Johns Hopkins University, Dept. of Biostatistics Working Papers

1-1-2014

Estimating population treatment effects from a survey sub-sample

Kara E. Rudolph
Johns Hopkins Bloomberg School of Public Health, kara.rudolph@gmail.com

Ivan Diaz
Department of Biostatistics, Johns Hopkins School of Public Health

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health, Departments of Mental Health and Biostatistics

Estimating population treatment effects from a survey sub-sample

Kara E. Rudolph[*1,2,3], Iván Díaz[†2], Michael Rosenblum[‡2] and Elizabeth A. Stuart[§2,3]

[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[3]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

May 14, 2014

Abstract

We consider the problem of estimating an average treatment effect for a target population from a survey sub-sample. Our motivating example is generalizing a treatment effect estimated in a sub-sample of the National Comorbidity Survey Replication Adolescent Supplement to the population of U.S. adolescents. To address this problem, we evaluate easy-to-implement methods that account for both non-random treatment assignment and a non-random two-stage selection mechanism. We compare the performance of a Horvitz-Thompson estimator using inverse probability weighting (IPW) and two double robust estimators in a variety of scenarios. We demonstrate that the two double robust estimators generally outperform IPW in terms of mean-squared error even under misspecification of one of the treatment, selection, or outcome models. Moreover, the double robust estimators are easy to implement, providing an attractive alternative to IPW for applied epidemiologic researchers. We demonstrate how to apply these estimators to our motivating example.

[*]krudolph@jhsph.edu
[†]idiaz@jhu.edu
[‡]mrosenbl@jhsph.edu
[§]estuart@jhsph.edu

1 Introduction

Population-based cohorts and nationally representative surveys lend external validity to a study: they allow inferences to be made about the target population of interest. In contrast, inferences drawn from studies that use non-representative samples may be valid for the study sample but may not generalize. External validity (also known as transportability (1)) of population-based cohorts and surveys is threatened when estimation is performed on a non-random sub-sample. Sub-sample effect estimates may not generalize to the population if selection probabilities depend on effect modifiers and if sub-sample sampling weights are not utilized (2-3). In this paper, we compare practical estimators of the population average treatment effect (PATE). These estimators simultaneously account for non-randomized treatment assignment and sub-sample selection from a population-based cohort, thereby addressing internal and external validity.

This paper was motivated by the problem of generalizing a treatment effect estimated in a sub-sample created by a two-stage selection process. In the first stage, adolescents were selected into a nationally representative survey of U.S. adolescent mental health, the National Comorbidity Survey Replication Adolescent Supplement (NCS-A) (4). In the second stage, a sub-sample of these participants had biomarker data measured. Our interest is in estimating the effect of a non-randomized treatment, residence in a disadvantaged neighborhood, on cortisol slope (rate of decline in cortisol levels over the course of an interview). Our scenario is different from the missing data pattern generally considered in the causal inference and missing data literature because we do not observe any data for individuals not in the survey. Our goal is to harness the available data and the nationally representative sample to generalize our results to the U.S. population of adolescents. This requires accounting for possible confounding due to the non-randomized treatment assignment and possible lack of external validity due to the two-stage selection mechanism.

Previous research has suggested and evaluated methods for generalizing results from randomized trials to target populations (2, 3), but there is little written extending this to observational studies. Double robust methods, which are consistent (converge to the true population average effect as sample size goes to infinity) under certain types of model misspecification, have been used to adjust for non-random treatment assignment and/or non-random selection, right-censoring, or missing data (5-13). However, implementation of these estimators can be challenging. This may contribute to the continued popularity of the simpler Horvitz-Thompson inverse-probability weighted (IPW) estimators despite efficiency and robustness concerns (14). A recent advance is that targeted maximum likelihood estimation (TMLE) was implemented in standard statistical software (11), thereby facilitating its accessibility. However, we know of no literature implementing TMLE in the context of survey data with weights. Furthermore, while a discussion of these methods is taking place in the biostatistics literature, it has yet to receive much attention in the epidemiology literature.

We first present results from a simulation study comparing performance of different estimators under correct model specification and various model misspecifications. We compare three estimators: IPW; a double robust, weighted least squares estimator (DRWLS); and a TMLE (15, 16). We then demonstrate how to apply these methods to the motivating example: using data from a

sub-sample of the NCS-A to estimate the effect of residence in a disadvantaged neighborhood on cortisol slope in the target population of U.S. adolescents. We aim to provide practical guidance on how to generalize average effect estimates from a survey sub-sample to a target population in the presence of measured confounders, effect heterogeneity, and non-random sub-sample selection.

2 Methods

We consider a scenario in which individuals are selected into a survey with known probabilities. Treatment information and covariates are fully observed for all participants selected into the survey, but outcome data are only available for a subset of the survey sample. Let:

W = vector of baseline covariates.

A = binary (0/1) variable indicating treatment.

Δ_{svy} = binary (0/1) variable indicating selection into survey sample.

Δ_{sub} = binary (0/1) variable indicating selection into sub-sample.

Y = continuous outcome of interest.

In the language of potential outcomes, Y_{1i} is the outcome for individual i under treatment $A = 1$; similarly, Y_{0i} is the outcome for individual i under treatment $A = 0$. The difference in these potential outcomes is the individual treatment effect.

Our estimand of interest is the PATE, $E(Y_1 - Y_0)$, with the expectation taken across the target population.(17) Other average treatment effects (ATEs) could be considered where the expectation is taken with respect to different target populations, e.g., the survey sample ATE, $E(Y_1 - Y_0|\Delta_{svy=1})$, and the sub-sample ATE, $E(Y_1 - Y_0|\Delta_{sub} = 1)$. In Appendix A, we show identification for each ATE under the assumptions of known survey sampling weights (a typical assumption in the survey literature (18)), no unmeasured confounders, consistency, and positivity.

We compare three methods of estimating the PATE. R code to implement each is provided in Appendix B.

2.1 IPW

The IPW estimator uses inverse probability of treatment and selection weights that are obtained by multiplying inverse probability of survey selection weights, inverse probability of treatment weights, and inverse probability of sub-sample selection weights as shown in Equation 1.
Inverse probability of survey selection weights are known and defined as:

$$w^{\Delta_{svy}=1} = \frac{1}{P(\Delta_{svy} = 1|\mathbf{W})}.$$

Inverse probability of treatment weights are defined as:

$$w^{A=a|\Delta_{svy}=1} = \frac{I(A=a)}{P(A=a|\Delta_{svy}=1,\mathbf{W})} \text{ for each } a \in \{0,1\}.$$

Inverse probability of sub-sample selection weights are defined as:

$$w^{\Delta_{sub}=1|A=a,\Delta_{svy}=1} = \frac{I(\Delta_{sub}=1)}{P(\Delta_{sub}=1|A=a,\Delta_{svy}=1,\mathbf{W})} \text{ for each } a \in \{0,1\}.$$

For each $a \in \{0,1\}$, define:

$$\begin{aligned} w^{A=a,\Delta_{svy}=1,\Delta_{sub}=1} &= \frac{1}{P(\Delta_{svy}=1|\mathbf{W})} \times \frac{I(A=a)}{P(A=a|\Delta_{svy}=1,\mathbf{W})} \\ &\times \frac{I(\Delta_{sub}=1)}{P(\Delta_{sub}=1|A=a,\Delta_{svy}=1,\mathbf{W})} \\ &= \frac{I(A=a,\Delta_{sub}=1,\Delta_{svy}=1)}{P(A=a,\Delta_{sub}=1,\Delta_{svy}=1|\mathbf{W})} \end{aligned} \tag{1}$$

The above weights are inverse conditional probabilities, which can be estimated using logistic regression. For example, $P(A=a|\Delta_{svy}=1,\mathbf{W})$ can be estimated using predicted probabilities from a logistic regression where A is the outcome and $\mathbf{W}$ is a vector of covariates among those in the survey.

The IPW estimator of the PATE is calculated using the above weights for the r individuals in the sub-sample:

$$\widehat{PATE} = \frac{\sum_{i=1}^{r} Y_i w_i^{A=1,\Delta_{svy=1},\Delta_{sub=1}}}{\sum_{i=1}^{r} w_i^{A=1,\Delta_{svy=1},\Delta_{sub=1}}} - \frac{\sum_{i=1}^{r} Y_i w_i^{A=0,\Delta_{svy=1},\Delta_{sub=1}}}{\sum_{i=1}^{r} w_i^{A=0,\Delta_{svy=1},\Delta_{sub=1}}}. \tag{2}$$

2.2 TMLE

For the TMLE estimator, we modify the implementation available in the tmle R package (11) as described in Appendix B. Additional details of TMLE implementation for estimating an ATE with survey sample data are provided in Appendix A. Below we summarize the main steps involved.

Below we summarize the main steps involved.

1. Obtain predicted values $\hat{Y}^0$ of the outcome conditional on the treatment and covariates using a linear regression of Y as a function of A and $\mathbf{W}$ among participants for whom Y is observed. Although we use a linear regression for comparability with DRWLS, described below, it is also possible to use data-adaptive methods (e.g., machine learning) and for a variety of outcome types.

2. For every individual i, compute the covariate: $H_i = A_i w_i^{A=1,\Delta_{sub}=1,\Delta_{svy}=1} - (1-A_i) w_i^{A=0,\Delta_{sub}=1,\Delta_{svy}=1}$.

3. Compute the estimated coefficient $\hat{\beta}$ in a regression of Y on H using $\hat{Y}^0$ as an offset. Using G-computation, compute the difference between the counterfactual outcomes predicted by this regression under assignment to $A = 1$ and to $A = 0$ for each participant in the survey sample. The TMLE is calculated as a weighted sum of this difference across the survey sample participants using the survey weights.

For readers unfamiliar with G-computation, Snowden et al (2011) provide an introduction (19). Briefly, G-computation uses the marginal distribution of covariates in a standardization procedure. This can be thought of as an extension of standardizing mortality rates by the age distribution in a standard population—a common epidemiologic practice. One fits a model of the outcome as a function of treatment and covariates in the observed sample and then applies the model to the distribution of covariates in the standard population to predict the counterfactual outcomes for each individual.

2.3 DRWLS

The DRWLS estimator combines weighted regression with G-computation. This estimator was first suggested by Marshall Joffe (14) and has been previously evaluated (5,14). It uses the following steps:

1. Use $w^{A=a,\Delta_{sub}=1,\Delta_{svy}=1}$ as weights in a weighted least squares (WLS) linear regression of Y given A and $\mathbf{W}$ among participants in the sub-sample. Using G-computation, predict counterfactual outcomes, standardized to the survey sample.

2. Use the counterfactual outcomes to estimate the survey sample ATE.

3. Weight this estimate by the survey weights to estimate the PATE.

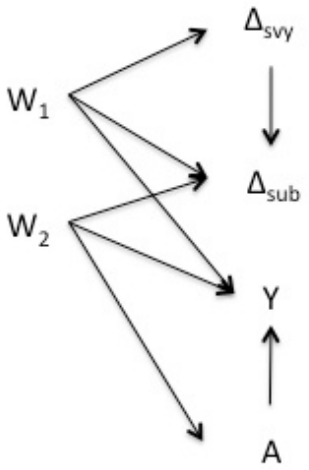
TMLE and DRWLS are double robust estimators. In our application, double robustness means that the estimators are consistent if either the outcome model is correct, or if the combined treatment-selection weights $w_i^{A=a,\Delta_{sub}=1,\Delta_{svy}=1}$ from Equation 1 are correct.

3 Simulation Study

3.1 Overview and set-up

We consider a simplified case with two continuous covariates: $\mathbf{W} = [W_1, W_2]'$. Let observed data O = ($\Delta_{svy} = 1$, $\mathbf{W}$, A, Δ_{sub}, $\Delta_{sub}Y$). We assume Δ_{svy} probabilities are known, there are no unobserved confounders, and no intermediate variables are observed between A and Y. Figure 1 depicts the data-generating mechanism.

Figure 1: Data generating mechanism.



The simulation is designed to be similar to the case study, detailed further below. Figure 2 provides a diagram of the complete data and the observed data. Under the causal inference framework we are using, it is assumed that the observed data generating process consists of several steps, the order of which is necessary for the identification result and corresponding methods implementation (20). First, selection into the survey is determined, where the probability of selection depends on W_1. For researchers designing the survey, W_1 is known for all individuals in the population. For all other analysts, W_1 is not observed for those not selected into the survey. Second, for all individuals selected into the survey, we observe W_2 and A, where the probability of $A = 1$ depends on W_2. Third, selection into the sub-sample (Δ_{sub}) is determined, where the probability of selection depends on W_1 and W_2. Fourth, for those in the sub-sample, we observe one of two counterfactuals, Y_0 or Y_1, corresponding to the treatment A actually received. We include a detailed description of the data generating process and code to implement it in Appendix C.

As seen in Figure 1, W_2 acts as a confounder. W_1 directly modifies the treatment effect and is related to selection into the survey and sub-sample. Figure 3 provides a summary of imbalance in W_1, W_2, and Y across treatment groups and a summary of imbalance of W_1, W_2, and A across selection groups. A consistent estimate of the sub-sample ATE requires adjusting for confounding by W_2. A consistent estimate of the PATE also requires adjusting for differential selection by W_1.

The simulation reflects practical positivity violations (i.e., when subsets of the sample have large weights) similar to the case study. Table 1 gives the treatment and selection weights for both the simulation and case study.

Figure 2: Simulation Set-up. X indicates data present.

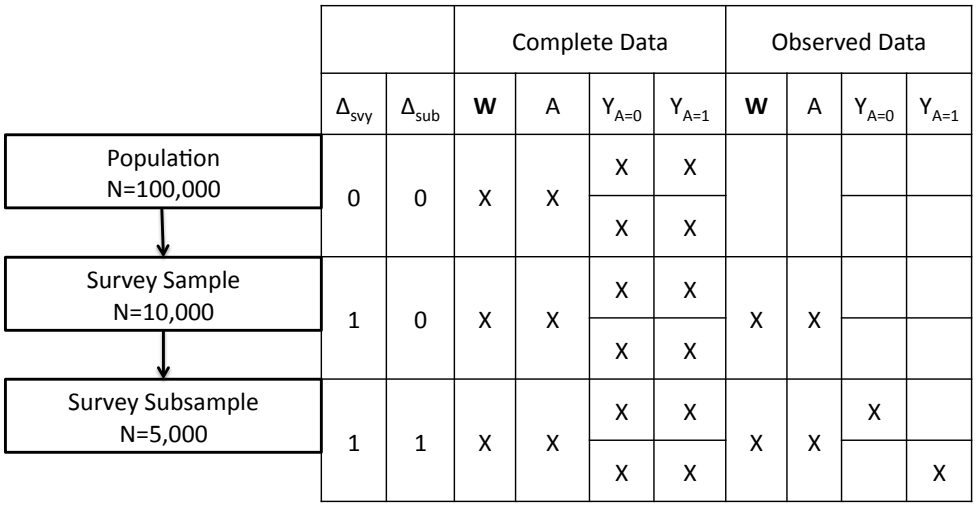


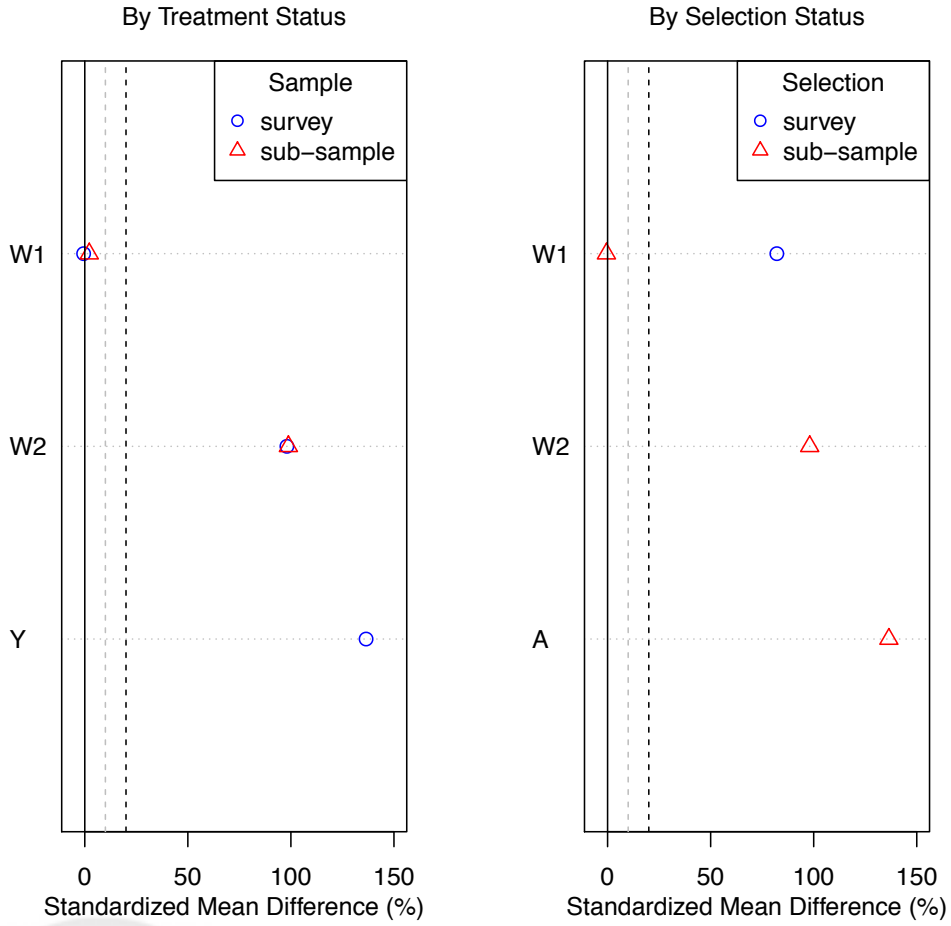
Table 1: Characteristics of Selection and Treatment Weights in Simulation and Case Study.[1][2]

	Simulation			Case Study		
Weights	Mean	SD	Min/Max	Mean	SD	Min/Max
$w^{\Delta_{svy}=1\|\mathbf{W}}\|\Delta_{svy}=1$	1.009	1.009	0.020/20.300	0.948	1.166	0.041/13.460
$w^{A=1\|\Delta_{svy}=1,\mathbf{W}}\|A=1$	0.994	0.607	0.529/13.600	0.931	0.878	0.377/8.799
$w^{A=0\|\Delta_{svy}=1,\mathbf{W}}\|A=0$	0.998	0.655	0.491/16.580	1.020	0.826	0.641/19.180
$w^{\Delta_{sub}=1\|A=1,\Delta_{svy}=1,\mathbf{W}}\|\Delta_{sub}=1$	1.009	0.562	0.339/7.148	0.995	0.631	0.332/5.914
$w^{\Delta_{sub}=1\|A=0,\Delta_{svy}=1,\mathbf{W}}\|\Delta_{sub}=1$	1.003	0.593	0.280/6.044	1.002	0.574	0.401/5.598
$w^{\Delta_{sub}=1,A=1\|\Delta_{svy}=1,\mathbf{W}}\|\Delta_{sub}=1,A=1$	0.992	1.177	0.182/22.620	0.883	1.110	0.152/14.560s
$w^{\Delta_{sub}=1,A=0\|\Delta_{svy}=1,\mathbf{W}}\|\Delta_{sub}=1,A=0$	0.995	0.565	0.269/7.907	1.069	1.558	0.315/24.860
$w^{\Delta_{sub}=1,A=1,\Delta_{svy}=1\|\mathbf{W}}\|\Delta_{sub}=1,A=1$	1.023	3.421	0.006/84.080	1.079	2.699	0.009/34.380
$w^{\Delta_{sub}=1,A=0,\Delta_{svy}=1\|\mathbf{W}}\|\Delta_{sub}=1,A=0$	0.993	1.687	0.012/17.910	1.303	2.990	0.019/53.160

[1]The simulation weights shown are the true weights. In the case study, the survey weights are assumed known and the remaining weights are estimated.

[2]Weights are stabilized by including the marginal probability in the numerator to facilitate comparison with simulation weights. For example, $w^{A=1|\Delta_{svy}=1}|A=1=\frac{mean(P(A=1|\Delta_{svy}=1,\mathbf{W}))}{P(A=1|\Delta_{svy}=1,\mathbf{W})}$.

Figure 3: Balance across treatment and selection groups in the simulation. The standardized mean difference is the difference in means between the two groups standardized by the standard deviation in the first group. The vertical black dashed line corresponds to 20% standardized mean difference and the grey dashed line corresponds to 10%.



We evaluate how well IPW, TMLE, and DRWLS perform in estimating the PATE when all models are correctly specified and when one or more models are misspecified (see Table 2 for model specifications). We include misspecification of multiple models simultaneously but caution that performance in these scenarios depends on the particulars of the data generating process and misspecifications (14). Performance is evaluated by mean percent bias, mean variance, mean-squared error (MSE), and 95% CI coverage across the 1,000 simulations. For each simulation iteration, variance and the 95% CI are estimated from 500 bootstrapped samples. The percentile method is used for the CI.

Table 2: Model Misspecification.

Description	Treatment	Sub-sample Selection	Outcome
Correct Specification	$A \sim W_2$, W_2=Sum(Z1:Z16)	$\Delta_{sub} \sim W_1$ + W_2, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)	$Y \sim A + W_2 + AW_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Moderately Misspecified Treatment	$A \sim W_{2mod}$, W_{2mod}=Sum(Z2:Z16)	$\Delta_{sub} \sim W_1$ + W_2, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)	$Y \sim A + W_2 + AW_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Majorly Misspecified Treatment	$A \sim W_1$, W_1=Sum(Z1:Z6)	$\Delta_{sub} \sim W_1$ + W_2, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)	$Y \sim A + W_2 + AW_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Moderately Misspecified Outcome	$A \sim W_2$, W_2=Sum(Z1:Z16)	$\Delta_{sub} \sim W_1$ + W_2, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)	$Y \sim A + W_{2mod} + AW_1$, W_1=Sum(Z1:Z6), W_{2mod}=Sum(Z2:Z16)
Majorly Misspecified (A) Outcome	$A \sim W_2$, W_2=Sum(Z1:Z16)	$\Delta_{sub} \sim W_1$ + W_2, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)	$Y \sim A + AW_1$, W_1=Sum(Z1:Z6)
Majorly Misspecified (B) Outcome	$A \sim W_2$, W_2=Sum(Z1:Z16)	$\Delta_{sub} \sim W_1$ + W_2, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)	$Y \sim A + W_2 + W_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Moderately Misspecified Selection	$A \sim W_2$, W_2=Sum(Z1:Z16)	$\Delta_{sub} \sim W_{1mod}$ + W_2, W_{1mod}=Sum(Z2:Z6), W_2=Sum(Z1:Z16)	$Y \sim A + W_2 + AW_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Majorly Misspecified Selection	$A \sim W_2$, W_2=Sum(Z1:Z16)	$\Delta_{sub} \sim W_1$, W_1=Sum(Z1:Z6)	$Y \sim A + W_2 + AW_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Misspecified Treatment and Selection	$A \sim W_1$, W_1=Sum(Z1:Z6)	$\Delta_{sub} \sim W_1$, W_1=Sum(Z1:Z6)	$Y \sim A + W_2 + AW_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Misspecified Treatment and Outcome	$A \sim W_1$, W_1=Sum(Z1:Z6)	$\Delta_{sub} \sim W_1$ + W_2, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)	$Y \sim A + W_2 + W_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Misspecified Selection and Outcome	$A \sim W_2$, W_2=Sum(Z1:Z16)	$\Delta_{sub} \sim W_1$, W_1=Sum(Z1:Z6)	$Y \sim A + W_2 + W_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)
Misspecified Treatment, Selection, and Outcome	$A \sim W_1$, W_1=Sum(Z1:Z6)	$\Delta_{sub} \sim W_1$, W_1=Sum(Z1:Z6)	$Y \sim A + W_2 + W_1$, W_1=Sum(Z1:Z6), W_2=Sum(Z1:Z16)

3.2 Results

Table 3 provides a summary of method performance. Generally, under correct and incorrect model specification, DRWLS outperforms TMLE and IPW—corroborating results of other simulations involving practical positivity violations (5, 14). Under correct specification of all models, DRWLS and TMLE perform similarly and outperform IPW in terms of variance and MSE. This result reflects known efficiency problems with IPW that have been thoroughly discussed in the biostatistics literature, but are perhaps less well known among epidemiology audiences (14). TMLE and IPW perform similarly and worse than DRWLS in terms of percent bias and 95% CI coverage under correct model specification. This result may seem surprising, and we discuss it further below.

The advantages of DRWLS and TMLE over IPW are pronounced under misspecification of the treatment or selection models. This result is expected, because IPW relies exclusively on the inverse probability weights to account for nonrandom sub-sample selection and nonrandom treatment assignment. In contrast, because DRWLS and TMLE are double robust, they will be consistent under misspecification of the treatment or sub-sample selection models if the outcome model is correctly specified.

Under correct specification of all models, we expect IPW, DRWLS, and TMLE estimates to be consistent. However, several authors (e.g., 5-7, 14, 21, 22) have warned that IPW and double robust estimators are sensitive in scenarios of practical positivity violations, as is the case in this simulation. We may expect IPW to be the most sensitive to positivity violations, because there is no outcome model to use for extrapolation. When the outcome model is correctly specified, we expect DRWLS and TMLE to outperform IPW due to successful extrapolation using the outcome model. This is true for DRWLS but not for this implementation of TMLE, which performs similarly to IPW in terms of percent bias and 95% CI coverage. This is because in fitting the model used in G-computation, DRWLS uses the combined treatment-selection weights for the treatment and selection conditions actually observed whereas TMLE uses the combined treatment-selection weights for the observed and unobserved counterfactual treatment and selection conditions. If individuals usually receive their most likely treatment and selection assignment, then using the counterfactuals can result in greater positivity violations, and thus, poorer performance of TMLE as compared to DRWLS.

We examined the extent to which there is a penalty for unnecessary adjustment for non-random treatment assignment or sample selection. We considered two scenarios for each of the treatment and sub-sample selection models (see Table 4). In this limited simulation, there are no noticeable penalties for over-adjusting. Table 5 shows the results from the more extreme second scenario. Results from the first scenario were similar.

Table 3: Method Performance Under Correct Specification and Misspecification. Mean % Bias, Mean Variance (Var), 95% CI Coverage (Cov), and Mean-squared Error (MSE) across the 1,000 Simulations.

	IPW				DRWLS				TMLE			
Correct Specification	-6.7	8.225	85.0	10.978	0.2	0.203	93.0	0.219	-8.4	0.252	77.9	0.402
Moderately Misspecified Treatment	-30.3	7.708	79.3	10.692	-0.1	0.203	94.4	0.207	-7.2	0.234	81.7	0.334
Majorly Misspecified Treatment	-176.0	5.154	15.8	55.133	0.0	0.202	94.4	0.206	-7.2	0.225	81.7	0.326
Moderately Misspecified Outcome	-6.7	8.225	85.0	10.978	-0.8	0.266	94.5	0.282	-6.1	0.306	87.2	0.376
Majorly Misspecified (A) Outcome	-6.7	8.225	85.0	10.978	-3.4	1.031	91.0	1.168	-0.2	0.920	93.5	1.104
Majorly Misspecified (B) Outcome	-6.7	8.225	85.0	10.978	-13.1	2.247	79.8	3.285	3.4	10.441	85.5	14.800
Moderately Misspecified Selection	-197.6	3.677	9.0	66.382	-0.2	0.193	94.9	0.193	-7.2	0.218	81.0	0.320
Majorly Misspecified Selection	-6.8	5.701	87.8	6.230	-0.1	0.201	94.2	0.208	-7.2	0.230	80.3	0.329
Majorly Misspecified Treatment and Selection	-169.2	4.174	15.1	50.242	-0.8	0.198	94.5	0.203	-7.7	0.220	80.8	0.326
Majorly Misspecified Treatment and Outcome	-176.0	5.154	15.8	55.133	-10.1	2.359	83.4	3.039	-14.9	9.796	84.6	14.640
Majorly Misspecified Selection and Outcome	-6.8	5.701	87.8	6.230	-15.3	2.148	80.8	3.009	-68.7	9.697	89.4	23.395
Majorly Misspecified Treatment, Selection, and Outcome	-169.2	4.174	15.1	50.242	-15.1	2.339	82.3	3.235	-6.6	8.986	85.9	13.278

Table 4: Model Misspecification, Overadjustment.

Description	Treatment	Sub-sample Selection	Outcome
Moderate Over-adjustment, Treatment			
True Model	random	$\Delta_{sub} \sim W_1 + W_2$	$Y \sim A + W_2 + AW_1$
Misspecified Model	$A \sim W_2$	$\Delta_{sub} \sim W_1 + W_2$	$Y \sim A + W_2 + AW_1$
Major Over-adjustment, Treatment			
True Model	random	$\Delta_{sub} \sim W_1 + W_2$	$Y \sim A + W_2 + AW_1$
Misspecified Model	$A \sim W_2 + W_2^2$	$\Delta_{sub} \sim W_1 + W_2$	$Y \sim A + W_2 + AW_1$
Moderate Over-adjustment, Selection			
True Model	$A \sim W_2$	$\Delta_{sub} \sim W_1$	$Y \sim A + W_2 + AW_1$
Misspecified Model	$A \sim W_2$	$\Delta_{sub} \sim W_1 + W_2$	$Y \sim A + W_2 + AW_1$
Major Over-adjustment, Selection			
True Model	$A \sim W_2$	random	$Y \sim A + W_2 + AW_1$
Misspecified Model	$A \sim W_2$	$\Delta_{sub} \sim W_1 + W_2$	$Y \sim A + W_2 + AW_1$

Table 5: Results Under Misspecification of the Treatment and Selection Models: Adjustment When the Treatment and Selection Mechanisms are Completely Random. Mean % Bias, Mean Variance (Var), 95% CI Coverage (Cov), and Mean-squared Error (MSE) across the 1,000 Simulations.

	Correct Specification			Overadjustment		
	IPW	DRWLS	TMLE	IPW	DRWLS	TMLE
Treatment						
% Bias	-1.887	-0.036	-6.554	-1.967	-0.037	-6.554
Var	88.8	95.1	82.6	89.4	95.1	82.5
Cov	6.625	0.195	0.227	6.619	0.195	0.227
MSE	7.330	0.197	0.296	7.291	0.197	0.296
Selection						
% Bias	-1.330	-0.731	-2.696	-1.855	-0.733	-2.699
Var	93.5	94.5	92.3	93.0	94.5	92.2
Cov	2.358	0.179	0.178	2.208	0.179	0.178
MSE	2.353	0.178	0.192	2.269	0.178	0.192

4 Case Study

4.1 Overview and set-up

We now apply the estimators evaluated in the simulation to generalize the effect of disadvantaged neighborhood residence on cortisol slope to the population of U.S. adolescents. The NCS-A has been described previously (4, 23-25). Neighborhood disadvantage was measured using an established scale (26) that has been used previously with NCS-A residence data geocoded to Census tracts (27). Cortisol is a hormone involved in the hypothalamic-pituitary-adrenal axis (28). Salivary cortisol samples were taken immediately before and after the survey interview. Cortisol slope was defined as (pre-interview level – post-interview level) / length of interview. Cortisol samples were assayed for a sub-sample of 2,490 participants because of budget limitations. Treatment and covariate data were available for all participants. Analysis of the relationship between neighborhood disadvantage and cortisol slope among the sub-sample of participants with cortisol data has been previously reported (27). We excluded those adolescents whose cortisol levels may not be at-risk to be influenced by a stressful neighborhood environment (e.g., current smokers, drug users, those on birth control, steroid inhalers) as well as eight influential outliers for a total of N=6,566 participants in the survey sample and N=1,600 participants with cortisol measures. Informed assent and consent were obtained from each adolescent and his/her parent or guardian. The Human Subjects Committees of Harvard Medical School and the University of Michigan approved recruitment and consent procedures.

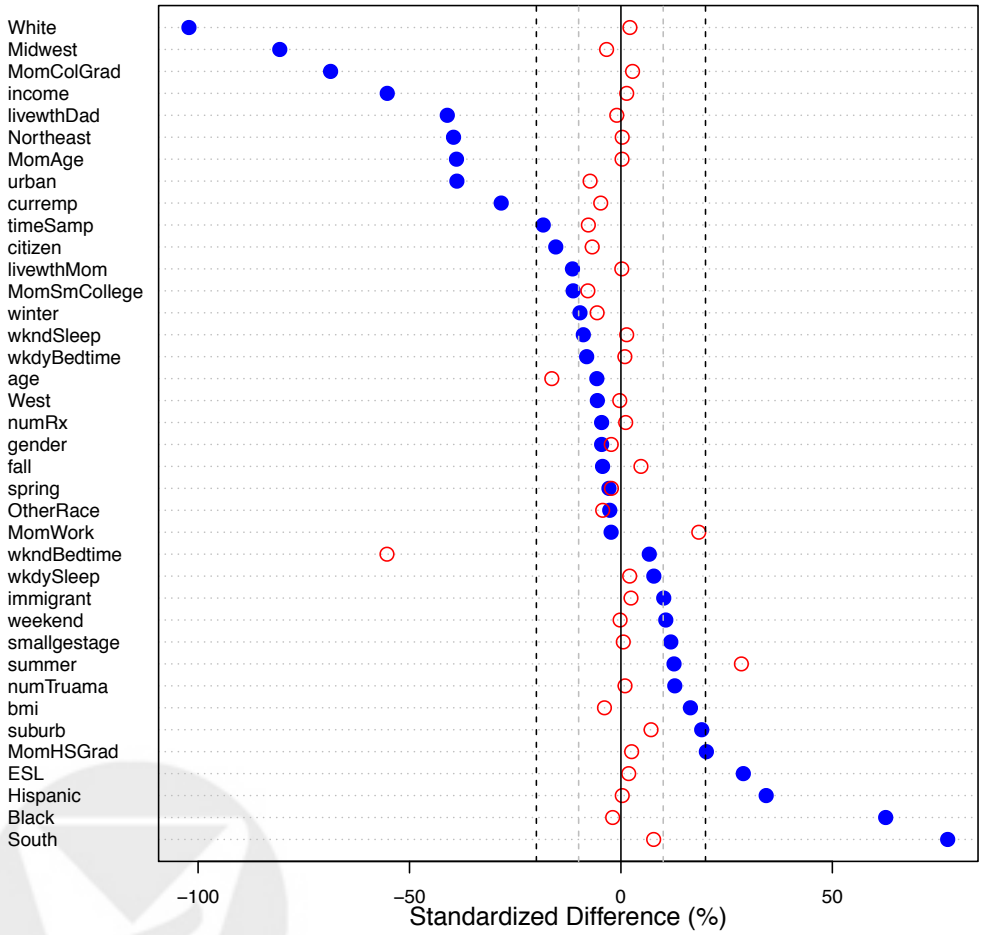
Figure 4 depicts the extent to which 1) NCS-A participants with cortisol measures compare to participants without across possible confounding variables (open dots), and 2) NCS-A participants living in disadvantaged neighborhoods compare to those in non-disadvantaged neighborhoods. Those with and without cortisol measures look similar with the exceptions of age, average bedtime on the weekends, maternal work, and the interview taking place in the summer. In contrast, participants living in disadvantaged neighborhoods differ from those in non-disadvantaged neighborhoods in terms of expected demographic variables like race, income, and maternal education.

The survey weights and estimated inverse probability of treatment and selection weights for these case study data are shown in Table 1. Positivity violations can be a substantial issue in observational studies (29, 30), and we see evidence of practical positivity violations here. The survey weights are given and assumed known (25). Unlike in the simulation, we do not know the true treatment and selection models. Consequently, it is likely that multiple models are misspecified with the degree of misspecification unknown. For details on model specification and weight estimation, see Appendix D.

4.2 Results

Figure 5 plots the estimates and 95% CIs for the expected effect of living in a disadvantaged neighborhood on cortisol slope using different methods. The 95% CIs are calculated by the percentile method using 1,000 bootstrapped samples. Results when the eight outliers were included are shown

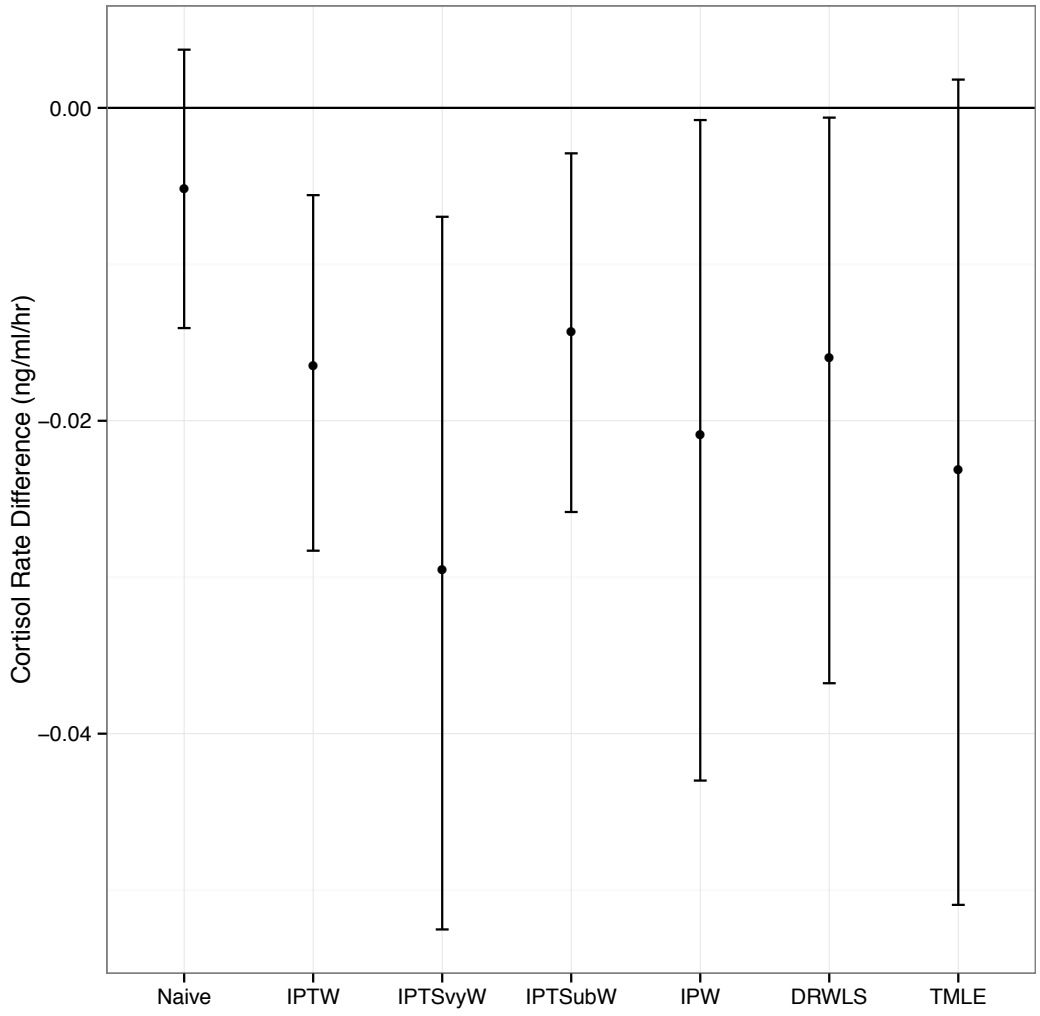
Figure 4: Covariate balance. Solid points represent the standardized mean differences between the disadvantaged neighborhood group and non-disadvantaged neighborhood group. Open points represent the standardized mean differences between those with cortisol measurement and those without. The standardized mean difference is the difference in means between the two groups standardized by the standard deviation in the first group. The vertical black dashed lines correspond to 20% standardized mean difference and the grey dashed lines correspond to 10%.



in Appendix E. The relative performance of the estimators was similar.

We first present simpler methods that adjust for none or only some of the sources of non-randomness. These methods may be biased for the primary estimand of interest—the PATE—but may consistently estimate other estimands, specifically, the survey sample ATE or the sub-sample ATE.

Figure 5: Illustrative example: population average effect estimates and 95% confidence intervals using data from the NCS-A sub-sample.



Under the naïve approach, there is no bias correction. The IPTW estimator adjusts for non-random assignment of the treatment only. If the treatment model is correctly specified, the IPTW estimate will be consistent for the NCS-A sub-sample ATE but may be biased for the PATE. IPTSvyW adjusts for non-random treatment assignment and selection into the survey. If there is nonrandom sub-sample selection, it may be biased for the population and survey sample ATEs. IPTSubW adjusts for non-random treatment assignment and non-random sub-sample selection. If the two models are correctly specified, IPTSubW will consistently estimate the survey sample ATE but may be biased for the PATE. In the presence of treatment effect heterogeneity when sub-sample selection probabilities depend on effect modifiers, the sub-sample ATE, survey sample ATE, and PATE may differ. We see evidence of this in Figure 5. Although the estimates of the sub-sample

ATE and survey sample ATE appear similar, they are slightly smaller than the estimate of the PATE.

The IPW, TMLE, and DRWLS estimators adjust for non-random assignment of the treatment, non-random selection into the survey sample, and non-random selection into the sub-sample. Under correct model specification, these methods are consistent estimators of the PATE. TMLE and DRWLS have the additional advantage of double robustness. Using DRWLS, we conclude that the cortisol rate difference comparing U.S. adolescents in disadvantaged versus non-disadvantaged neighborhoods likely falls between -3.68 and -0.06×10^{-2} ng/mL/hour.

The wider TMLE 95% CI relative to IPW may seem surprising given that TMLE was more efficient than IPW in the simulation under correct model specification and under most model misspecifications. However, in scenarios where the outcome regression model was misspecified to exclude treatment effect heterogeneity (either alone or in addition to other misspecified models, including the realistic scenario where all models are misspecified), TMLE was not more efficient than IPW. In these scenarios, the TMLE CI was wider than the IPW and DRWLS CIs by amounts as or more extreme than the case study in several hundred of the 1,000 simulation draws. This reflects the fact that under practical positivity violations and model misspecification, TMLE is not necessarily expected to have smaller CI width than IPW. Just as we assess whether there are penalties for unnecessarily adjusting for non-random treatment and non-random sub-sample selection in the simulation study (see Table 5), we compare case study results using more parsimonious and less parsimonious models. When more parsimonious sub-sample selection models are used, the CI of each of the estimators slightly narrows but relative performance stays the same. Interval width is insensitive to parsimony in the treatment model (see Appendix F).

5 Discussion

We evaluate estimators of the PATE in the presence of treatment effect heterogeneity, non-random treatment assignment, a two-stage selection process, and practical positivity violations. Using a simulation study, we find that a DRWLS estimator and a TMLE estimator have lower MSE than an IPW estimator under correct model specification and in all but two model misspecification scenarios. DRWLS has the lowest percent bias, variance, and best CI coverage under correct model specification and in most model misspecification scenarios. We derive the efficient influence function and present a TMLE estimator incorporating survey sampling weights in Appendix A, which can be easily implemented using the available tmle package in R (code presented in Appendix B).

We agree with others (2, 3, 6, 14) that estimating an average effect standardized to a population of interest is a practical goal. It can aid in the interpretability and applicability of a study's conclusions, provided one recognizes the assumptions and limitations involved. First, a PATE will not provide information about treatment effect heterogeneity. Second, estimation can be difficult in the presence of positivity violations. In cases where the weights are highly variable, a sensitivity analysis varying model specifications is recommended (14), and there exist methods to identify possible resulting biases (22, 31). Non-parametric methods of model specification may improve robustness to model misspecification (32).

Our demonstration of the poor performance of IPW is not new. IPW estimators have well-known efficiency problems and can be biased due to structural or practical positivity violations (14). Much has been published on this in the biostatistics literature (e.g., 7, 14), but IPW continues to be widely used by epidemiologists—perhaps because it is straightforward to implement in standard statistical software. We hope by demonstrating the similarly straightforward implementation of DRWLS and TMLE coupled with their superior performance over IPW in terms of MSE, use of these estimators may gain popularity.

We evaluated the robustness of our simulation results in a series of sensitivity analyses. First, we truncated the most extreme 2% of treatment and selection weights. This may lessen both bias and variance due to extreme weights, though it may also increase bias due to misspecification (22, 33). Generally, this resulted in higher percent bias across estimators (bias was particularly high for IPW), smaller variance, larger MSE for IPW, and lower MSE for DRWLS and TMLE. Optimal truncation strategies have been examined (34); identifying the best one for the data generating mechanism considered here is an area for future work. Second, we reran the simulations removing positivity violations in the data generating distributions. This resulted in consistent estimates, 95% CI coverage of approximately 95%, and substantially lower variance and MSE across estimators when the models were correctly specified. In this case, TMLE and DRWLS clearly outperformed IPW. Third, we modified the data generating mechanism so that both W_1 and W_2 were associated with probability of treatment and probability of selection. Performance of DRWLS was similar and performance of IPW and TMLE worsened due to greater positivity violations. Fourth, we repeated simulations under no effect heterogeneity. Weights were unchanged in this scenario, but finite sample bias improved due to less data sparsity. Percent bias and CI coverage improved for IPW and TMLE. Variance and MSE improved for all estimators.

In our simulation and example, we considered a scenario where the full set of covariates was measured in the larger survey sample and selection into the sub-sample only affected missingness of the outcome variable. One could also conceive of scenarios where the general missing data pattern (due to non-response, sample selection, or right-censoring) extends to some subset of covariates. We explain how such a scenario would alter our assumptions and estimator performance in Appendix F.

Our simulation study has some limitations. First, simulations can only give a rough approximation of the sampling distribution of the estimators (14). Second, there are nearly a dozen or more estimators that could have been assessed and compared, including other implementations of TMLE (6, 13, 35). We chose to focus on a smaller set of estimators that are particularly straightforward to implement and used the implementation of TMLE that is available in the R software package (11). Other estimators and TMLE implementations may have outperformed those we considered, in particular the TMLE implementation where the weights are incorporated into a weighted logistic regression model for the updated outcome expectation instead of as part of covariate H (defined in Step 2 of the TMLE description) (36). Future work should develop easy-to-use software packages implementing these estimators. Third, the approach shown in this paper is not a fully design-based survey analysis. For example, we ignore survey sampling strata in our bootstrapping procedure. This is another area for future work.

In conclusion, we compared estimators of an average effect standardized to a target population in the presence of non-random treatment assignment, a two-stage selection process, treatment effect heterogeneity, and practical positivity violations. This scenario can apply to generalizing results from a survey sub-sample to a specified target population (2, 3). We demonstrated that DRWLS and TMLE estimators outperform an IPW estimator in terms of MSE and that DRWLS generally performs best in terms of percent bias, variance, and CI coverage in our practical positivity scenario, even under misspecification of one or more of the treatment, selection, or outcome models. Moreover, DRWLS and TMLE are easy to implement. Lastly, we demonstrated how DRWLS and TMLE estimators can be applied to everyday research questions, providing an attractive alternative to IPW for applied epidemiologic researchers.

References

1. Pearl J, Bareinboim E. Transportability of causal and statistical relations: a formal approach [technical report R-372-A]. Presented at the 25th AAAI Conference on Artificial Intelligence, San Francisco, California, August 7-11, 2011.

2. Stuart EA, Cole SR, Bradshaw CP, et al. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc Ser A Stat Soc. 2011;174(2):369–386.

3. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations the actg 320 trial. Am J Epidemiol. 2010;172(1):107–115.

4. Merikangas KR, Avenevoli S, Costello EJ, et al. National comorbidity survey replication adolescent supplement (NCS-A): I. background and measures. J Am Acad Child Adolesc Psychiatry. 2009;48(4):367–379.

5. Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007;22(4):523– 539.

6. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc. 1999;94(448):1096–1120.

7. Robins JM, Rotnitzky A, Zhao LP. Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc. 1995;90(429):106–121.

8. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–973.

9. van der Laan MJ. Targeted maximum likelihood based causal inference: Part I. Int J Biostat. 2010;6(2).

10. Tsiatis AA, Davidian M, Zhang M, et al. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. Stat Med. 2008;27(23):4658–4677.

11. Gruber S, van der Laan M. tmle: An R package for targeted maximum likelihood estimation. J Stat Softw. 2012;51(13).

12. Petersen ML, Schwab J, Gruber S, et al. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. UC Berkeley Working Paper Series. 2013:312.

13. Kim JK, Haziza D. Doubly robust inference with missing data in survey sampling. Presented at the Joint Statistical Meetings: Section on Survey Research Methods, Vancouver, British Columbia, July 31-August 5, 2010.

14. Robins J, Sued M, Lei-Gomez Q, et al. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. Stat Sci. 2007;22(4):544–559.

15. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47(260):663–685.

16. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat. 2006; 2(1).

17. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.

18. Lumley T. Complex surveys: a guide to analysis using R. 2010. In Couper MP, Kalton G, Rao JNK, Schwarz N, Skinner C, ed., Wiley Series in Survey Methodology, pp2-3. Wiley and Sons, Inc., Hoboken, NJ.

19. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. Am J Epidemiol. 2011;173(7):731.

20. Simon H. Causal ordering and identifiability. 1953. In W.C. Hood and T. Koopmans, ed., Studies in Econometric Method, vol. 11, pp 65-69. Wiley and Sons, Inc., New York.

21. Porter KE, Gruber S, van der Laan MJ, et al. The relative performance of targeted maximum likelihood estimators. Int J Biostat. 2011;7(1).

22. Petersen ML, Porter KE, Gruber S, et al. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res. 2012;21(1):31-54.

23. Kessler RC, Avenevoli S, Costello EJ, et al. National comorbidity survey replication adolescent supplement (NCS-A): II. Overview and design. J Am Acad Child Adolesc Psychiatry. 2009;48(4):380–385.

24. Kessler RC, Avenevoli S, Green J, et al. National comorbidity survey replication adolescent supplement (NCS-A): III. Concordance of DSM-IV CIDI diagnoses with clinical reassessments. J Am Acad Child Adolesc Psychiatry. 2009;48(4):386–399.

25. Kessler RC, Avenevoli S, Costello EJ, et al. Design and field procedures in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A). Int J Methods Psychiatr Res. 2009;18(2):69–83.

26. Roux A, Kiefe CI, Jacobs Jr DR, et al. Area characteristics and individual-level socioeconomic position indicators in three population-based epidemiologic studies. Ann Epidemiol. 2001;11(6):395–405.

27. Rudolph KE, Wand GS, Stuart EA, et al. The association between cortisol and neighborhood disadvantage in a US population-based sample of adolescents. Health Place. 2014: 25;68-77.

28. McEwen BS. Physiology and neurobiology of stress and adaptation: central role of the brain. Physiol Rev. 2007;87(3):873–904.

29. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat. 2002;84(1):151–161.

30. Messer LC, Oakes JM, Mason S. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. Am J Epidemiol. 2010;171(6):664–673.

31. Wang Y, Petersen ML, Bangsberg D, et al. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. UC Berkeley Working Paper Series. 2006:211.

32. Chaffee P, Hubbard AE, van der Laan ML. Permutation-based pathway testing using the super learner algorithm. UC Berkeley Working Paper Series. 2010:263.

33. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology. 2008; 168(6):656-664.

34. Bembom O, van der Laan ML. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical Report 230, Division of Biostatistics, University of California, Berkeley, 2008.

35. Leon S, Tsiatis AA, Davidian M. Semiparametric estimation of treatment effect in a pretest-posttest study. Biometrics. 2003;59(4):1046–1055.

36. Stitelman OM, De Gruttola V, van der Laan ML. A general implementation of TMLE for longitudinal data applied to causal inference in survival analysis. Int J Biostat. 2012;8(1).

Appendix A

Identification Results

We observe the following vector of data for each participant:

$$O = (\Delta_{svy} = 1, W_1, W_2, A, \Delta_{sub}, \Delta_{sub} \times Y),$$

where Δ_{svy} is an indicator of sample selection that depends on a set of covariates W_1; W_2 is a vector of post-sample covariates; A is a binary treatment or exposure of interest; Δ_{sub} is an indicator of the outcome Y being observed. We only observe O for those with $\Delta_{svy} = 1$; for those not in the survey (i.e., who have $\Delta_{svy} = 0$) we observe no data. Let $\mathbf{W} = (W_1, W_2)$.

We assume each participant's data vector O is an independent, random draw from the unknown distribution P_0 on O. We assume the sampling selection probabilities $P(\Delta_{svy} = 1|W_1)$ are known, and denote them by $\pi(W_1)$.

The objective is to estimate the population average treatment effect (PATE) defined to be

$$\psi^f = E(Y_1 - Y_0),$$

where for each $a \in \{0, 1\}$, Y_a denotes the counterfactual outcome that would be observed if treatment $A = a$ were assigned and $\Delta_{svy}, \Delta_{sub}$ were both set to 1. Under assumptions given below, the PATE can be represented as follows:

$$E(Y_1 - Y_0) = E\left[\frac{\Delta_{svy}}{P(\Delta_{svy}|W_1)}\, E\left(\{\mu(1, \mathbf{W}) - \mu(0, \mathbf{W})\}\middle|\, \Delta_{svy} = 1, W_1\right)\right], \tag{3}$$

where we define $\mu(A, \mathbf{W}) = E(Y|A, \Delta_{sub} = 1, \Delta_{svy} = 1, \mathbf{W})$.

To explain the intuition for (3), first consider the term $\{\mu(1, \mathbf{W}) - \mu(0, \mathbf{W})\}$ on the right side of (3). It is a contrast between the mean outcome distributions under treatment versus control for the sub-sample population, conditioned on the variables $\mathbf{W}$. The inner expectation on the right side of (3) standardizes this contrast to the distribution of $\mathbf{W}$ given W_1 for the survey sample population. Lastly, the outer expectation standardizes this to the distribution of W_1 for the target population based on the survey-sampling weights. This overall process adjusts for measured confounders among variables in $\mathbf{W}$, and standardizes the average treatment effect to the target population. Since W_1 is observed only for individuals in the survey sample, the outer expectation cannot be directly estimated from the observed data. However, we have the following relationships:

$$P(W_1|\Delta_{svy} = 1) = \pi(W_1)P(W_1)/P(\Delta_{svy} = 1),$$

and

$$P(\Delta_{svy} = 1) = 1/E\left[1/\pi(W_1)|\Delta_{svy} = 1\right],$$

where $\pi(W_1) \equiv P(\Delta_{svy} = 1|W_1)$ are known sample selection probabilities. These relationships imply the following alternative representation of the PATE, which can be estimated from the observed data:

$$E(Y_1 - Y_0) = \frac{E(\pi^{-1}(W_1)[\mu(1, \mathbf{W}) - \mu(0, \mathbf{W})]|\Delta_{svy} = 1)}{E(\pi^{-1}(W_1)|\Delta_{svy} = 1)}. \tag{4}$$

For comparison to the PATE in Equation 3, we next define two related quantities. Let the subsample ATE be defined as $E(Y_1 - Y_0|\Delta_{sub} = 1)$. Under the assumptions at the end of this section, it is identified by the equation:

$$E(Y_1 - Y_0|\Delta_{sub} = 1) = E(Y|A = 1, \Delta_{sub} = 1, \Delta_{svy} = 1) - E(Y|A = 0, \Delta_{sub} = 1, \Delta_{svy} = 1).$$

Next, we define the survey sample ATE to be $E(Y_1 - Y_0|\Delta_{svy} = 1)$, which under the same assumptions is identified by the equation:

$$E(Y_1 - Y_0|\Delta_{svy} = 1) = E[\mu(1, \mathbf{W}) - \mu(0, \mathbf{W})|\Delta_{svy} = 1].$$

In general, the above three ATEs differ. The goal of this paper is to estimate the PATE. To highlight the contrast among the above estimands, we examine them in the Case Study Section.

Assumptions

The above identification results rely on the following assumptions:

1. Known survey sampling weights, denoted $\pi^{-1}(W_1)$. That is, for each person in the survey sample, we are given his/her corresponding sampling weight. This is a typical assumption in the survey literature.[3]

2. No unmeasured confounders: for each $a \in \{0, 1\}$, we have

$$Y_a \perp\!\!\!\perp \Delta_{svy}|W_1; \tag{5}$$

$$Y_a \perp\!\!\!\perp (A, \Delta_{sub})|\mathbf{W}, \Delta_{svy} = 1. \tag{6}$$

3. Consistency: for each $a \in \{0, 1\}$, we have $Y_a = Y$ on the event $A = a, \Delta_{sub} = 1, \Delta_{svy} = 1$.

4. Positivity: for each $a \in \{0, 1\}$, we have $P(A = a, \Delta_{sub} = 1, \Delta_{svy} = 1|\mathbf{W})$ and $P(\Delta_{svy} = 1|W_1)$ are strictly positive.

The no unmeasured confounders assumption is analogous to the sequential randomization assumption (SRA) needed when estimating parameters of longitudinal data generating processes [1].
We next prove the relationship (3) holds under the above assumptions. We first have

$$\begin{aligned} E(Y|A = 1, \Delta_{sub} = 1, \Delta_{svy} = 1, W_1, W_2) &= E(Y_1|A = 1, \Delta_{sub} = 1, \Delta_{svy} = 1, W_1, W_2) \\ &= E(Y_1|\Delta_{svy} = 1, W_1, W_2), \end{aligned} \tag{7}$$

where the first equality follows from the consistency assumption, and the second follows from (6). Integrating (7) with respect to the distribution of W_2 conditional on $(W_1, \Delta_{svy} = 1)$, we get

$$E(Y_1|W_1, \Delta_{svy} = 1) = E(E(Y|A = 1, \Delta_{sub} = 1, \Delta_{svy} = 1, W_1, W_2)|W_1, \Delta_{svy} = 1).$$

By (5), the left side of the above display equals $E(Y_1|W_1)$, which implies

$$\begin{aligned} E(Y_1) &= E_{W_1}E(Y_1|W_1) \\ &= E_{W_1}\{E[E(Y|A=1,\Delta_{sub}=1,\Delta_{svy}=1,W_1,W_2)|W_1,\Delta_{svy}=1]\} \\ &= E\left\{\frac{\Delta_{svy}}{P(\Delta_{svy}|W_1)}E[E(Y|A=1,\Delta_{sub}=1,\Delta_{svy}=1,W_1,W_2)|W_1,\Delta_{svy}=1]\right\} \end{aligned}$$

where the positivity assumption is needed to justify the last line. By an analogous argument, we have

$$E(Y_0) = E\left\{\frac{\Delta_{svy}}{P(\Delta_{svy}|W_1)}E[E(Y|A=0,\Delta_{sub}=1,\Delta_{svy}=1,W_1,W_2)|W_1,\Delta_{svy}=1]\right\}.$$

This completes the proof that the above assumptions imply (3).

We next describe the relationship between the no unmeasured confounders assumption and a non-parametric structural equation model (NPSEM). The assumptions (5) and (6) of no unmeasured confounders follow if we assume the following NPSEM:

$$\begin{aligned} W_1 &= f_{W_1}(U_{W_1}) \\ \Delta_{svy} &= f_{\Delta_{svy}}(W_1, U_{\Delta_{svy}}) \\ W_2 &= \Delta_{svy} \times f_{W_2}(W_1, U_{W_2}) \\ A &= \Delta_{svy} \times f_A(W_2, W_1, U_A) \\ \Delta_{sub} &= \Delta_{svy} \times f_{\Delta_{sub}}(A, W_2, W_1, U_{\Delta_{sub}}) \\ Y &= \Delta_{sub} \times \Delta_{svy} \times f_Y(A, W_2, W_1, U_Y), \end{aligned}$$

In the NPSEM, the functions f are assumed unknown but fixed, and the error variables $U_{W_1}, U_{\Delta_{svy}}, U_{W_2}, U_A, U_{\Delta_{sub}}, U_Y$ are assumed to be independent. The variable $U_{\Delta_{svy}}$ represents a set of covariates used in computing the survey selection probabilities. For example, $U_{\Delta_{svy}}$ may represent the size of the county or the household, variables that are often taken into account in multi-stage survey sampling.

TMLE Implementation

The efficient influence function of the parameter ψ^f in the model for O with $\pi(W_1)$ known is given by

$$D(O) = \frac{1}{\pi(W_1)}\left(\frac{(2A-1)\Delta_{sub}}{g(A,\mathbf{W})}(Y-\mu(A,\mathbf{W})) + \mu(1,\mathbf{W}) - \mu(0,\mathbf{W}) - \psi\right),$$

where $g(A,\mathbf{W}) \equiv P(A,\Delta_{sub}=1|\mathbf{W})$. It can be verified that this function is double robust, i.e., it has expectation zero if either μ or g are correctly specified.

We assume that Y is bounded, i.e., that for some values a, b we have $P(Y \in [a,b]) = 1$. We transform Y to be in the interval $[0,1]$ by defining the new variable $Y^* = (Y-a)/(b-a)$. The

TMLE implementation below uses Y^* in place of Y, and at the end the estimate $\hat{\psi}$ is transformed back to the original scale.

An implementation of TMLE for ψ is computed as follows (and is double robust):

1. Obtain initial estimates $\hat{\mu}(A, \mathbf{W})$ and $\hat{g}(A, \mathbf{W})$. In this paper we use logistic regression but the TMLE template can be used with other data-adaptive estimation methods.

2. Define the following logistic regression model:

$$\text{logit}\, \mu(A, \mathbf{W}) = \text{logit}^{-1} \left[\text{logit}\, \{\hat{\mu}(A, \mathbf{W})\} + \beta H(A, \Delta_{sub}, \mathbf{W})\right],$$

   where

$$H(A, \Delta_{sub}, \mathbf{W}) = \frac{(2A-1)\Delta_{sub}}{\pi(W_1)\hat{g}(A, \mathbf{W})}.$$

   The maximum likelihood estimator $\hat{\beta}$ of β is computed by logistic regression of Y on $H(A, \mathbf{W})$ with offset $\text{logit}\, \hat{\mu}(A, \mathbf{W})$, among all participants with $\Delta_{sub} = \Delta_{svy} = 1$. (Although Y may take values in the interval $[0, 1]$, logistic regression software in R still computes a solution $\hat{\beta}$ to the corresponding score equations, which is used in what follows.)

3. Compute

$$\hat{\mu}^1(A, \mathbf{W}) = \text{expit} \left\{\text{logit}\, \hat{\mu}(A, \mathbf{W}) + \hat{\beta} H(A, \Delta_{sub}, \mathbf{W})\right\}$$

4. The TMLE of ψ is defined as

$$\hat{\psi} = \frac{\frac{1}{n}\sum_{i=1}^{n} \pi^{-1}(W_{1i})(\hat{\mu}^1(1, \mathbf{W}_i) - \hat{\mu}^1(0, \mathbf{W}_i))}{\frac{1}{n}\sum_{i=1}^{n} \pi^{-1}(W_{1i})}.$$

In this paper we estimate the variance of the TMLE using the bootstrap.

Appendix B

```
## These functions can be used to replicate the methods used in: Rudolph KE,
   Diaz I, Rosenblum M, Stuart EA. Estimating population treatment effects
   from a survey sub-sample.


# Install tmle Rpackage and load
require(tmle)

#survey weights are assumed known. put the weights in the following object:
data$ssvywt<-data$surveyWeights

#treatment weights
```

```
txwts <- function(data, txmodel){
        sampled.data<-data

        #estimate treatment probabilities from survey sample
        sampled.data$pscore.svy <- predict(glm(formula=txmodel , data=sampled.
            data, family="binomial"), type="response")
        sampled.data$t.wt <- ifelse(sampled.data$t==1, 1/sampled.data$pscore.
            svy, 1/(1-sampled.data$pscore.svy))
        return(sampled.data$t.wt)
}

#selection weights
selwts <- function(data, selmodel){
        sampled.data<-data

        #estimate sub-sample selection probabilities from survey sample
            separately for those treated (A=1) and untreated (A=0)
        #for treated A=1
        sampled.data$pscore.subsamp1[sampled.data$t==1]<-predict(glm(formula=
            selmodel, data=sampled.data[sampled.data$t==1,], family="binomial")
            , type="response")
        sampled.data$subsamp.wt[sampled.data$t==1]<-1/sampled.data$pscore.
            subsamp1[sampled.data$t==1]

        #for untreated A=0
        sampled.data$pscore.subsamp1[sampled.data$t==0]<-predict(glm(formula=
            selmodel, data=sampled.data[sampled.data$t==0,], family="binomial")
            , type="response")
        sampled.data$subsamp.wt[sampled.data$t==0]<-1/sampled.data$pscore.
            subsamp1[sampled.data$t==0]
        return(sampled.data$subsamp.wt)
}

#IPW estimator
ipw <- function(data, txmodel, selmodel){
        sampled.data<-data

        sampled.data$t.wt <- txwts(sampled.data, txmodel)
        sampled.data$subsamp.wt <- selwts(sampled.data, selmodel)

        #combine weights
        sampled.data$subsvytrtwt<-ifelse(sampled.data$insubsample==1, sampled.
            data$t.wt*sampled.data$subsamp.wt*sampled.data$ssvywt, 0)

        #estimate ATE
        lm.subsvytrt <- lm(y ~ t, weights=sampled.data[sampled.data$
            insubsample==1,]$subsvytrtwt, data=sampled.data[sampled.data$
            insubsample==1,])
        ate.ipw <- summary(lm.subsvytrt)$coef[2,1]
```

```
48
49          return(ate.ipw)
50      }
51
52  #DRWLS estimator
53  drwls <- function(data, txmodel, selmodel){
54          sampled.data<-data
55
56          sampled.data$t.wt <- txwts(sampled.data, txmodel)
57          sampled.data$subsamp.wt <- selwts(sampled.data, selmodel)
58
59          #combine weights
60          sampled.data$subsvytrtwt<-ifelse(sampled.data$insubsample==1, sampled.
                data$t.wt*sampled.data$subsamp.wt*sampled.data$ssvywt, 0)
61
62          #specify outcome model for g-computation step
63          model<-glm(y ~ t + z2 + t:z1, data=sampled.data, weights=sampled.data$
                subsvytrtwt, family="gaussian")
64
65          data_new0<-sampled.data
66          data_new0$t<-0
67          data_new1<-sampled.data
68          data_new1$t<-1
69
70          #g-computation step
71          sampled.data$y1hat<-predict(model, newdata=data_new1, type="response")
72          sampled.data$y0hat<-predict(model, newdata=data_new0, type="response")
73          sampled.data$dif<-sampled.data$y1hat - sampled.data$y0hat
74          sampled.data$num<-sampled.data$dif*sampled.data$ssvywt
75
76          ate.drwls <-sum(sampled.data$num)/sum(sampled.data$ssvywt)
77
78          return(ate.drwls)
79      }
80
81  #TMLE estimator
82  tmlefxn<-function(data, txmodel, selmodel, outmodel){
83      sampled.data<-data
84
85      sampled.data$pscore.svy <- predict(glm(formula=txmodel, data=sampled.data,
            family="binomial"), type="response")
86
87      #estimate subsample probabilities from survey sample, separately for those
            treated and untreated
88
89      modelc<-glm(formula=selmodel, data=sampled.data, family="binomial")
90
91      data_new0<-sampled.data
92      data_new0$t<-0
```

```
    data_new1<-sampled.data
    data_new1$t<-1

    sampled.data$a1s1.ps<-predict(modelc, newdata=data_new1, type="response")
    sampled.data$a0s1.ps<-predict(modelc, newdata=data_new0, type="response")

    W<-as.matrix(cbind(W1=sampled.data$z1, W2=sampled.data$z2))

    fit.tmle<-tmle(Y=sampled.data$y, A=sampled.data$t, W=W, Delta=sampled.data
        $insubsample, Qform=outmodel, g1W=sampled.data$pscore.svy, pDelta1=(1/
        sampled.data$ssvywt)*cbind(sampled.data$a0s1.ps, sampled.data$a1s1.ps),
         gbound=c(0,1))

    sampled.data$dif<-fit.tmle$Qstar[,2] - fit.tmle$Qstar[,1]
    sampled.data$num<-sampled.data$dif*sampled.data$ssvywt

    ate.tmle<-sum(sampled.data$num)/sum(sampled.data$ssvywt)
    return(ate.tmle)
}
```

Functions.R

Appendix C

For each of 1,000 simulations, we generate 20 million population members (similar to the population of U.S. adolescents), each with W_1. W_1 is the sum of six normally distributed covariates. We generate an indicator of survey selection from a Bernoulli distribution with probability: $P(\Delta_{svy} = 1|W_1) = Logit^{-1}(-7.5 + 0.07W_1)$. Approximately 12,000 individuals (0.06%) are retained in the survey. For those in the survey, we generate 1) W_2 as the sum of 16 normally distributed covariates, and 2) a treatment variable from a Bernoulli distribution with probability: $P(A = 1|\mathbf{W}) = Logit^{-1}(-2.5 + 0.13W_2)$. Approximately one-half of those in the survey sample are exposed ($A = 1$). We then generate an indicator of sub-sample selection from a Bernoulli distribution with probability: $P(\Delta_{sub} = 1|\Delta_{svy} = 1, \mathbf{W}) = Logit^{-1}(-2.5 + 0.05W_1 + 0.05W_2)$. Approximately 3,000 (25% of survey sample) are retained in the sub-sample. Finally, we generate a continuous outcome variable that exhibits linear effect heterogeneity: $Y = -3 + W_2 + 3A + 2AW_1$, $\epsilon \sim N(0, 2)$. The true PATE is -4.0. R code for generating this simulation scenario is included below.

```
## These functions can be used to replicate the data generating process used
    in: Rudolph KE, Diaz I, Rosenblum M, Stuart EA. Estimating population
    treatment effects from a survey sub-sample.

library(MASS)

expit<-function(p){
```

```
  exp(p)/(1+exp(p))
}

#dataset with n*10 = 20 million individuals.
set.seed(132)
n=2000000

#variance-covariance matrices for the two summary covariates, W1 and W2
sigmaW1<-matrix(data=c
    (100,7,2,4,-3,1,7,20,.3,.6,-.6,.2,2,.3,4,.6,-.3,.1,4,.6,.6,1.7,-.5,.1,-3,-.6,-.3,-.5
    , nrow=6, byrow=TRUE)

sigmaW2<-matrix(data=c( 1.91 , -0.34 , 0.44 , -0.29 , 0.70 , -0.17 , 0.40 ,
    -0.44 , 0.80 , 0.19 , 0.19 , 1.00 , -0.30 , 0.46 , -1.03 , -0.56,  -0.34 ,
     1.89 , 0.99 , 1.06 , 0.78 , -0.17 , -0.22 , 0.26 , 0.37 , -0.10 , -0.91 ,
    -0.92 , -0.02 , -0.29 , 0.45 , 0.16  ,0.44 , 0.99 , 2.26 , 0.96 , 1.83 ,
    -0.30 , 0.78 , 1.28 , 0.15 , 0.04 , -0.64 , -0.20 , 1.21 , -0.78 , 1.34 ,
    -0.63  , -0.29 , 1.06 , 0.96 , 3.26 , 1.38 , 0.57 , 1.18 , 0.94 , -0.48 ,
    -0.04 , -2.16 , 0.10 , -0.72 , -1.78 , 0.35 , -1.06,  0.70 , 0.78 , 1.83 ,
     1.38 , 7.18 , -1.25 , 0.91 , 1.52 , -3.41 , 0.82 , -2.03 , 0.04 , 0.63 ,
    -1.22 , 1.83 , 0.09  ,-0.17 , -0.17 , -0.30 , 0.57 , -1.25 , 1.76 , 0.65 ,
    -0.66 , -0.04 , -0.83 , 0.17 , -0.32 , -1.15 , -1.60 , 0.62 , -0.12, 0.40 ,
     -0.22 , 0.78 , 1.18 , 0.91 , 0.65 , 4.14 , 1.75 , -1.85 , -0.94 , 0.33 ,
    -0.07 , 1.26 , -0.91 , 1.67 , -1.57 ,  -0.44 , 0.26 , 1.28 , 0.94 , 1.52 ,
    -0.66 , 1.75 , 8.23 , -2.65 , 0.84 , 0.12 , 0.32 , 0.33 , -0.45 , 2.53 ,
    -1.19  ,0.80 , 0.37 , 0.15 , -0.48 , -3.41 , -0.04 , -1.85 , -2.65 , 9.62 ,
     0.64 , 0.50 , -0.53 , -1.95 , 1.24 , -1.88 , -2.26 ,  0.19 , -0.10 , 0.04
    , -0.04 , 0.82 , -0.83 , -0.94 , 0.84 , 0.64 , 1.93 , -1.50 , 0.28 , -0.65
    , 1.52 , 0.28 , -0.14  , 0.19 , -0.91 , -0.64 , -2.16 , -2.03 , 0.17 , 0.33
     , 0.12 , 0.50 , -1.50 , 5.23 , -0.24 , 1.50 , 1.07 , -1.94 , 0.58, 1.00 ,
    -0.92 , -0.20 , 0.10 , 0.04 , -0.32 , -0.07 , 0.32 , -0.53 , 0.28 , -0.24 ,
     2.07 , -0.79 , 0.70 , -1.12 , 0.76  , -0.30 , -0.02 , 1.21 , -0.72 , 0.63
    , -1.15 , 1.26 , 0.33 , -1.95 , -0.65 , 1.50 , -0.79 , 5.90 , -0.47 , 1.09
    , -0.88  , 0.46 , -0.29 , -0.78 , -1.78 , -1.22 , -1.60 , -0.91 , -0.45 ,
    1.24 , 1.52 , 1.07 , 0.70 , -0.47 , 5.96 , -2.40 , 1.79 , -1.03 , 0.45 ,
    1.34 , 0.35 , 1.83 , 0.62 , 1.67 , 2.53 , -1.88 , 0.28 , -1.94 , -1.12 ,
    1.09 , -2.40 , 7.18 , 0.31,  -0.56 , 0.16 , -0.63 , -1.06 , 0.09 , -0.12 ,
     -1.57 , -1.19 , -2.26 , -0.14 , 0.58 , 0.76 , -0.88 , 1.79 , 0.31 , 5.53
    ),nrow=16,byrow=TRUE)

#the following loop is for speed, since 20 million is a lot of people
svysamp1<-list(rep(NA,10))
  for (j in 1:10){
    #draw W1 covariate
    w1vars<-mvrnorm(n=n, mu=c(20, -15,-10, 2.5,-1,0), Sigma=sigmaW1)
    #make summary covariate
    z1<-rowSums(w1vars)

    #selection into survey sample
```

```
    beta0 <- -7.5
    beta1 <- log(1.07)
    prob.sel <- expit(beta0+beta1*z1)
    ss<-rbinom(n,1, prob.sel)
    ssvywt<-mean(prob.sel)/prob.sel
    pop<-data.frame(z1, prob.sel, ss, ssvywt, w1vars)
    colnames(pop)<-c("z1","prob.svysel", "insample", "ssvywt", "w1" , "w2", "
        w3", "w4", "w5", "w6")

  svysamp1[[j]]<-pop[ss==1,]
  }

svysamp<-rbind(svysamp1[[1]],svysamp1[[2]],svysamp1[[3]],svysamp1[[4]],
    svysamp1[[5]],svysamp1[[6]],svysamp1[[7]],svysamp1[[8]],svysamp1[[9]],
    svysamp1[[10]])
rm(svysamp1)

#draw W2 covariate
w2vars<-mvrnorm(n=nrow(svysamp), mu=c(0,0,2,0,10 ,5,-5,10,-2,0, 0,0,0,0,0,0),
    Sigma=sigmaW2)
svysamp$z2<-rowSums(w2vars)

# make treatment variable
svysamp$prob.trt<-expit(-2.5 + (log(1.14)*svysamp$z2) )
svysamp$t<-rbinom(nrow(svysamp), 1, svysamp$prob.trt)

# make outcome variable
svysamp$meany0<- -3 + svysamp$z2
svysamp$y0<-rnorm(nrow(svysamp),svysamp$meany0,2)
svysamp$y1<-svysamp$y0 + 3 + 2*svysamp$z1 + rnorm(nrow(svysamp),0,.5)
svysamp$y<-ifelse(svysamp$t==1, svysamp$y1, svysamp$y0)

# selection into subsample to determine which outcome variables are observed
beta0 <- -2.5
beta1 <- log(1.05)
beta2 <- log(1.05)
svysamp$prob.subsel <- expit(beta0+ beta1*svysamp$z1 + beta2*svysamp$z2 )
svysamp$insubsample<-rbinom(nrow(svysamp), 1, svysamp$prob.subsel)
```

DataGenerationCode.R

Appendix D

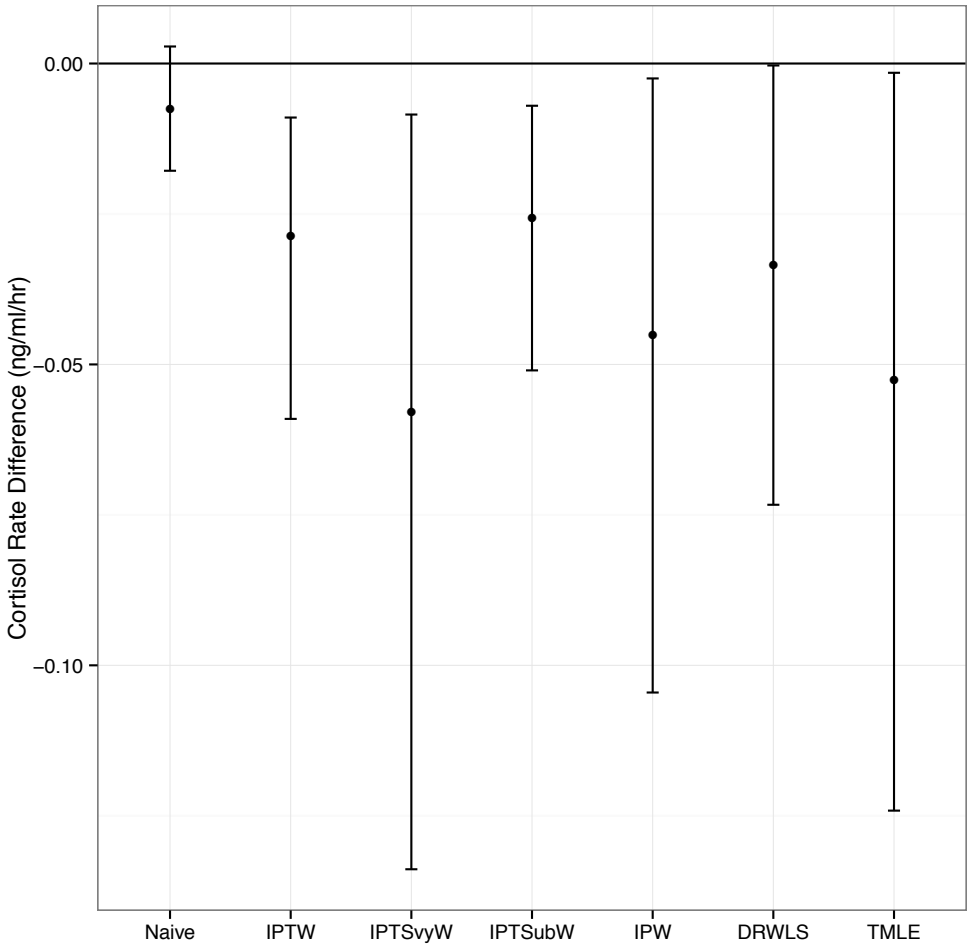
Just as in the simulation study, we estimate each adolescent's treatment weight as the inverse predicted probability of living in the type of neighborhood (disadvantaged or non-disadvantaged) that the adolescent lives in from a logistic regression model with neighborhood disadvantage status as the outcome and a vector of potential confounders as covariates. These potential confounders in-

clude: age, sex, race/ethnicity (Black, White, Hispanic, Other), maternal age at birth of the child, maternal level of education, whether or not the adolescent lived with his/her father for his/her whole life, urbanicity (urban center, urban fringe, non-urban), region of the country (Northeast, South, Midwest, West), season (winter, spring, summer, fall), whether the sample was taken on a weekday or weekend, and time of day that the sample was taken. Inclusion of these variables was supported by both theory[2] and backward stepwise selection. For example, race/ethnicity and measures of socioeconomic status (e.g., maternal age and education) are thought to be associated with (but not caused by) residence in a disadvantaged neighborhood and also associated with cortisol levels. In addition, several strong predictors of cortisol levels (e.g., season, weekday vs. weekend, sampling time) may be differentially distributed by neighborhood disadvantage status and thereby act as confounders. By assuming that these variables are potential confounders and not mediators, we assume they are not affected by residence in a disadvantaged neighborhood. If they were affected, our results may be biased.[4] It could be argued that maternal age at birth of the adolescent and maternal level of education are influenced by neighborhood disadvantage if considered as part of a multi-generational feedback loop. Ultimately, we included these variables to provide some control for family socioeconomic status.

As in the simulation study, we estimate the sub-sample selection weights as inverse predicted probabilities of being included in the sub-sample with cortisol measures, separately for those treated (living in a disadvantaged neighborhood) and untreated (living in a non-disadvantaged neighborhood). We use a logistic regression model with inclusion in the sub-sample as the outcome and a vector of predictive variables as covariates. These predictive variables included: race/ethnicity, father presence, urbanicity, region, season, sampling time (all of which were also included in the treatment model), as well as: current anxiety or depressive disorder, history of psychological abuse, parental employment, family income, weekend bedtime, and small for gestational age at birth. These variables were determined by backward stepwise selection. We assume that selection into the sub-sample conditional on these variables is independent of selection into the sub-sample conditional on these variables and the vector of confounders included in the treatment model. This assumption allows us to multiply the conditional treatment and sub-sample selection probabilities to obtain the joint conditional probability of treatment status and sub-sample selection.

Appendix E

Figure 6: Illustrative example: population average effect estimates and 95% confidence intervals using the NCS-A sub-sample. Eight outlying, influential observations included.

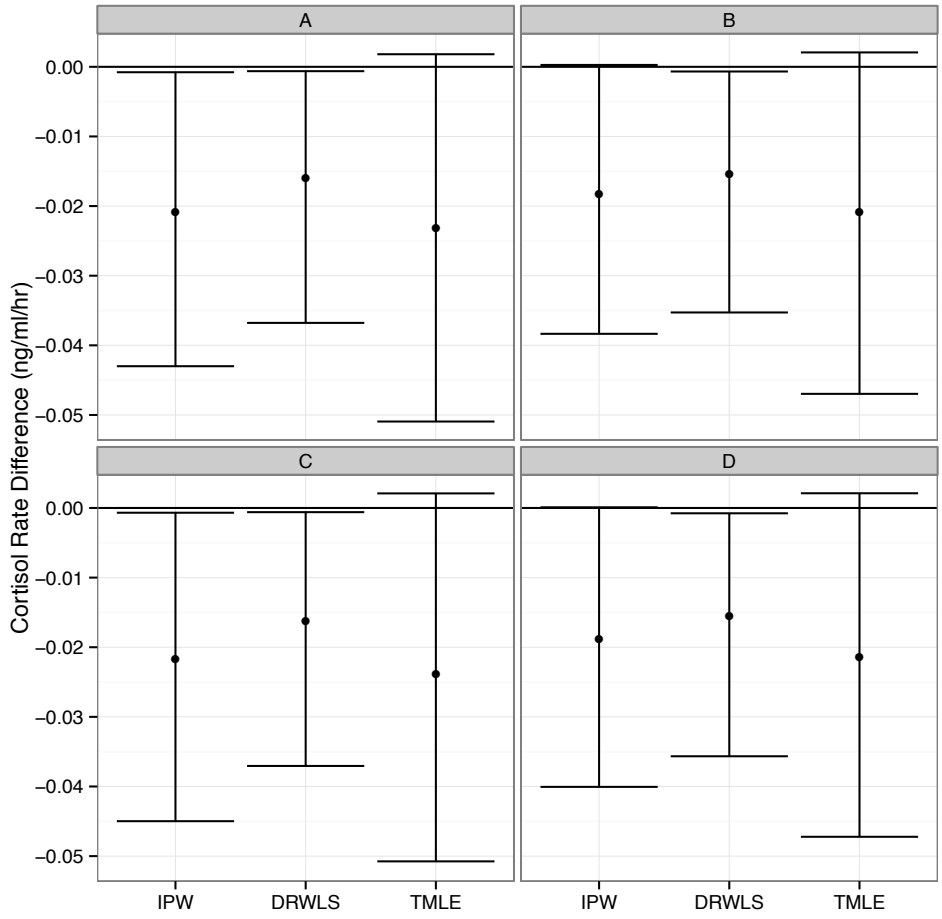


Appendix F

Unlike in the simulation, we do not know the true parametric model forms for the treatment, selection, and outcome models in this case study. We chose the models using Akaike information criterion (AIC) in a stepwise algorithm. Using the AIC may result in models that have better predictive ability but are less parsimonious. In the simulation, we paid no noticeable penalty for over-adjusting or unnecessarily adjusting for non-random treatment and non-random selection (see Table 5). It is possible, though, that we may pay an efficiency penalty in the more complicated real-world case study. To examine this, we compare (A) the IPW, TMLE, and DRWLS estimators that use the

full, AIC-optimized treatment and selection models to (B) those that use the full, AIC-optimized treatment model and a more parsimonious Bayesian information criterion (BIC)-optimized selection model, (C) those that use the BIC-optimized treatment model and the AIC-optimized selection model, and (D) those that use the BIC-optimized selection and treatment models (see Figure 7). When the more parsimonious selection model is used in panels B and D, the CI of each of the estimators slightly narrows but relative performance stays the same. Interval width is insensitive to parsimony in the treatment model.

Figure 7: Illustrative example: marginal mean effect estimates and 95% confidence intervals under different levels of parsimony in model specification. A=full treatment model, full selection model; B=full treatment model, smaller selection model; C=smaller treatment model, full selection model; D=smaller treatment and selection models.



Appendix G

For example, let the full set of covariates be represented by $\mathbf{W} = (W_1, W_2)$. Participants in the sub-sample have the full set of covariates (W_1, W_2), but participants in the survey sample have only the subset (W_1). In this case, the observed data would be O = (W_1, $\Delta_{svy} = 1$, A, Δ_{sub}, $\Delta_{sub} \times W_2$, $\Delta_{sub} \times Y$). In order for the exchangeability assumption to hold, Δ_{svy} and Δ_{sub} must be random conditional on W_1, which implies $P(\Delta_{svy} = 1|W_1) = P(\Delta_{svy} = 1|W_1, W_2)$ and $P(\Delta_{sub} = 1|\Delta_{svy} = 1, W_1) = P(\Delta_{sub} = 1|\Delta_{svy} = 1, W_1, W_2)$. Some additional assumptions and modifications would be required for each of the three estimators evaluated to maintain their unbiasedness properties, as described below.

For the IPW estimator, if W_1 and W_2 are both confounders of the treatment effect, then the conditional probabilities with which we construct the combined inverse probability of treatment and selection weights would need to change to:

$$\begin{aligned} w^{A=a,\Delta_{svy}=1,\Delta_{sub}=1} &= \frac{1}{P(\Delta_{svy}=1|W_1)} \times \frac{I(\Delta_{sub}=1)}{P(\Delta_{sub}=1|\Delta_{svy}=1,W_1)} \\ &\quad \times \frac{I(A=a)}{P(A=a)|\Delta_{svy}=1,\Delta_{sub}=1,W_1,W_2)} \\ &= \frac{I(A=a,\Delta_{sub}=1,\Delta_{svy}=1)}{P(A=a,\Delta_{sub}=1,\Delta_{svy}=1|\mathbf{W})} \end{aligned}$$

for $a \in \{0, 1\}$. The IPW estimate would be unbiased if the treatment and selection models were correctly specified. For the TMLE and DRWLS estimators, if W_1 and W_2 are both confounders of the treatment effect, then these estimators would no longer be robust to misspecification of the selection model. This is because the G-computation step could only standardize to the sub-sample, meaning that the predicted individual treatment effects would need to be weighted by the combined survey weights and sub-sample weights to compute the PATE. TMLE and DRWLS would be consistent under correct model specification and either treatment or outcome model misspecification.

Appendix References

[1] R. Gill and J.M. Robins. Causal inference in complex longitudinal studies: continuous case. *Ann. Stat.*, 29(6), 2001.

[2] Tama Leventhal and Jeanne Brooks-Gunn. The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. *Psychological bulletin*, 126(2):309, 2000.

[3] Thomas Lumley. *Complex surveys: A guide to analysis using R*, volume 565. John Wiley & Sons, 2011.

[4] 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, 1983.