Journal of Econometrics and Statistics
Vol. 4, Issue 2, 2024, pp. 217-229

Exploration of Propensity Score Adjustment in Logistic Regression via Simulation Study

Erick Nguyen*[a] and Andrew G Chapple[b]

[a]*Biostatistics Program, School of Public Health, LSU Health Sciences Center, New Orleans, LA, US*
[b]*Department of Interdisciplinary Oncology, LSU School of Medicine, LSU Health Sciences Center , New Orleans, LA, US*

Using propensity scores as covariates can control the effect of confounders in observational studies. However, the methods of variable selection for propensity score modeling are still under debate. To gain insight on the variables that should be used in the propensity model, a simulation study with randomly generated scenarios was conducted to examine confounding variables with varying effect sizes on exposure and outcome. We found that there was a negative effect for including variables related to exposure but not outcome (aka instrumental variables). The inclusion of variables related to outcome, but not related to exposure has little to no detrimental effect on the propensity model. All other relationships did not have an appreciable negative effect either. However, including variables related to both exposure and outcome is necessary for a strong propensity model. Therefore, we recommend including all possible confounders except instrumental variables into the propensity model. In terms of hypothesis testing, we recommend inclusion of all possible confounders to avoid inflation of type I error rates.

keywords: Observational Studies, Propensity Scores, Logistic Regression

1 Introduction

Randomized control trials are considered the gold-standard of establishing a causal relationship between a treatment and an outcome (Hariton and Locascio, 2018). The randomized allocations of participants in the treatment and control groups allows for even distribution of known and unknown confounders to be distributed across each group. Since the distribution of the confounders between the groups are assumed to be equal, the differences in outcomes between the groups can be attributed to the differences in treatment, not the differences resulting from the confounders. Not all studies can be conducted through randomized control trials due to costs or ethical concerns. In such cases, an observational study will be carried out instead. In observational studies, subjects are not randomly allocated into treatment groups and, therefore, do not undergo any interventions (Gilmartin-Thomas et al., 2018). They are instead placed into the treatment groups based on their current exposures. Due to the lack of randomization, there may be systematic differences between the groups, which may obfuscate the true effect of the treatment on the outcome. To accommodate for this, propensity scores, formalized by Rosenbaum and Rubin (1983), are used in observational studies to control for the effect of confounding.

Propensity score methods allow observational studies to mimic some of the characteristics of the randomized control trials. The propensity score was defined to be the probability of exposure (treatment) given observed factors (Rosenbaum and Rubin, 1983). In more detail, patient factors are used to estimate the probability

*Corresponding authors: mahashweta259@gmail.com

Article History
Received : 14 October 2024; Revised : 16 November 2024; Accepted : 30 November 2024; Published : 28 December 2024

To cite this paper

Erick Nguyen & Andrew G. Chapple (2024). Exploration of Propensity Score Adjustment in Logistic Regression via Simulation Study. *Journal of Econometrics and Statistics.* 4(2), 217-229

of exposure, which is then used in adjustment. Individuals with similar probabilities are considered similar in characteristics as they are both as likely to have experience the exposure. Using this, individuals with similar probabilities but different exposures can be matched, a method known as propensity score matching (Jupiter, 2017). When matched in this way, the differences in outcomes can be more confidently attributed to the differences in exposure, not personal characteristics, as they have been matched based on their similarity.

There are several methods that utilize propensity scores to adjust for confounders, some of which will be described (Austin, 2011). The first method is propensity score matching, as described in the previous paragraph. Individuals with similar propensity scores are matched into pairs of treated and untreated. Afterwards, the outcome can be compared between the pair to estimate the treatment effect. The second method is stratification upon the propensity scores. Mutually exclusive strata with defined thresholds are created where individuals are placed in strata according to propensity score. Since individuals within the same strata have approximately the same propensity score, the treated and untreated within the strata can be compared to estimate treatment effect. Lastly, the propensity score method can be included as a covariate in the regression model for the outcome. This last method is the one that will be used in this study.

Given the correct variables in the propensity score model, the propensity score can reduce the effect of confounding in the trials. However, there is a lack of consensus on which variables to include in the propensity score model that was used for covariate adjustment (Austin, 2017). Historically, variables related to exposure were primarily used over variables related to outcome (Adelson, 2017). A study conducted by Brookhart et al. (2006) suggested that variables strongly related to outcome and weakly related to exposure should be included while variables strongly related to exposure but weakly to outcome should be excluded. More recent studies found additional different criteria for selecting variables to be included into the propensity score. Austin, Grootendorst, and Anderson (2007) suggested that the inclusion of true confounders (i.e. related to both exposure and outcome) is necessary for a strong propensity model, and the failure to include all confounders resulted in biased estimate of treatment effects. Myers et. al. (2017) found that inclusion of instrumental variables - that is, variables related to exposure but not outcome - may result in increases in bias of treatment effect. The objective of this study is to conduct simulations similar to Brookhart's but with the intention to explore more general scenarios to determine the type of variables that should be selected for the best performance in bias and precision.

2 Simulation settings, generation, and methods

The plan of our study is to expand upon the work of Brookhart, who had used three variables to examine their effects on the propensity score model: X_1 which is related to both outcome and exposure, X_2 which is related to only outcome, and X_3 which is related to only exposure. For these three variables, they had only used a fixed value for their effect size ($\alpha_0^t = .5, \alpha_1^t = 4, \alpha_2^t = 1, \alpha_3^t = 0, \beta_0^t = 0, \beta_1^t = 0.50, \beta_2^t = 0, \beta_3 = 0.75$). Our plan is to conduct a more thorough investigation by using two variables potentially related to outcome, exposure, or both. Rather than fixing coefficient values and making general suggestions based on these results, we simulated 1,000 different magnitudes of effect sizes for the variables. The goal of this study is to explore different scenarios of the variables effect sizes to determine the type of variables to include that will produce low bias while maintaining high precision. All of the following simulations were conducted using R statistical software version 4.0.0, along with the corresponding regression models and analyses of results.

2.1 True Data Generation Model

Here we describe the specific data generation models that establish the causal mechanism. This consists of a true exposure model as a function of X_{1i}, X_{2i} and a separate outcome model conditional on exposure and X_{1i}, X_{2i}. First, a true propensity score model will be established between the continuous covariates X_{1i}, X_{2i} and a binary exposure E_i assuming the form:

$$logit\left\{P[E_i = 1|X_i, \alpha_0^t, \alpha_1^t, \alpha_2^t]\right\} = \alpha_0^t + \alpha_1^t X_{1i} + \alpha_2^t X_{2i}, \tag{1}$$

where α_1^t is true effect of X_{1i} on exposure, and α_2^t is true effect of X_{2i} on exposure. With a large $|\alpha_1^t|$, X_{1i} has a large effect of whether or not a patient will have the exposure E_i. The same applies to α_2^t and X_{2i}. We

assume outcome model of the following form:

$$logit\left\{P[Y_i = 1|X_i, \beta_0^t, \beta_1^t, \beta_2^t, \beta_E^t]\right\} = \beta_0^t + \beta_1^t X_{1i} + \beta_2^t X_{2i} + \beta_E^t E_i, \tag{2}$$

where β_1^t is true effect of the confounder X_{1i} on outcome, and β_2^t is true effect of confounder X_{2i} on outcome. With a β_1^t of large magnitude, X_{i1} will have a larger effect on the outcome. The same applies to β_2^t and X_{i2}. E_i is the binary indicator for the exposure for the individual i while β_E^t is the effect for the exposure indicator on outcome. Lastly, $X_i = (X_{1i}, X_{2i})$ are obtained from a bivariate normal distribution with mean (0, 0) and correlation of ρ^t. These X_i will represent the confounding variables .

2.2 Data Generation

Data generation for each true data generation model consists of 1,000 simulations and a sample size of 1,000. For each of the 1,000 simulations under a true simulation scenario, the following procedures are conducted 1,000 times:

1. Using the randomly generated α_0^t, α_1^t, and α_2^t along with the sampled values of X_{1i} and X_{2i}, E_i is obtained from a binomial distribution with probability equal to that from equation (1)

2. Using the randomly generated β_0^t, β_1^t, β_2^t, and β_E^t along with the sampled values of X_{1i}, X_{2i}, and E_i, Y_i is obtained from a binomial distribution with probability equal to that from the equation (2)

The data generated from this section will be used to fit the propensity score and outcome models in the model fitting stage.

2.3 True Parameter Generation

An even distribution of generated values was desired in order to gain a complete understanding of the effects of α's and β's at all magnitudes. To achieve this, the parameter magnitudes, ($|\alpha_1^t|$, $|\beta_1^t|$, $|\alpha_2^t|$, $|\beta_2^t|$), were sampled with equal probabilities from the intervals (0, .5), (.5, 1), (1,2), and (2, 3).

For each randomly generated sets of parameters, we explored 10 different options which set various α's, β's to 0 to observe the effect of their absence on the resulting model. For example with X_1, we consider choices where $(\alpha_1^t, \beta_1^t) = (0, \beta_1^t), (\alpha_1^t, 0), (0, 0)$, which respectively indicates that X_1 has no effect on exposure, X_1 has no effect on outcome, and X_1 has no effect on neither exposure nor outcome. This applies to X_2 as well. Finally, we look at $(\alpha_1^t, \alpha_2^t) = (0, 0)$ which indicates neither variables has an effect on exposure, $(\beta_1^t, \beta_2^t) = (0, 0)$ which indicates neither variables has an effect on outcome, and $(\alpha_1^t, \beta_1^t, \alpha_2^t, \beta_2^t) = (0, 0, 0, 0)$ which indicates that both variables have no effect on exposure nor outcome. These settings allow us to examine the effect that X_1 and X_2 have on the propensity model when they have no effect or only a single effect - either on exposure or outcome.

For each of the three α^t's and β^t's, a bin was randomly sampled, and a value was generated from uniform distribution bounded by the bounds of the bin that was sampled. For example, if α_1^t had sampled the interval (2, 3), then the magnitude for α_1^t will be drawn from a uniform distribution between 2 and 3 (for example, a value of 2.34). A sample from (-1,1) will determine the positive or negative symbol for each of the three α^t's and β^t's. Lastly, β_E^t is sampled from a uniform distribution bounded by (-2, 2), and the correlation between X_1, X_2, ρ^t, is sampled from a uniform distribution bounded by (-1, 1).

2.4 Checking for Separation

Due to issues with separation, a series of checks are in place to discard any potentially trouble causing pre-generated parameter values. First, a cross-table between E_i and Y_i is examined to determine if there is any separation. If any of the 4 cells are 0, then the current pre-generated parameter values and simulation is marked as having separation, and the corresponding parameter settings will be discarded.

The next check involves examining for separation between confounders X_1, X_2 and the outcome Y_i. This is done by ordering by descending X_1 values and then a second separate ordering by descending X_2 values. If either of the ordering results in Y_i flipping (that is, going down the ordering by either X, the Y_i turning from

1 to 0 or 0 to 1) one or less times, then the current pre-generated parameter values and simulation is marked as having separation.

The last check for separation involves ordering Y_i by the estimated propensity score probabilities, denoted $\hat{\pi}$ for a given propensity score model. Like with the previous separation check, if the Y_i values flips one time or less, then the current simulation and pre-generated parameters will be marked as having separation. Any simulation and pre-generated parameter marked with separation are immediately discarded, and the next simulations will begin with the next set of pre-generated parameters.

2.5 Model Fitting and Operating Characteristics

The model fitting stage involves model fitting in two steps - the first predicts the exposure E_i, and the second predicts the outcome Y_i, using the estimated propensity scores from step 1. The first model fitting step involves obtaining a propensity score estimate $\hat{\pi}$ as a function of X_1 and X_2. These four propensity score models will be investigated to determine how including propensity scores in the outcome model affects estimation of β_E^t:

$$\text{Model (X1X2):} \quad logit(P[E_i = 1|X_i]) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} \tag{3}$$

$$\text{Model (X1):} \quad logit(P[E_i = 1|X_i]) = \alpha_0 + \alpha_1 X_{1i} \tag{4}$$

$$\text{Model (X2):} \quad logit(P[E_i = 1|X_i]) = \alpha_0 + \alpha_2 X_{2i} \tag{5}$$

$$\text{Model (\_\_):} \quad logit(P[E_i = 1|X_i]) = \alpha_0 \tag{6}$$

Afterwards, the second model fitting step uses $\hat{\pi}$ in the outcome model via:

$$logit\left[P(Y_i = 1|E_i, \beta_0, \beta_E, \hat{\pi})\right] = \beta_0 + \beta_E E_i + g(\hat{\pi}), \tag{7}$$

where $g(\hat{\pi})$ is a polynomial function up to degree 4 chosen through forward variable selection and a deviance based test. The estimated propensity score is used in the outcome model to obtain estimates of β_E^t.

3 Simulation Results

After the model has been fitted, the estimate (denoted as $\hat{\beta}_E$), the estimated standard error of $\hat{\beta}_E$, and the p-value of E from the model will be saved. For every simulation within a scenario, the bias and mean square error is calculated then, afterwards, averaged across the entire scenario's simulation replications. Explicitly for each scenario, we compute:

$$\text{B}IAS^* = |\beta_E^t - \hat{\beta}_E| \tag{8}$$

$$\text{M}SE^* = \hat{SE}(\hat{\beta}_E)^2 + (\beta_E^t - \hat{\beta}_E)^2 \tag{9}$$

At the end of each simulation, the operating characteristics are calculated. The mean of the BIAS*, MSE*, and $\hat{SE}(\hat{\beta}_E)$ of each of the four models (X1X2, X1, X2, __) is calculated for each simulation. BIAS* and MSE* are the estimates of bias and mean squared error without taking the expectation. Lastly, the Power of each of the four models is obtained by the proportion of simulations that the null hypothesis of $H_0 : \beta_E = 0$ when β_E^t is greater than .02; and Type I Error is obtained similarly when $\beta_E^t = 0$. These operating characteristics will be examined to determine the effect of X_1 and X_2.

3.1 Overall Simulation Results

Operating characteristics are shown in Table 1 with subtables broken down by different structural α^t, β^t 0's. Table 1.1, which displays the overall results across all structural 0 adjustments, shows that model (X1X2) outperforms all other models in every single measurement observed, except for standard error. Not only is the mean squared error and bias much lower for model (X1X2), but they are also much more tightly controlled with a standard deviation of 0.10 and 0.11 respectively. The average estimated standard error is marginally higher compared to the other models. While the Power of model (X1X2) is marginally lower compared to the other 3 models, the Type I Error is significantly lower at 0.05, with a standard deviation of 0.01. When considering all simulation settings at once, it appears that including both X_{1i} and X_{2i} into the propensity

score model significantly improves the model mean squared error and bias at a minor cost to standard error and power. Furthermore, the Type I Error for model (X1X2) is by far the lowest of any model.

Moving forward to the rest of the settings in Table 1, it appears that (X1X2) consistently appears as a strong performer with the Type I Error staying at .05 or lower. In examining Table 1.2 and 1.3 where there is no effect on exposure or outcome (respectively), when either X_{1i} or X_{2i} only affects outcome, their inclusion, i.e. model (X1X2), improves the model performance. When either X_{1i} or X_{2i} only affect exposure, their removal from the propensity model provides a slight improvement in performance. Table 1.4 represents no effects from the confounders at all, and it can be seen that their inclusion or exclusion has no effect on the performance. When X_{1i} and X_{2i} both affect exposure and outcome, the inclusion of both, i.e. (X1X2), is necessary for a strong propensity score model.

Table 2 examines the individual X's effect on propensity score performance. As X_{1i} and X_{2i} are both generated from the same settings, it would be expected that they would have the same performance, except flipped for (X1) and (X2). By examining and comparing Table 2.1, 2.2, and 2.3 to Table 2.4, 2.5, 2.6 (respectively), this expectation has been proven correct, with the values being very similar. As such, only Table 2.1, 2.2, and 2.3 will be examined.

As seen before, model (X1X2) is consistently a strong performer no matter the circumstances, having only favorable values for the operating characteristics. In Table 2.1 where X_{1i} only influences outcome, it can be seen that X_{1i} should be included when in combination with X_{2i}, i.e. model (X1X2); otherwise, the improvement is small on its own. This same pattern can be seen with Table 2.2 and 2.3. From Table 2.3, the addition of a variable with no relationship to either the exposure or outcome - in this case, X_{1i} - has little to no negative effects on the performance; however, this variable should not be the only variable in the propensity score. Table 2 indicates that the inclusion of the true confounder in the propensity is necessary for strong model performance.

The addition of the non confounder to a true confounder causes little to no negative effect to the propensity model performance in terms of an increase of type II error. However, failure to include a variable that truly affects exposure can result in disastrous increases of type I error rates between .4 and .8. This makes hypothesis testing meaningless for exposure effects. If one errs on the side of caution and includes all possible covariates, they can expect type II error increases of around .01-.05 compared to propensity score models without one of the covariates, regardless of scenario type, with increases around .02-.09 compared to propensity score models without any covariates. We recommend inclusion of all possible exposure effects to avoid type I error inflation.

3.2 Simulation Results by Varying Effect Magnitude

Figure 2 examines the impact on mean squared error when varying the magnitude of the effect of X_{1i} or X_{2i} on outcome or exposure (i.e. $|\alpha_1^t|$, $|\beta_1^t|$, $|\alpha_2^t|$, $|\beta_2^t|$). The dot represents the median while the line represents the 25% percentile and 75% percentile mean squared error. Looking at the top left plot of Figure 2, the first bin (0%, 0%) examines the scenarios where α_1^t is exactly 0. The next bin (0%, 10%) looks at scenarios where α_1^t is between the 0% percentile and 10% percentile out of all randomly generated the simulated α_1^t. In this instance, this bin contains α_1^t between the interval (0.003, 0.178). The next bin (10% - 20%) contains scenarios where α_1^t is between the interval (0.178, 0.374), and so on for the remaining 8 bins.

As before, X_{1i} and X_{2i} portray the same effects but flipped for their (X1) and (X2) model. The first thing to notice is that (X1X2) is the best or among the best performers in all four tables and at all effect sizes. Additionally, (X1X2) consistently has very tight 25% - 75% percentile bounds even as the effect sizes increases. When examining the increasing effect of X_{1i} on either exposure or outcome, the (X1) and (X1X2) mean squared error stays low and constrained while (X2) and (_) increasingly performs poorly and erratically as the X_{1i} effect size increases. This same pattern is seen with increasing effect of X_{2i} on either exposure or outcome - both (X1X2) and (X2) perform well while (X1) and (_) perform poorly.

The top left plot of Figure 3 examines the confounding effect of X_{1i} followed by the effect of the correlation between $X1$ and $X2$, denoted ρ^t, on the right. Since the confounding effect of X_{2i} will mirror the top left plot but with the mean squared error of model (X1) and (X2) switched, it has been omitted. The bottom two plots examine the effect of the confounding effect of X_{1i} at absolute value of the correlation $|\rho^t| < .5$ and $|\rho^t| > .5$ respectively. Looking at the top left plot, the mean squared error at varying levels of the confounding effect of X_{1i} is portrayed. This plot indicates that as the confounding effect of X_{1i} increases in magnitude, the more necessary it is to have X_{1i} in the model, i.e. model (X1X2) and (X1). The models (X2) and (_)

Model	**MSE***	**BIAS***	**SE**	**Power**	**Type I Error**
1.1	**All Scenario Settings Considered**				
X1X2	0.12 (0.10)	0.23 (0.11)	0.23 (0.06)	0.77 (0.33)	0.05 (0.01)
X1	0.38 (0.89)	0.38 (0.43)	0.22 (0.06)	0.79 (0.32)	0.21 (0.33)
X2	0.36 (0.83)	0.37 (0.41)	0.22 (0.06)	0.79 (0.32)	0.23 (0.33)
None	0.84 (1.49)	0.64 (0.62)	0.19 (0.06)	0.81 (0.31)	0.52 (0.45)
1.2	**X1,X2 don't affect exposure:** $\alpha_1^t = 0$**;** $\alpha_2^t = 0$				
X1X2	0.14 (0.14)	0.27 (0.27)	0.21 (0.27)	0.75 (0.27)	0.04 (0.27)
X1	0.13 (0.13)	0.25 (0.25)	0.21 (0.25)	0.76 (0.25)	0.05 (0.25)
X2	0.13 (0.13)	0.25 (0.25)	0.21 (0.25)	0.76 (0.25)	0.05 (0.25)
None	0.22 (0.22)	0.36 (0.36)	0.19 (0.36)	0.72 (0.36)	0.05 (0.36)
1.3	**X1,X2 don't affect outcome:** $\beta_1^t = 0$**;** $\beta_2^t = 0$				
X1X2	0.10 (0.10)	0.19 (0.19)	0.23 (0.19)	0.79 (0.19)	0.05 (0.19)
X1	0.08 (0.08)	0.17 (0.17)	0.21 (0.17)	0.81 (0.17)	0.05 (0.17)
X2	0.08 (0.08)	0.17 (0.17)	0.21 (0.17)	0.81 (0.17)	0.05 (0.17)
None	0.07 (0.07)	0.15 (0.15)	0.19 (0.15)	0.83 (0.15)	0.05 (0.15)
	X1,X2 don't affect outcome or exposure				
1.4	$\alpha_1^t = 0$; $\alpha_2^t = 0$; $\beta_1^t = 0$; $\beta_2^t = 0$				
X1X2	0.08 (0.08)	0.17 (0.17)	0.21 (0.17)	0.81 (0.17)	0.05 (0.17)
X1	0.08 (0.08)	0.17 (0.17)	0.21 (0.17)	0.81 (0.17)	0.05 (0.17)
X2	0.08 (0.08)	0.17 (0.17)	0.21 (0.17)	0.81 (0.17)	0.05 (0.17)
None	0.08 (0.08)	0.17 (0.17)	0.21 (0.17)	0.81 (0.17)	0.05 (0.17)
	X1,X2 both affect everything in some way				
1.5	$\alpha_1^t \neq 0$; $\alpha_2^t \neq 0$; $\beta_1^t \neq 0$; $\beta_2^t \neq 0$				
X1X2	0.15 (0.15)	0.27 (0.27)	0.23 (0.27)	0.74 (0.27)	0.05 (0.27)
X1	0.79 (0.79)	0.65 (0.65)	0.22 (0.65)	0.79 (0.65)	0.44 (0.65)
X2	0.74 (0.74)	0.62 (0.62)	0.22 (0.62)	0.80 (0.62)	0.50 (0.62)
None	1.54 (1.54)	0.98 (0.98)	0.18 (0.98)	0.83 (0.98)	0.79 (0.98)

Table 1: Average operating characteristics (sd) are reported across all randomly generated scenarios.

Model	**MSE***	**BIAS***	**SE**	**Power**	**Type I Error**
2.1	**X1 doesn't affect exposure:** $\alpha_1^t = 0$				
X1X2	0.13 (0.13)	0.25 (0.25)	0.22 (0.25)	0.75 (0.25)	0.05 (0.25)
X1	0.78 (0.78)	0.64 (0.64)	0.21 (0.64)	0.80 (0.64)	0.44 (0.64)
X2	0.14 (0.14)	0.25 (0.25)	0.23 (0.25)	0.75 (0.25)	0.06 (0.25)
None	1.17 (1.17)	0.83 (0.83)	0.18 (0.83)	0.81 (0.83)	0.69 (0.83)
2.2	**X1 doesn't affect outcome:** $\beta_1^t = 0$				
X1X2	0.12 (0.12)	0.22 (0.22)	0.23 (0.22)	0.76 (0.22)	0.05 (0.22)
X1	0.78 (0.78)	0.64 (0.64)	0.21 (0.64)	0.80 (0.64)	0.45 (0.64)
X2	0.09 (0.09)	0.19 (0.19)	0.23 (0.19)	0.80 (0.19)	0.06 (0.19)
None	1.03 (1.03)	0.77 (0.77)	0.18 (0.77)	0.82 (0.77)	0.74 (0.77)
2.3	**X1 doesn't affect outcome or exposure:** $\alpha_1^t = 0$**;** $\beta_1^t = 0$				
X1X2	0.10 (0.10)	0.20 (0.20)	0.24 (0.20)	0.79 (0.20)	0.05 (0.20)
X1	0.78 (0.78)	0.64 (0.64)	0.20 (0.64)	0.81 (0.64)	0.45 (0.64)
X2	0.10 (0.10)	0.20 (0.20)	0.24 (0.20)	0.79 (0.20)	0.06 (0.20)
None	1.15 (1.15)	0.82 (0.82)	0.19 (0.82)	0.83 (0.82)	0.65 (0.82)
2.4	**X2 doesn't affect exposure:** $\alpha_2^t = 0$				
X1X2	0.14 (0.14)	0.25 (0.25)	0.23 (0.25)	0.75 (0.25)	0.04 (0.25)
X1	0.14 (0.14)	0.25 (0.25)	0.23 (0.25)	0.75 (0.25)	0.05 (0.25)
X2	0.73 (0.73)	0.61 (0.61)	0.21 (0.61)	0.81 (0.61)	0.49 (0.61)
None	1.10 (1.10)	0.8 (0.80)	0.18 (0.80)	0.80 (0.80)	0.72 (0.80)
2.5	**X2 doesn't affect outcome:** $\beta_2^t = 0$				
X1X2	0.13 (0.13)	0.23 (0.23)	0.23 (0.23)	0.76 (0.23)	0.04 (0.23)
X1	0.09 (0.09)	0.19 (0.19)	0.23 (0.19)	0.80 (0.19)	0.05 (0.19)
X2	0.73 (0.73)	0.62 (0.62)	0.21 (0.62)	0.81 (0.62)	0.51 (0.62)
None	1.00 (1.00)	0.76 (0.76)	0.18 (0.76)	0.82 (0.76)	0.78 (0.76)
2.6	**X2 doesn't affect outcome or exposure:** $\alpha_2^t = 0$**;** $\beta_2^t = 0$				
X1X2	0.11 (0.11)	0.20 (0.20)	0.24 (0.20)	0.78 (0.20)	0.05 (0.20)
X1	0.10 (0.10)	0.20 (0.20)	0.24 (0.20)	0.79 (0.20)	0.05 (0.20)
X2	0.73 (0.73)	0.61 (0.61)	0.20 (0.61)	0.81 (0.61)	0.52 (0.61)
None	1.09 (1.09)	0.79 (0.79)	0.19 (0.79)	0.83 (0.79)	0.69 (0.79)

Table 2: Average operating characteristics (sd) are reported across all randomly generated scenarios.

begin to drastically rise in mean squared error and widen its range as the confounding effect of X_{1i} increases. Furthermore, it can be seen that the addition of another confounder X_{2i} of unknown effect size does not induce a salient negative effect of the performance as seen with the performance of model (X1X2).

The top right plot of Figure 2 portrays the mean squared error across varying magnitude of correlation. There are no plots in the (0% - 0%) bin, as there are no instances of exactly 0 correlation, since we generated $\rho^t \sim U[-1, 1]$. In all models, the mean squared error stays roughly the same across all values of ρ^t. This is interesting in that it shows that multicollinearity is not an issue when selecting highly correlated variables for the propensity model. The bottom two plots examine the mean squared error at varying levels of the confounding effect of X_{1i} when $|\rho^t| < .5$ and $|\rho^t| > .5$ respectively. When the correlation $|\rho^t|$ is low, then model (X2) performance increasingly deteriorates as the confounding effect of X_{1i} increases; however, when $|\rho^t|$ is high, the performance of model (X2) degrades by a lower rate. This suggests that when X_{1i} and X_{2i} are highly correlated, then the usage of only the other variable can be viable. However, in both graphs, the inclusion of both variables, i.e. model (X1X2), will have consistently low mean squared error at all levels of confounding effect of X_{1i} as well as different levels of correlation $|\rho^t|$.

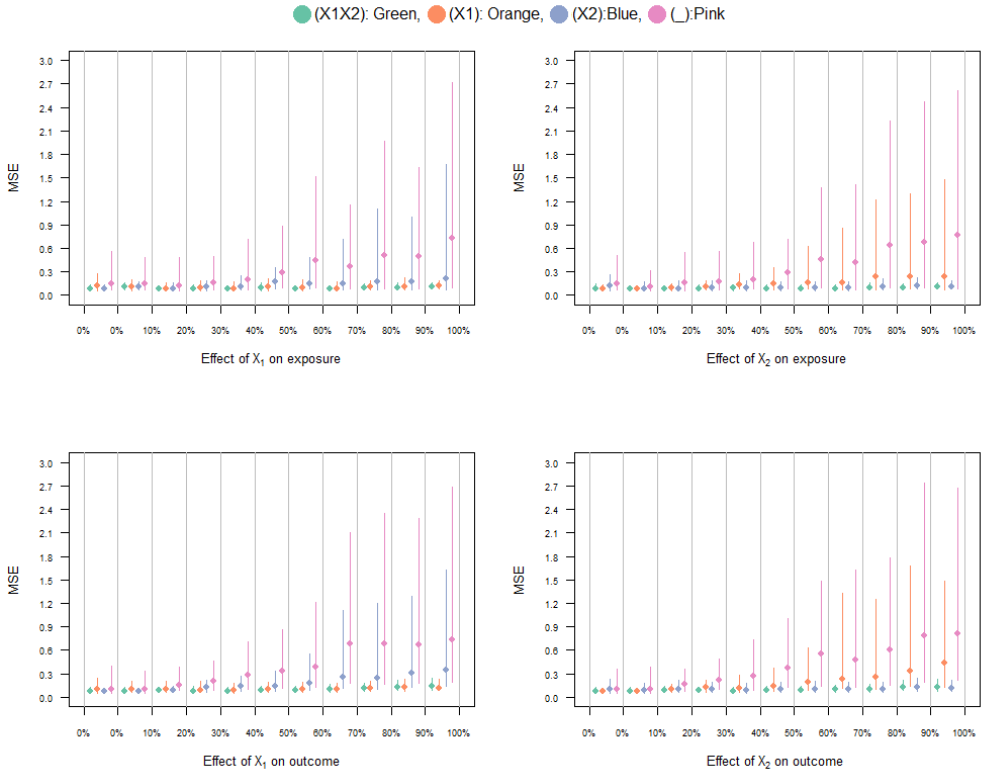

Figure 1: These four table portray the mean squared error of model (X1X2), (X1), (X2), (_) at varying effects that X_{1i} or X_{2i} has on exposure ($|\alpha_1^t|$, $|\alpha_2^t|$) or outcome ($|\beta_1^t|$, $|\beta_2^t|$). In each table, there are 11 bins. The first bin represents when the effect is exactly 0. The remaining bins represents when the effect size is between the left and right percentile of the respective effect. The dot represents the median while the line begins at the 25% mean squared error percentile and ends at 75% mean squared error percentile.

3.3 Simulation Results by Best Model Performance of Mean Squared Error

When going through the results of the simulations, the best model of each simulation was of interest. Since a mean squared error of .052 is not considerably different from a mean squared error of .053, we decided that denoting "best" model by the lowest mean squared error wasn't prudent. Therefore, instead of looking for the absolute best model across each simulation scenario, the best set of models were of interest instead. A model was included among the best set as long as it was within .01 mean squared error of the model with the lowest mean squared error. For example, if the mean squared error for the models were (X1X2) = .052, (X1) = .055, (X2) = .074, (_) = .10, then the best set of models would be (X1X2) and (X1).

The percentage of times a model was among the best performing models are 69.1% (X1X2), 57.9% (X1), 57.6% (X2), 38.9% (_). This indicates that in about 70% of the simulations, the propensity score model that included the both X_{1i} and X_{2i} was among the best models. It has an appreciable improvement (approximately 11%) over the propensity score model that only included X_{1i} and X_{2i}, and a drastic improvement over the propensity score model with neither.

4 Conclusion

The results of the simulation study indicate that the inclusion of true confounders (i.e. related to both exposure and outcome) is necessary for a strong propensity score model. Furthermore, the inclusion of additional variables in the propensity score has negligible negative effect on the performance as seen in Table 1 and 2 and Figure 2. Variables that are related to exposure but not outcome (i.e. instrumental variables) have a slight negative effect when included into the propensity model. As the strength of the true confounder increases, the more necessary it becomes to include it into the propensity score model. As long as a true confounder is included in the propensity score model, the addition of variables with unknown confounding effect does not cause a salient negative effect in the performance of the propensity score.

Since Brookhart's simulation study in 2006, there have been other studies that examined different selection methods such as Bayesian selection, Lasso, or other machine learning methods (see: Zigler and Francesca, 2014; Shortreed and Ertefaie, 2017; Schneeweiss, 2018; Judkins et al., 2007; Cheng et al., 2020; Cefalu et. al., 2017; Parast et al., 2017). However, we will be comparing similar simulation studies that looked at different scenarios. They differ from our paper in that they only examine one or few fixed parameter scenarios as well as their propensity score method. Austin, Grootendorst, and Anderson (2007) conducted a simulation with a different set-up and used the propensity method of stratification. From their results, they found "mean squared error was the lowest when only the true confounders were included in the propensity score model" and "failure to include all confounders can result in ... biased estimation of treatment effects" (Austin et al., 2007). Examining our Table 2.1 (X1 doesn't affect exposure) and 2.2 (X1 doesn't affect outcome), we see that including the non confounder X1 together with X2 in model (X1X2) results in higher average mean squared error and bias when X1 only affect exposure (Table 2.2) - which is often referred to instrumental variable. We can examine Table 1.5 for the second statement where only including one of the confounder X1 and X2 and foregoing the other only give a fractional of the benefit of including both. For the best performing propensity score model, all confounders should be included. Our findings agrees with both of Austin's results.

Adelson et. al. in 2017 also had a different structure as well as focusing on stratification. Their results found that "if a propensity score model does not include variables strongly related to both outcome and assignment, bias will not decrease and may possibly increase when using the stratification method" and "variables with weak associations either to outcome, to assignment, or to both tends to result in a greater decrease in bias ... , without a variable (or composite of variables) strongly related to both outcome and assignment, the propensity score is ineffective at removing bias from the estimated effect size" (Adelson et al., 2017). To make to comparison for these statements, a subset from our simulation data was done where $|\alpha_1^t|$ and $|\beta_1^t| > .8$ while $|\alpha_2^t|$ and $|\beta_2^t| < .5$. For the first statement, we examine the (X1) and (X2) model which had mean squared error = (0.12, 1.72) and bias = (0.22, 1.13), respectively, down from the None (_) Model which had mean squared error = 2.6 and bias = 1.48. This contradicts Adelson's first claim as the weaker confounder X2 still had a positive effect on bias, albeit not as strong as including the true confounder; but Adelson does include that these findings may be limited to stratification. However, for the second statement, our findings match with Adelson's. The model with both the strong and weak confounder (X1X2) with the same subset mentioned previously had mean squared error = 0.12 and bias = 0.22. We see that without the strong confounder X1,

the reduction in bias is minor compared to including it as with model (X1) and (X1X2).

Another study, by Myers et al. focused primarily on instrumental variables - where a variable is associated with the exposure but not with the outcome except through exposure. They stated "estimating an exposure effect conditional on a perfect instrument can increase the bias and standard error of the exposure effect estimate, but these increases were generally small" and "increases in bias and standard error were observed when conditioning on a variable that was strongly associated with exposure and weakly associated with outcome" (Myers et al., 2011). In order to make our data comparable to Myers, the results data set was subsetted where X1 was a perfect instrumental variable (where $\alpha_1^t = 0$) and where X1 had very little correlation to X2 ($|\rho^t| < .05$). The respective average results for (X1X2), (X1), (X2), and (_) are bias = (0.22, 0.97, 0.19, 0.02) and standard error = (0.23, 0.2, 0.22, 0.19). For Myers' first statement, we can see that our results match - conditioning on X1 will cause an increase in BIAS* and SE compared to using no variable at all. Including X1 with X2 (i.e. model (X1X2)) results in a slightly worse performance than X2 (i.e. model (X2)) by itself. For the second statement, a subset was taken where $|\alpha_1^t| > .8$ and $|\beta_1^t| < .2$ with $|\rho^t| < .05$. The results for (X1X2), (X1), (X2), and (_) respectively are bias = (0.2, 0.56, 0.03, 0.51) and standard error = (0.21, 0.18, 0.19, 0.16). We see that comparing the model with a variable highly related to exposure and weakly to outcome (i.e. model (X1)) with model (_) both the BIAS* and SE* increase with the addition of the X1 compared to the model with X2 alone. The also matches with Myer's results across randomly generated relationships.

Lastly, we'll examine VanderWeele (2019) who had suggested a criterion for confounder selection. The results of their simulation had led to suggest "choosing as confounders those variables that are causes of the exposure or outcome or both, then, additionally, discarding any variable known to be an instrumental variable, and including variables that do not satisfy the criterion but are good proxies for unmeasured common causes of the exposure and the outcome" (VanderWeele, 2017). Our findings confirm what VanderWeele (2019) is suggesting - to include confounders with any relationships except for instrumental variables. As mentioned previously with Myers' study, including instrumental variables may increase bias and standard error; but including any other variables will only potentially have a little to negative effects as seen with our examination of Table 1 and 2.

In summary, the five main points after comparing our generalizable results (due to exhaustive simulation) to similar studies are:

1. Excluding variables that relate to exposure (whether or not they relate to outcome) from the propensity score model can drastically increase type I error rates above 30 %. If one includes all variables to be cautious, they can expect a drop in power of around 2-10 % compared to other models.
2. Including true confounders (related both to exposure and outcome) reduces bias.
3. Including instrumental variables (related to only exposure but not outcome) increases bias.
4. Including variables related to outcome but not exposure can reduce bias and mean squared error.
5. Weak confounders may not be needed if strong confounders are controlled for.

As a result, our simulation study suggests that it is recommended to include all possible confounders except for those known to be instrumental variables into the propensity score more in order to achieve the best performance. As seen in table 2, the inclusion of a variable with relationship to only to outcome or neither exposure nor outcome has no appreciable negative effect on the performance of the propensity score model. However, the exclusion of necessary variables in the propensity score model might increase type I errors drastically, indicated by tables 1 and 2.

Conflict of Interest

The authors confirm that this article content has no conflict of interest.

Acknowledgement

This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science (Pordes et al 2007, Sfiligoi et al

2009). This research was also done using resources from the TigerFish Cluster (NSF #2018936).

References

Adelson, J. L., McCoach, D., Rogers, H., Adelson, J. A., and Sauer, T. M. (2017). Developing and applying the propensity score to make causal inferences: variable selection and stratification. *Frontiers in psychology*, 8:1413.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine*, 26(4):734–753.

Bhide, A., Shah, P. S., and Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta obstetricia et gynecologica Scandinavica*, 97(4):380–387.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156.

Cefalu, M., Dominici, F., Arvold, N., and Parmigiani, G. (2017). Model averaged double robust estimation. *Biometrics*, 73(2):410–421.

Cheng, D., Chakrabortty, A., Ananthakrishnan, A. N., and Cai, T. (2020). Estimating average treatment effects with a double-index propensity score. *Biometrics*, 76(3):767–777.

Collier, R. (2009). Legumes, lemons and streptomycin: A short history of the clinical trial.

Cramer, J. S. (2002). The origins of logistic regression.

DeMaris, A. (1995). A tutorial in logistic regression. *Journal of Marriage and the Family*, pages 956–968.

Gilmartin-Thomas, J. F., Liew, D., and Hopper, I. (2018). Observational studies and their utility for practice. *Australian prescriber*, 41(3):82.

Hariton, E. and Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716.

Harsoor, S. and Bhaskar, S. B. (2016). Learning research methodology: Revisiting the evidence. *Indian journal of anaesthesia*, 60(9):619.

Hart, P. D. (1999). A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s. *Bmj*, 319(7209):572–573.

Judkins, D. R., Morganstein, D., Zador, P., Piesse, A., Barrett, B., and Mukhopadhyay, P. (2007). Variable selection and raking in propensity scoring. *Statistics in medicine*, 26(5):1022–1033.

Jupiter, D. C. (2017). Propensity score matching: retrospective randomization? *The Journal of Foot and Ankle Surgery*, 56(2):417–420.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11):1213–1222.

Parast, L., McCaffrey, D. F., Burgette, L. F., de la Guardia, F. H., Golinelli, D., Miles, J. N., and Griffin, B. A. (2017). Optimizing variance-bias trade-off in the twang package for estimation of propensity scores. *Health Services and Outcomes Research Methodology*, 17(3):175–197.

Pordes, R. e. a. (2007). The open science grid. *J. Phys. Conf. Ser. 78*.

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

Schneeweiss, S. (2018). Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical epidemiology*, 10:771.

Sfiligoi, I., Bradley, D. C., Holzman, B., Mhashilkar, P., Padhi, S., and Wurthwein, F. (2009). The pilot way to grid resources using glideinwms. *2009 WRI World Congress on Computer Science and Information Engineering*, 2:428–432.

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122.

Sibbald, B. and Roland, M. (1998). Understanding controlled trials. why are randomised controlled trials important? *BMJ: British Medical Journal*, 316(7126):201.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica: Biochemia medica*, 24(1):12–18.

UCLA: Statistical Consulting (2020). Faq what is complete or quasi-complete separation in logistic/probit regression and how do we deal with them? Accessed: 2021-2-26.

VanderWeele, T. J. (2019). Principles of confounder selection. *European journal of epidemiology*, 34(3):211–219.

Wilson, J. R. and Lorenz, K. A. (2015a). Short history of the logistic regression model. In *Modeling Binary Correlated Responses using SAS, SPSS and R*, pages 17–23. Springer.

Wilson, J. R. and Lorenz, K. A. (2015b). Standard binary logistic regression model. In *Modeling Binary Correlated Responses using SAS, SPSS and R*, pages 25–54. Springer.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505):95–107.

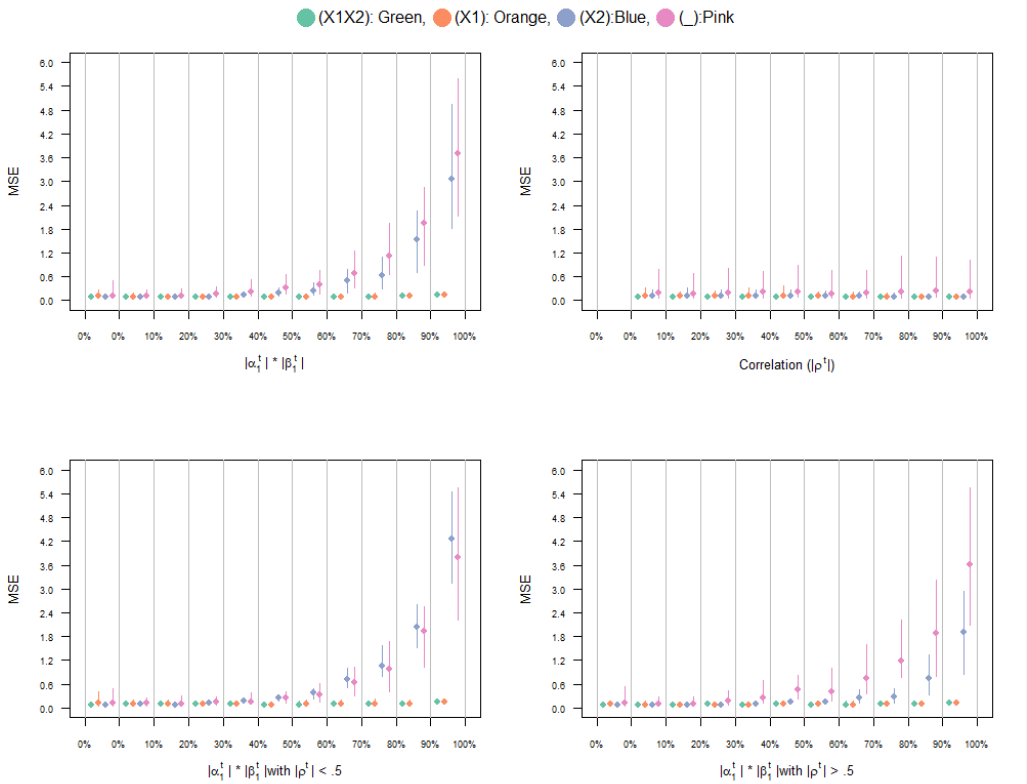

Figure 2: (Top left): portrays the mean squared error of model (X1X2), (X1), (X2), (_) at varying magnitude of the X_{1i} confounding effect ($|\alpha_1^t| * |\beta_2^t|$). (Top right): portrays the mean squared error at varying level of $|\rho^t|$. (Bottom left): table portrays the mean squared error at varying the X_{1i} confounding effect where the correlation $|\rho^t|$ is less than .5. (Bottom right): portrays the mean squared error at varying the X_{1i} confounding effect where the correlation $|\rho^t|$ is greater than .5. In each table, there are 11 bins. The first bin represents when the effect is exactly 0. The remaining bins represents when the effect size are between the left and right percentile of the respective effect. The dot represents the median while the line begins at the 25% mean squared error percentile and ends at 75% mean squared error percentile.