Causal Project

Milena Silva, Aidan Neher, Tiankai Xie

November 28, 2023

Introduction

For researchers and policymakers looking to see whether a policy or treatment is worthwhile to implement, a common problem is that the research conducted on the issue might have taken place within a setting that is substantively different from the setting in which they seek to apply the policy or treatment. In certain situations, generalizability or transportability methods can serve as a solution, and allow us to estimate the average treatment affect within the target population (TATE).

In order to do this, however, several stipulations must hold. Firstly, we must have individual participant data for a well-defined target population of interest (or a sample thereof) and for a study/trial where the intervention of interest was studied. The trial data must include the treatment assignment, the observed outcome, and a set of covariates that are meaningfully linked to the outcome and/or the treatment difference. The target data does not need to include the assignment or outcome (if it did, there would be no need to apply transportability methods), but it does need to share covariates with the trial. Secondly, certain identifiability conditions, such as positivity, consistency, and conditional exchangeability of mean must apply.

A potential problem for researchers looking to apply these methods, however, is that there might be covariates that hypothetically signify the same thing, and therefore could potentially be utilized in this type of analysis, but that are defined in ways that differ between the study and trial population. In this case, the concern would be that blind utilization of these covariates could create bias in our estimation of the TATE.

We chose to focus on a specific subtype of this issue: namely, the case in which there is an underlying continuous variable, U, that is reported as a binary variable depending on whether the

realized value of U falls above or below a certain cutpoint. In this situation, differential coding occurs if there were different cutpoints used in the trial and target datasets. Our goal for this paper is to run a simulation study to see what the impact of increasing the severity of differential coding will be in practice.

Methods

Data Simulation Description

In transportability, we have two datasets: a sample from a trial or study, and a sample from a target population of interest. Transportability methods assume that we have a common set of covariates, X, in both of our datasets. We sought to include common covariates into our X; sex, age, and race were incorporated because these specific variables are commonly used, while BMI was incorporated because continuous variables are common in many settings. Our dataset also includes our latent variable U, a differentially coded binary variable V based on the latent variable U, treatment and study indicators (A and S), potential outcomes Y^1 and Y^0, and the observed outcome Y.

Our variable for sex was modeled as a Bernoulli random variable, using proportions from two NHANES datasets (2018-2020) and (2011-2018) so that we could have realistically different covariate balances in the trial and target populations. Age was generated through beta distributions, adjusted to fit within predefined age ranges, mirroring a realistic age distribution in the target population in NHANES data (2018-2020) [3] and (2011-2018) [2]. Race/ethnicity is categorized into four distinct groups, Non-Hispanic White, Non-Hispanic Black, Hispanic, and Other. The probabilities for each of these groups was derived from a specific study related to obesity from NHANES data (2018-2020) for the trial population, and the probabilities for target population is derived from NHANES data (2011-2018) which is a study about cardiovascular disease. One-hot encoding applied to race/ethnicity for analytical purposes. BMI was calculated using age and gender slopes sourced from NHANES data, with an acknowledgment of gender-related differences in BMI average.

We also simulated the latent variables, U, which is a continuous latent variable that generated with varying distributions between the trial and target population. We want to use U to represent a single underlying continuous variable which might be slightly different distributed in the study

and target populations. U can be introduced with two parts, U_{study} and U_{target}. Both of them are considered i.i.d. normally distributed around their respective means with a variance of 1. The difference between their means is given by the 'Udiff' parameter, determined using a uniform distribution between -1 and 1.

The latent variable U is dichotomized into a new variable V based on 2 cut-points determined by the 'k' parameter, which represents the severity of the difference in variable specification. These cut-points differ between the trial and target populations, with an even chance of being assigned the higher cut point.

We also created treatment and study indicators, which used to mark subjects under treatment and differentiate between trial and target populations. Treatment indicator A is simulated from Bernoulli draws with probability equal to 1/2, and study indicator S is fixed to 1 for subjects in the study dataset and 0 for subjects in the target dataset.

Finally, the outcome variable, Y is created. It signifies the outcome of interest and is conditional on the treatment effect. It's formulated based on the potential outcomes ($Y0$ and $Y1$) with Y being observed only for subjects in the trial population. $Y0$ and $Y1$ represent potential outcomes, capturing the predicted values under control and treatment conditions, respectively. Factors influencing those outcomes include gender, age, BMI, race, the latent variable effect (U), and random variation. To display the treatment effect between trial and target group, we simulate a treatment effect (ATE). It represents the impact of treatment on potential outcomes, calculated by scaling the latent variable (U) with a constant, which we set as 3. $Y1$ incorporates the treatment effect ('ATE') on top of the factors influencing $Y0$.

Models

The factors that we intend to vary across our simulations are i) the method for transportability and ii) the severity of the difference in variable specification, k. Here, the estimators we are using are the ones defined in Dahabreh et al's 2020 paper; specifically we use the outcome regression estimator[1], two IOW estimators[2], and the first two doubly robust estimators described on pages[3]. [1]

[1]Dahabreh, 2020, pg 5.
[2]Dahabreh, 2020, pg 6.
[3]Dahabreh, 2020, pgs 6-7.

We varied K over the following set of values: 0, 0.05, 0.2, 0.5 and 0.8. For each of the models described above, we fit our models for $P(S = 1|X)$, $P(A = a|X, S = 1)$ (for our weights), and $E[Y|X, S = 1, A = a]$ on the full set of covariates and without including any interaction terms; we want to ensure that any changes we observe in bias or coverage are a result of changes in K or in the estimator, and not the results of variable selection, so we went for a simple, consistent model.

We conducted 100 simulations and bootstrapped all simulation for 200 times. For each Monte Carlo simulation, we generated a dataset, and then, for each method of interest, we collected the point estimate $\hat{TATE}$ and corresponding confidence intervals. To compare the results of the simulation, the metrics considered are Bias* and CI coverage. The Bias* is the bias between $\hat{TATE}$ and $TATE^*$, where $TATE^*$ is calculated from a given dataset as $1/n \sum(Y_i^1 - Y_i^0 * I[S = 0])$. We compare the bias* between each k setting, and define bias* closer to 0 as better performance. The CI coverage represents the proportion of Monte Carlo datasets where the confidence interval covers the TATE* or not. CI coverage closer to 1 is considered better performance.

Results

We conducted simulations under five different K, which are K = 0, 0.05, 0.2, 0.5 and 0.8. The estimator we are interested in are treatment affect within the target population (TATE), bias and absolute bias between true treatment affect (Truth) and TATE (Bias and Abs Bias), and coverage and length of confidence interval (CI Coverage and CI length). All average of those estimated values were calculated from 100 simulations within which we bootstrapped 200 times to find the confidence intervals. Related results are stored in Table 1 and Table2.

Table 1: This table shows the exact, true TATE ("Truth") at each level of K as well as the estimated TATE, bias, and absolute bias for each estimator.

		Outcome Regression			IOW			IOW2		
	Truth	TATE	Bias	Abs Bias	TATE	Bias	Abs Bias	TATE	Bias	Abs Bias
K=0	30.28	29.50	-0.78	2.21	29.40	-0.88	2.94	29.62	-0.66	2.43
K=0.05	29.43	30.17	0.74	2.79	30.16	0.73	3.48	30.30	0.87	2.86
K=0.2	30.23	29.72	-0.51	2.93	30.46	0.23	4.03	29.47	-0.76	3.30
K=0.5	29.73	31.09	1.37	4.61	31.00	1.28	6.85	31.19	1.46	4.90
K=0.8	29.80	32.24	2.44	7.71	33.81	4.01	12.00	33.08	3.28	8.84
		Doubly Robust			Doubly Robust 2					
	Truth	TATE	Bias	Abs Bias	TATE	Bias	Abs			
K=0	30.28	29.68	-0.61	2.28	29.67	-0.61				
K=0.05	29.43	30.30	0.87	2.77	30.30	0.87				
K=0.2	30.23	29.50	-0.73	3.22	29.50	-0.73	3.23			
K=0.5	29.73	31.24	1.51	4.85	31.24	1.52	4.83			
K=0.8	29.80	33.08	3.27	8.82	33.05	3.24	8.73			

Table 2: This table shows the exact coverage and length of confidence interval at each level of K for each estimator.

	Outcome Regression		IOW		IOW2	
	CI Coverage	CI Length	CI Coverage	CI Length	CI Coverage	CI Length
K–0	0.96	10.14	0.98	15.20	0.94	10.73
K–0.05	0.84	10.05	0.96	15.33	0.82	11.16
K–0.2	0.86	10.26	0.90	16.39	0.84	11.26
K–0.5	0.72	13.20	0.76	25.40	0.74	15.36
K–0.8	0.78	24.60	0.96	61.40	0.80	30.19
	Doubly Robust		Doubly Robust 2			
	CI Coverage	CI Length	CI Coverage	CI Length		
K–0	0.94	10.57	0.94	10.55		
K–0.05	0.80	10.98	0.80	10.95		
K–0.2	0.86	11.19	0.86	11.17		
K–0.5	0.78	15.26	0.76	15.09		
K–0.8	0.76	30.65	0.76	29.74		

For our analysis, we were primarily interested in seeing how the bias and coverage of our estimators were affected by increases in K, our term that specifies the amount of differential specification. We initially hypothesized that the average absolute bias might increase and coverage might worsen as K increased, but that some estimators under specific model might perform better than others.

The scenario where K=0 is especially important since that is the case where there is no misspecification from V (although there might still be differences in the distributions in U between the target and the trial from random chance). This is effectively our 'base case' to which we are comparing all the other cases.

In our graph for average absolute bias (Figure 1), we can see that we have relatively low bias at K=0 (our base case), and that it increases substantially as K does. Our outcome regression model (OR) is consistently the best and our non-normalized inverse odds weighting model (IOW) is consistently the worst at each K for average absolute bias.

From examining the the average actual bias for each estimator and each level of k, we can get a sense for the direction of the bias in each situation (Figure 2). Interestingly, for all estimators as a group, this appears to change directions as k increases. But it seems that when k exceeds a certain value, the change of direction may stop. To test this hypothesis it may be necessary to choose more k values from 0.2 to 0.8.

For coverage, the results are a bit more muddled, but coverage seems to generally decrease and worsen with K (Figure 3). At k=0.5, the coverage is the lowest in almost all cases, except for the doubly robust estimator 1 (DRE1), which has the lowest coverage at k=0.8. The non-normalized inverse odds weighting estimator (IOW) has the best coverage for high values of K,

Figure 1: This shows the average absolute bias across each of our M simulations for each estimator and level of k. Here, we are defining bias as the difference between an estimation of the TATE and the truth TATE, which we defined as the average difference between our potential outcomes in our target population.

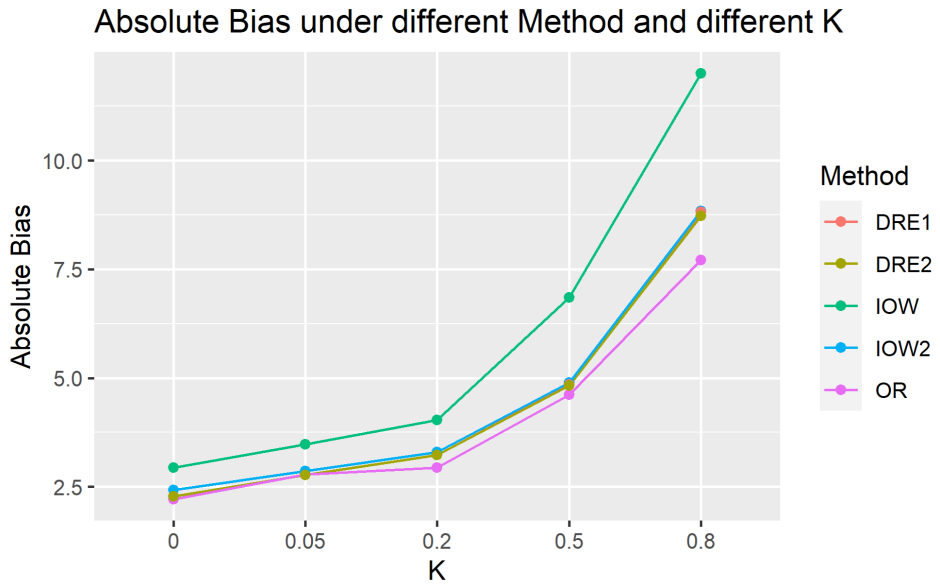
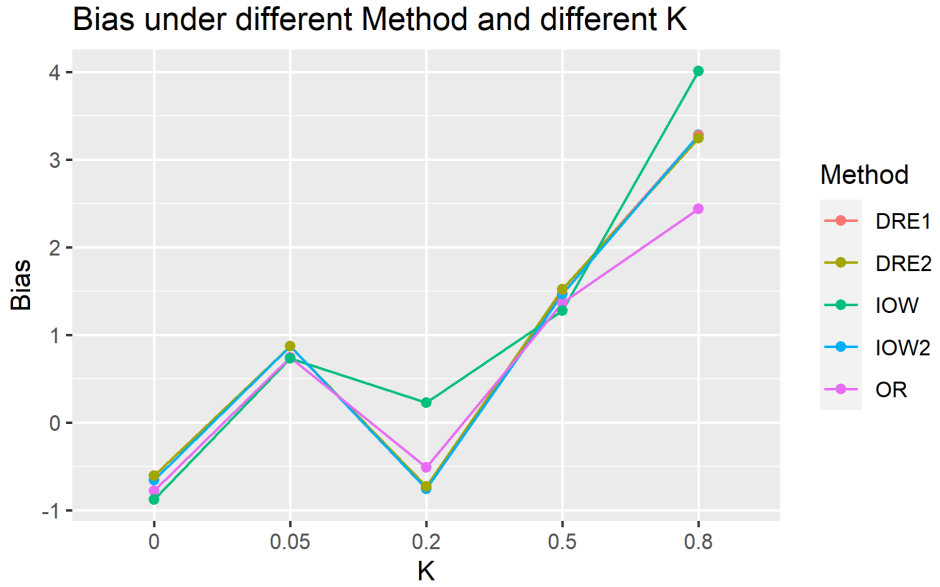
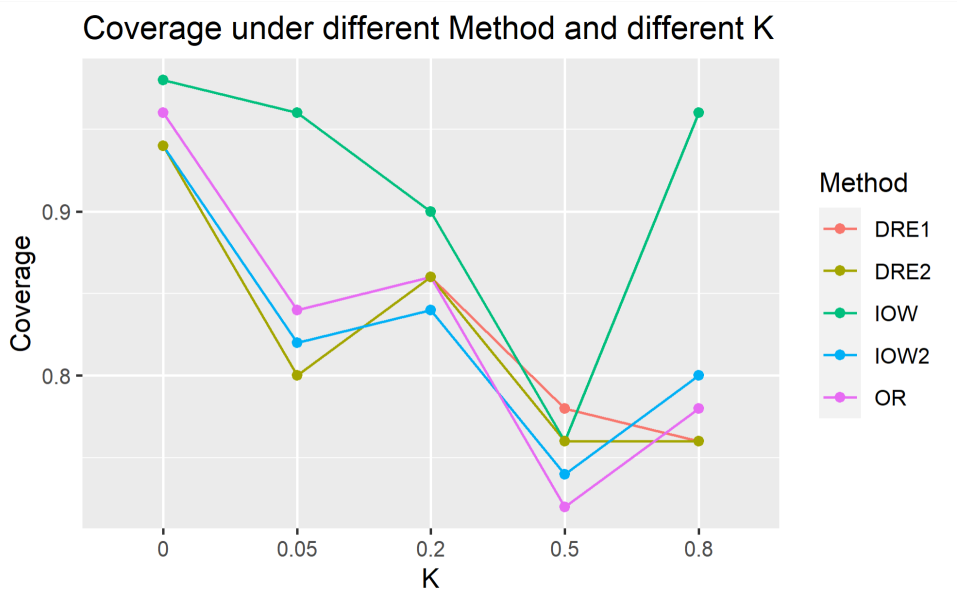


Figure 2: This shows the average actual bias across each of our M simulations for each estimator and level of K.



although not for lower values. The coverage for the doubly robust estimators (DRE1 and DRE2), outcome regression estimator (OR), and the normalized inverse odds weighting model (IOW2) follow a pattern of generally decreasing coverage. The outcome regression estimator (OR) starts

Figure 3: This shows the coverage of bootstrap confidence intervals across each of our M simulations for each estimator and level of k.



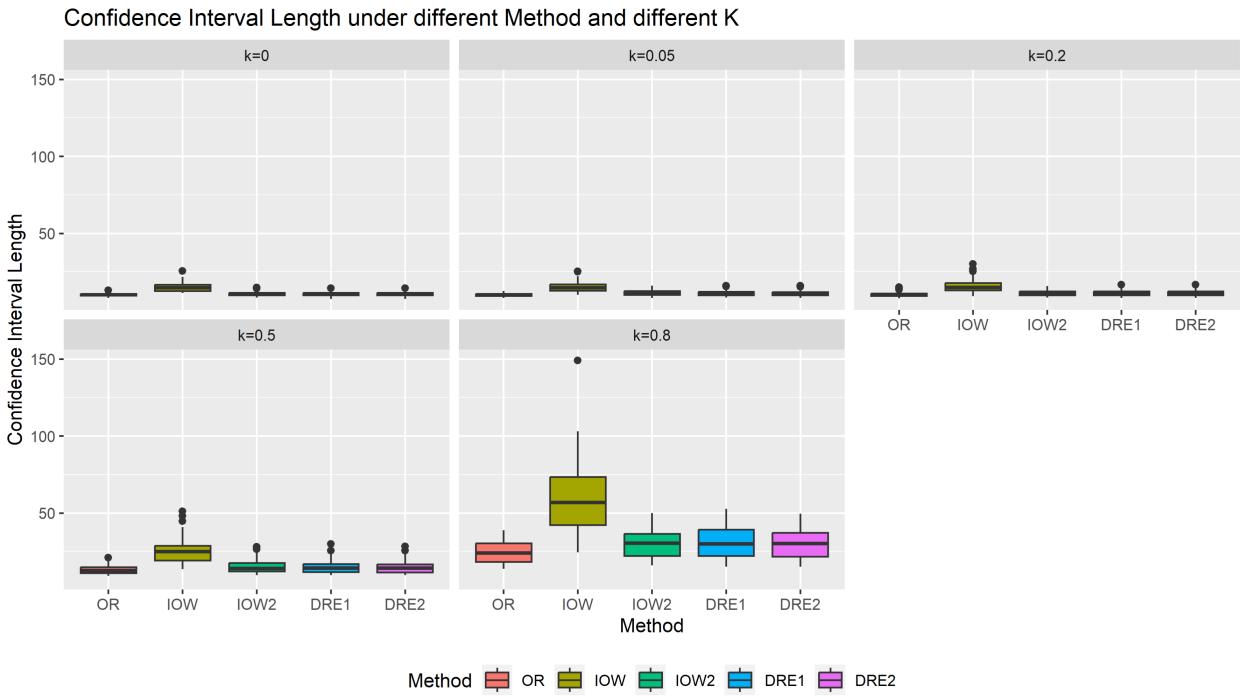
with relatively high coverage, but falls further away from .95 than the other estimators do for higher values of K.

Discussion

Trial participants can differ substantively from the population that policymakers or researchers are targeting with a specific intervention. Generalizability or transportability methods can be used in this context to estimate the average treatment affect within the target population (TATE). This simulation study examined the performance of 5 generalizability or transportability methods from Dahabreh et al. 2020 when specific latent continuous variables U_trial and U_target are differentially coded as "low" or "high" between the trial and the target samples. K in this study represents the severity of this differential specification, with K=0 representing the base case where the cutpoint for "low" or "high" specification is the same between trial and target. Larger values representing a greater degree of disagreement across the trial and target coding.

Primarily, we analyzed performance by average absolute bias (Figure 1), average bias (Figure 2), and CI coverage relative to TATE* (Figure 3). Secondarily, we analyzed bootstrap CI length across simulations (Figure 4). IOW1 had the worst average absolute bias relative to all other methods considered at all values of K. The other 4 methods did not show much difference in

Figure 4: This shows the length of bootstrap confidence intervals across each of our M simulations for each estimator and level of k.



average absolute bias. All methods increased in average absolute bias as K increased, as expected. IOW1 has the largest average bias at K=0.8. Interestingly, all methods have the same average bias direction across K. IOW1, despite performing the worst in terms of average absolute bias across K and average bias at K=0.8, had the best coverage across K. This is likely attributable to its largest confidence interval lengths across K and relative to the other methods.

The finding that IOW1 has the greatest variability in its TATE estimation yet the best coverage across K might be attributed to large variability in its weights. In fact, IOW2 mitigates this by normalizing by the sum of the weights, which results in better finite-sample performance when the weights are highly variable (Dahabreh et. al. 2020). Whether or not the variability in IOW1's weights are driven by our simulation study's design, or the degree of differential specification K is worth exploration. The relative performance of other methods OR, IOW2, DRE1, and DRE2 is not easily determined given their proximity in the metrics of interest. This suggests that it might not matter which of these methods is chosen in the context of target and trial covariate differential specification, or that further analysis is required to determine substantial differences between these methods.

Conclusion

The high variability in IOW1 estimates relative to the other four methods considered points at the issue that can come with high variability in the weights it uses for inference. We think that IOW2 adjusts for this variability, consequently performing similar to the OR, DRE1, and DRE2 across metrics considered. Our simulation study suggests that IOW1 might be used with caution and that the other 4 methods perform similarly in the context of target and trial covariate differential specification. However, additional analyses are required to establish this firmly.

References

[1] Issa J Dahabreh et al. "Extending inferences from a randomized trial to a new target population". In: *Statistics in medicine* 39.14 (2020), pp. 1999–2014.

[2] Jiang He et al. "Trends in cardiovascular risk factors in US adults by race and ethnicity and socioeconomic status, 1999-2018". In: *Jama* 326.13 (2021), pp. 1286–1298.

[3] Jason H Karnes et al. "Racial, ethnic, and gender differences in obesity and body fat distribution: An All of Us Research Program demonstration project". In: *PloS one* 16.8 (2021), e0255583.