Extended Abstract

Background:

Randomized Control Trials (RCTs) are believed to be the golden standard for assessing causal treatment effects. By randomly assigning individuals to treatment and control groups, RCTs eliminate all possible confounding effects and consequently the difference between the treatment and the control group is attributed as the causal effect of the treatment. The primary strength is that randomization controls for both observed and unobserved relevant factors by producing balance in these factors across treatment groups or levels. This creates a true counterfactual in the control group. Of course, RCTs are not available for many of the causal effects of interest because random assignment is not possible. In recent decades, there has been a flurry of development in the theory and methods of causal validity in observational studies with non-random selection into exposures. In this study, we are comparing several of these methods. Two methods that have seen very large growth in use in the most recent 10 to 15 years are propensity score matching (PSM) and inverse probability weighting (IPW).

Rosenbaum and Rubin (1983) proposed a propensity score method to address the unavailability of RCTs in practical research. The propensity score reflects the probability of an individual receiving treatment given the values of this individual's measured confounding variables. This technique applies a dimension deduction to form a single confounding covariate (i.e., the propensity score) to replace multiple observed confounding variables. Two frequently used propensity score analyses are propensity score matching (PSM) and inverse probability weighting (IPW; also known as inverse probability treatment weighting, IPTW).

In PSM, individuals from the treatment group are matched to individuals in the control group that have the same or similar propensity scores. With a matched sample, researchers can directly compare the differences between the treated and controlled individuals in the outcomes of interests because the confounding is eliminated by the balanced propensity scores across the two groups. PSM works well when the sample size is large, and the number of controlled individuals exceeds that of the treated individuals. The large control group affords a possibility that each treated individual can be paired with someone with similar levels of propensity scores in the control group. However, when the control group is small, some treated individuals may be discarded because of the failure to find a match for them from the control group.

Compared to PSM that creates subgroups with similar propensity scores, IPW instead uses propensity scores to weight sample observations by the inverse probability of being treated or experiencing an exposure. Specifically, IPW creates "copies" of individuals in the treatment or control groups based on the inverse probability weights (1/propensity score for the treatment group and 1/ (1-propensity score) for the control group) to eliminate the association between the confounding variables and treatment conditions. To estimate the treatment effect, one can fit a

linear regression using the treatment condition to predict the outcomes of interests with the weights applied in the modeling.

One question is whether these computationally intensive methods are better at estimating treatment effects than basic statistical control. If not, they may have more practical limitations especially when external validity is also of concern. Many population studies have data collected using complex sample designs that include unequal probabilities of selection, clustered, and stratified selection. Estimation corrections that allow for proper inference to populations include weighted estimation and standard error estimation that accounts for clustering and stratification. However, these estimators have not been included in the literature on PSM and IPW methods for internal validity. These estimation methods are available in traditional modeling frameworks and therefore, a model-based approach to treatment selection using traditional statistical control can be combined with estimation corrections for proper inference to populations. This results in the ability to obtain average treatment effects (ATEs) that are scaled to the population (population ATEs or PATEs).

In this study, we compare the three methods for increasing internal validity in a longitudinal population study with data collected using complex sampling. We compare the ATE estimates to PATE estimates using the traditional modeling approach. We also compare quality of estimation under the three methods using a simulation experiment.

Methods:

Pilot Simulation:

We first demonstrate the equivalence of four average treatment effect (ATE) methods: traditional control (TC), propensity score matching (PSM), inverse-probability weighting (IPW), and traditional control with centering (TCC) methods by simulating one million observations from a model with selection into the treatment/exposure. Selection occurs when a variable that is associated with treatment is also associated with the outcome thereby confounding the relationship between treatment and outcome. This approach provides evidence of equivalence of the ATE methods in expectation or without sampling error confounding.

The generating model with selection is depicted below. Where Y is the outcome, T is a dichotomous treatment variable, and S is the selection variable. The treatment effect is 1 and the selection variable effects on treatment and the outcome are both 0.80. Variables were generated with standard normal distributions where the treatment variable is dichotomized using the mean (0) as a threshold.

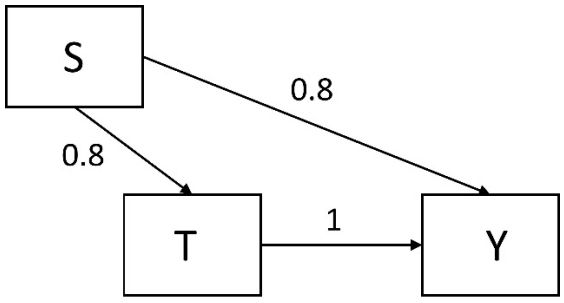


Figure 1. The generating model used to simulation treatment selection.

Real Population Cohort Data Example:

Next, we compare the four ATE methods using an empirical example. Data from a large, longitudinal, nationally representative cohort sample of individuals in the U.S. was used to evaluate the effects of college degree attainment (dichotomous treatment) on mental and physical health in adulthood (ages 24-43). We also assess the effect of years of formal education or schooling completed (continuous treatment) on the health outcomes. Data come from the National Longitudinal Study of Adolescent to Adult Health (Add Health; https://addhealth.cpc.unc.edu/). The sample used in our study includes over 11,000 participants each with data collected across 5 time points over 24 years. Data was first collected in 1994 when participants were adolescents in grades 7 to 12. Subsequent follow up was obtained in 1996, 2001, 2008, and 2017.

Measures

Education treatments and health outcomes were measured in early to middle adulthood. Most participants completed their formal education in early adulthood. Poor physical health is measured using a single self-reported assessment of physical health with higher values representing poorer health. Poor mental health is a depression scale measured as a mean of 5 questions.

Selection variables that were used in the four ATE methods as either model control variables (TC & TCC), matching (PSM), or weighting (IPW) variables included: sex/gender, age at wave 1, white race, learning disability, physical disability, mental health in adolescence, physical health in adolescence, parents' educational degree attainment, family receipt of welfare in adolescence, repeated a grade in school, self-reported intelligence, math grade in high school, and English grade in high school. These variables were related to both the education treatments and both adult health outcomes. Therefore, they represent relevant selection variables.

The unconditional treatment effects of college degree attainment and number of years of schooling are first estimated to show the degree of bias in the sample average treatment effect

(SATE) due to selection. The unconditional model is provided in equation 1. The traditional control (TC) model is estimated with all selection variables included as controls (Equation 2). Equation 2 is also used for the TCC method where model-based means are estimated for the treatment and control groups (college degree or no college degree) and a for a one year increase in years of schooling using centering at the sample mean values from the combined sample. The difference in these models predicted means represent the SATE. These means also represent the potential outcomes for the treatment and control group and treatment levels, respectively. The TCC method is also used in models that allow for moderation of the treatment groups/levels by selection variables. This can be accomplished by estimating Equation 2 stratified by group or by including interactions of the treatment variable with the selection variables.

$$\mathrm{Y} = \beta_1 \mathrm{T} + error \qquad (1)$$

$$\mathrm{Y} = \beta_1 \mathrm{T} + \sum_{p=2}^{P} \beta_p S_p + error \qquad (2)$$

where Y is poor mental or physical health in adulthood, T is college degree attainment or years of schooling in young adulthood and S are the selection variables. β_1 represents the SATE. Stata/SE v. 18 was used to estimate these models.

The PSM and IPW methods are estimated using Stata/SE v18 and the *teffect* command with default settings. These methods were only available for the categorical treatment, college degree attainment. The same S selection variables were included for matching across treatment/control groups in the case of PSM and for weighting across treatment/control groups in the case of IPW. R statistical software was used to estimate the PSM and IPW SATEs for the continuous treatment, years of schooling.

The Add Health Sample was collected using a complex sample design where schools were the primary sampling unit, stratification was used, and unequal selection of observations was present both by design and due to non-participation, etc. Therefore, the SATEs of education on health may not be equal to the population average treatment effects (PATEs). Unequal sample selection would be especially important to the point estimate. The unconditional model, TC, and TCC methods were used again accounting for the complex sample design using sample-weighted estimators and linearized variance estimators that correct for weighting, nesting, and stratification. For the TCC method, model predicted means are based on centering at the population (weighted) means of the selection variables rather than the sample means. As far as we know, little work has been done with the PSM or IPW methods to incorporate sampling corrections to obtain externally valid PATEs and these corrections were not available in the software we used. Stata v18 *svy* commands were used to correct estimates for the sampling design when estimating the PATEs.

Simulation Experiment:
We evaluated the bias, variance (SEs), and 95% confidence interval coverage in estimates across the TC, PSM, and IPW methods. These are evaluated as a function of several factors, including samples size, degree of selection, and number of selection variables. We generated data for a 3X4X3X3 design (3 methods X 4 sample sizes X 3 degrees of selection X 3 number of selection variables). Each cell has 1,000 replicate samples (N = 108,000).

Select Pilot Results:
The biased treatment effect from the simple simulation was 1.8, which should be compared to the true effect of 1. The PSM method matching on S resulted in a treatment effect of 0.999. The IPW method weighting on S resulted in a treatment effect of 1.027. The TC and TCC methods controlling on S resulted in a treatment effect of 0.999. All methods recovered the true treatment effect.

Tables 1 and 2 provide initial results from the empirical analysis for SATEs for the physical health outcome only. There was about a 67% and 71% bias in the uncorrected estimate for the effect of college degree and years of education, respectively, due to sample selection into education. All the methods recovered similar SATEs. Note that the TC and TCC methods provide identical results and therefore we presented the TCC methods with moderation in the tables. There does not seem to be a large difference in the estimate after allowing treatment effects to differ across levels of the selection variables for these examples. Notice the R-package for PSM in the continuous treatment case had some precision issues and some extreme weights emerged in the IPW method for continuous treatment. These extreme weights were trimmed. In general, the PSM and IPW methods are less well developed in the software for continuous treatment effects.

Tables 3 and 4 display the PATEs for the uncorrected, the TC, and TCC methods. In this case, the PATEs are not that different from the SATEs. It may be the case that the same variables that are controlled to account for selection into education are also accounting to some degree for selection into the sample. For example, the Add Health study did have unequal selection across race and disability. This would make the model robust to both sample selection and treatment selection. In other situations, the PATE could be quite different from the SATE. Note that standard error estimates for the PATEs are also corrected for the nesting and therefore are less biased estimates of the precision of the treatment effect estimates.

Select Simulation Results:
All methods resulted in unbiased ATEs in expectation, but the TC method was the only method with bias that was not detectable from zero (Figure 1). The methods were similarly efficient as measured by the standard deviation of the sample estimates, again with the TC method performing marginally better (Figure 2). Confidence interval coverage (95% CI) was best in the TC method (0.948), next best in the PSM method (0.933), and least good in the IPW method (0.924).

The TC method outperformed PSM and IPW in terms of bias in smaller sample sizes, with higher degrees of selection, and with more selection variables. The TC method was more efficient (smaller variance of sample estimates) compared to the PSM method for all levels of selection and all numbers of selection variables. The IPW method fell somewhere between the other two methods with respect to efficiency and was more impacted by the degree of selection and number of selection variables than the other two methods.

Summary:
Our initial findings suggest that using traditional modeling methods may be the most practical and flexible way to improve causal validity, particularly when estimating population level effects. Population researchers and especially life-course researchers using large, representative samples with data collected longitudinally are at an advantage for incorporating the most relevant selection variables. None of the ATE methods for observational studies can account for unmeasured selection. This is the advantage of randomization. Nevertheless, combining the TC or TCC method with within-person models that control for unmeasured between-person factors would result in strong causal validity. Combining this with methods for improved external inference allows for good estimates of treatment (or exposure) effects in populations.

Table 1. The effect of having earned a college degree (treatment) on self-reported poor physical health in adulthood (n = 13,989; n_control = 8,558; n_treatment = 5,431) using four causal validity methods

	effect	SE	t	p-value	95% CI	
uncorrected	-0.434	0.014	-30.35	0.000	-0.463	-0.406
traditional control (TC)	-0.260	0.016	-16.21	0.000	-0.291	-0.228
propensity score matching (PSM)	-0.263	0.026	-10.06	0.000	-0.314	-0.212
inverse-probability weighting (IPW)	-0.262	0.028	-9.36	0.000	-0.317	-0.207
centering with moderation (TCC)	-0.248	0.017	-14.92	0.000	-0.281	-0.216

Table 2. The effect of years of school completed (continuous treatment) on self-reported poor physical health in adulthood (n = 11,495) using four causal validity methods

	effect	SE	t	p-value	95% CI	
uncorrected	-0.097	0.004	-24.65	0.000	-0.105	-0.090
traditional control (TC)	-0.057	0.005	-12.15	0.000	-0.066	-0.048
propensity score matching* (PSM)	-0.061	0.027	-2.26	0.024		
inverse-probability weighting* (IPW)	-0.058	0.007	-8.85	0.000		
centering with moderation (TCC)	-0.057	0.005	-11.77	0.000	-0.067	-0.048

*R software was used

*Table 3. The **population** effect of having earned a college degree (treatment) on self-reported poor physical health in adulthood (n = 13,989; n_control = 8,558; n_treatment = 5,431) using four causal validity methods*

	effect	SE	t	p-value	95% CI	
unconditional	-0.466	0.020	-22.81	0.000	-0.507	-0.426
control (TC)	-0.275	0.024	-11.64	0.000	-0.321	-0.228
centering no moderation (TCC)	-0.275	0.024	-11.64	0.000	-0.321	-0.228
centering with moderation (TCC)	-0.264	0.028	-9.49	0.000	-0.319	-0.209

*Table 4. The **population** effect of years of school completed (dosage treatment) on self-reported poor physical health in adulthood (n = 11,495) using four causal validity methods*

	effect	SE	t	p-value	95% CI	
unconditional	-0.101	0.006	-17.17	0.000	-0.113	-0.089
control (TC)	-0.055	0.007	-7.89	0.000	-0.068	-0.041
centering no moderation (TCC)	-0.055	0.007	-7.89	0.000	-0.068	-0.041
centering with moderation (TCC)	-0.054	0.008	-6.87	0.000	-0.070	-0.039

Figure 1:

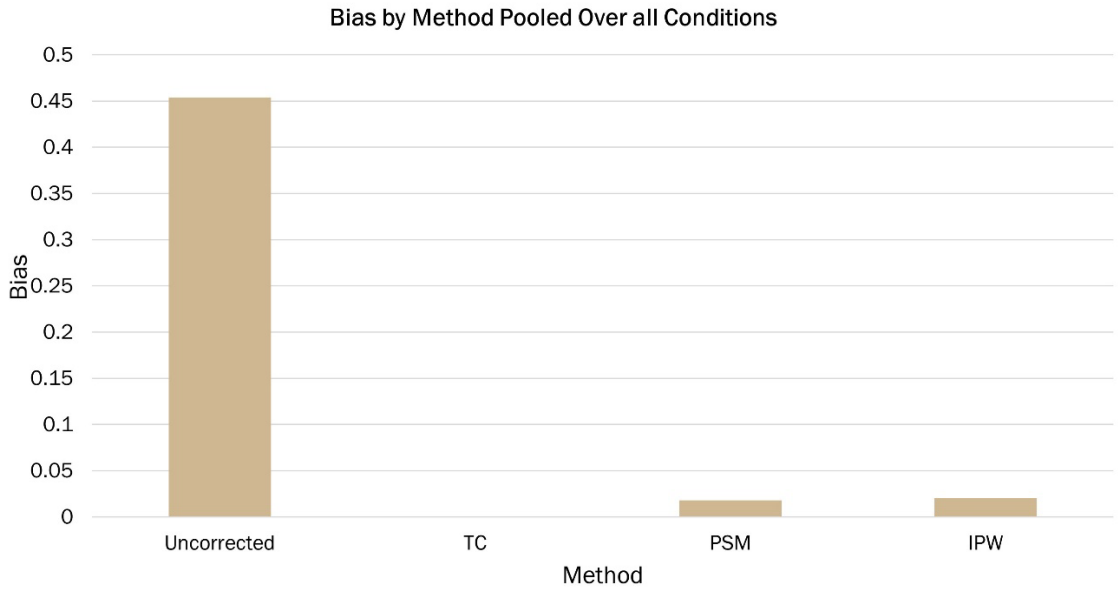
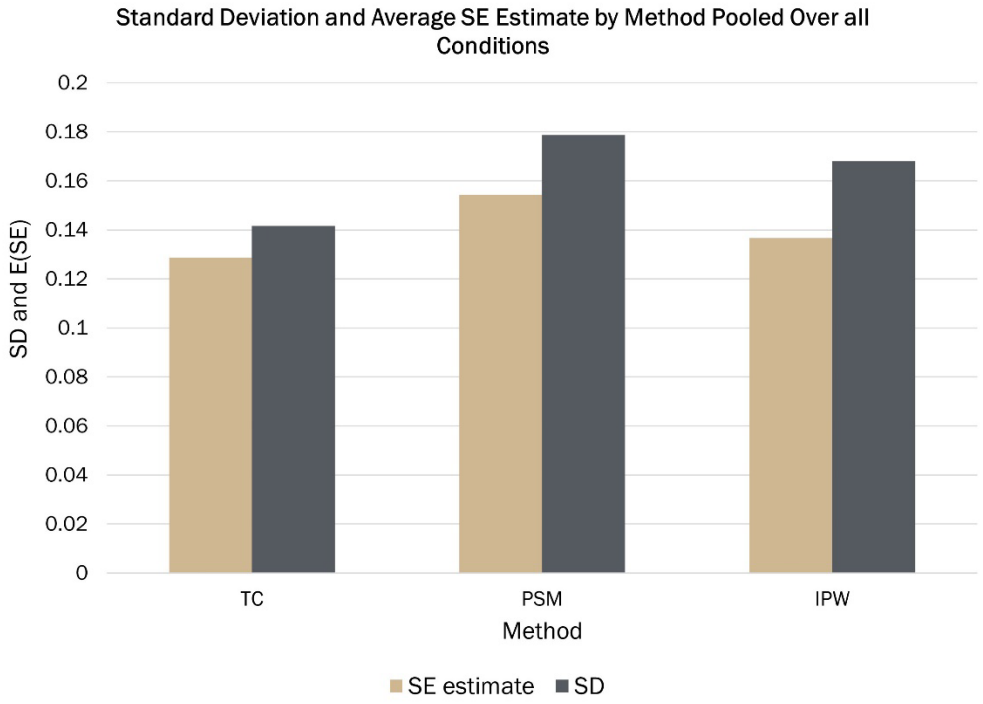


Figure 2:


Standard Deviation and Average SE Estimate by Method Pooled Over all Conditions
SD and E(SE)
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
TC
PSM
IPW
Method
SE estimate
SD