.annepidem.2008.04.003.

Stephen R. Cole and Elizabeth A. Stuart. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. American Journal of Epidemiology, 172(1):107–15, jul 2010. doi: 10.1093/aje/kwq084.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. Journal of the National Cancer Institute, 22:173–203, 1959.

Peng Ding and Tyler J. VanderWeele. Generalized Cornfield conditions for the risk difference. Biometrika, 101(4):971–977, 2014. doi: 10.1093/biomet/asu030.

Peng Ding and Tyler J. VanderWeele. The differential geometry of homogeneity spaces ccross effect scales. ArXiv: 1510.08534v2. 2015.

Peng Ding and Tyler J. VanderWeele. Sensitivity analysis without assumptions. Epidemiology, 27(3):368–377, 2016. doi: 10.1097/EDE.0000000000000457.

Joseph L Gastwirth, Abba M Krieger, and Paul R Rosenbaum. Dual and simultaneous sensitivity analysis for matched pairs. Biometrika, 85(4):907–920, 1998.

Joel B Greenhouse, Eloise E Kaizar, Kelly Kelleher, Howard Seltman, and William Gardner. Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users. Statistics in Medicine, 27:1801–1813, 2008. doi: 10.1002/sim.

S Greenland. Basic methods for sensitivity analysis of biases. International Journal of Epidemiology, 25(6):1107–1116, 1996. doi: 10.1093/ije/25.6.1107.

Guanglei Hong. Ratio of mediator probability weighting for estimating natural direct and indirect effects. Proceedings of the American Statistical Association, Biometrics Section, pages 2401–2415, 2010.

Holger L. Kern, Elizabeth A. Stuart, Jennifer L Hill, and Donald P. Green. Assessing methods for generalizing experimental impact estimates to target samples. Journal of Research on Educational Effectiveness, 9(1):103–127, 2016. doi: 10.1080/19345747.2015.1060282.

Robert B Olsen, Larry L Orr, Stephen H Bell, and Elizabeth A Stuart. External validity in policy evaluations that choose sites purposively. Journal of Policy Analysis and Management, 32(1):107–121, 2013. ISSN 15206688. doi: 10.1002/pam.21660.

Malcolm S Reid, Bryan Fallon, Susan Sonne, Frank Flammino, Edward V Nunes, Huiping Jiang, Eva Kuoniotis, Jennifer Lima, Ron Brady, Cynthia Burgess, Eric Pihlgren, Louis Giordano, Aron Starosta, James Robison, and John Rotrosen. Smoking cessation treatment in community-based substance abuse rehabilitation programs. Journal of Substance Abuse Treatment, 35(1):68–77, 2008.

P R Rosenbaum. Sensitivity analysis for certain permutational inferences in matched observational studies. Biometrika, 74:13–26, 1987.

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society, 45(2):212–218, 1983a.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983b.

Sebastian Schneeweiss. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiology and Drug Safety, 15:291–303, 2006. doi: 10.1002/pds.1200.

Elizabeth A Stuart and Anna Rhodes. (In press). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. Evaluation Review.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174(2):369–386, 2011. doi: 10.1111/j.1467-985X.2010.00673.x.

Elizabeth A. Stuart, Catherine P. Bradshaw, and Philip J. Leaf. Assessing the generalizability of randomized trial results to target populations. Prevention Science, 16(3):475–485, 2015. doi: 10.1007/s11121-014-0513-z.

Ryoko Susukida, Rosa M Crum, Elizabeth A Stuart, Cyrus Ebnesajjad, and Ramin Mojtabai. Assessing sample representativeness in randomized controlled trials: Application to the National Institute of Drug Abuse Clinical Trials Network. Addiction, 2016. doi: 10.1111/add.13327.

E. Tipton. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics, 38(3):239–266, 2013. doi: 10.3102/1076998612441947.

Tyler J Vanderweele and Onyebuchi a Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology, 22(1):42–52, 2011. doi: 10.1097/EDE.0b013e3181f74493.

Herbert I Weisberg, Vanessa C Hayden, and Victor P Pontes. Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? Clinical Trials, 6(2):109–118, 2009. doi: 10.1177/1740774509102563.

Acknowledgements

Supplement A: R-code for the simulation study

Supplement B: R-code for implementing the outcome-model-based and weighted-outcome-model-based methods on the data example