Covariate Balance in Propensity Score Models: Much Ado about Nothing?

Jessica Montgomery, Eun Sook Kim, Jeffrey D. Kromrey, Rheta E. Lanehart, Patricia Rodriguez de Gil, Derrick Saddler, Yan Wang
University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33620

Abstract

Conventional wisdom states that sample covariate balance is needed to obtain unbiased treatment effect estimates in propensity score models; yet the literature offers little empirical evidence of a relationship between sample covariate balance and the quality of the estimated treatment effect. The present study used simulation to investigate this relationship. The factors investigated include correlation among covariates, the strength of relationships between covariates and both treatment assignment and outcome, the number and reliability of covariates, the magnitude of the population treatment effect, sample size, and accuracy of model specification. Each sample was analyzed for both the degree of covariate balance and estimation error in the treatment effect estimate. Results indicate increased balance only yields improved estimates in naïve models. No relationship is evident between sample covariate balance and estimation error in models that adjust for covariate differences. Results are interpreted in terms of the discrepancy between sample estimates and population parameters, and the potential for sample balance estimates to provide useful information about the quality of propensity score models.

Key Words: Propensity score, covariate balance, estimation error

1. Introduction

The investigation of causal relations is often the goal of research enterprises across a wide array of disciplines. Causal relationships can be estimated from non-experimental studies within Rubin's potential outcomes framework (1974), in which for each unit of analysis two potential outcomes are estimated: the response to exposure to treatment and the response to exposure to the control condition. To assess the efficacy of a given treatment, *X*, on an outcome, *Y*, for individual, *i*, we would need two observations, one in which individual *i* was given the treatment and one in which individual *i* was not given the treatment. The measure of the treatment effect (τ_i) would be determined by finding the difference for individual *i* when treatment is applied (Y_{ti}) as compared to when it is withheld (Y_{ci}):

$$\tau_i = Y_{ti} - Y_{ci}$$

However, it is not possible for an individual to be simultaneously assigned to both treatment and control groups (Fundamental Problem of Causal Inference; Holland, 1986). As such for each unit one outcome will be present while the other will be missing.

Holland (1986) stated that by using the framework provided by Rubin's Causal Model, it is possible to derive estimates of the counterfactual situation and thus generate estimates of the treatment effect. While each unit is still only observed as having a singular outcome (i.e., that either experienced following treatment or lack thereof), the aggregate outcomes of units experiencing the alternative assignment provide the missing data needed to calculate the difference between treatment and control outcomes in a similar manner as shown in the equation above. This model thereby "replaces the impossible-to-observe causal effect of *Z* on a specific unit, *i*, with the possible-to-estimate average causal effect of *Z* over a population of units, *U*" (Holland, 1986, p. 947).

For the derived estimates of the treatment effect to accurately reflect the true population parameter, several assumptions must be met, primary among them being the stable unit treatment value assumption (SUTVA) and strongly ignorable treatment assignment. SUTVA refers to the independence between units and asserts that the outcome for a given unit when exposed to treatment should be the same regardless of the selection process employed and regardless of the treatment status of other units. Strongly ignorable treatment assignment focuses more on the actual process used to determine whether a given individual will become part of the treatment group or the control group. An assignment mechanism is said to be strongly ignorable if group membership is not associated with treatment outcome or any other factor.

Experimental designs incorporating random assignment are able to satisfy these assumptions as samples become sufficiently large due to the law of large numbers. Yet experimental designs are not always possible, leading researchers to develop several methods meant to allow observational data to approximate those from randomized experiments. The use of propensity scores is one such method that has gained popularity in recent years (Pearl, 2009). Propensity score methods can aid in the estimation of treatment effects from observational data by accounting for the fact that selection was not random. If the propensity score is properly calculated and an adequate conditioning method is selected it is then possible to calculate treatment effects assuming that control and treatment populations, after conditioning, are similar and thus can be used in conjunction to create the necessary counterfactuals. If the covariate differences between treatment and control groups are minimal, then balance has been achieved.

The requirement that treatment and control populations become adequately matched has led to a focus in the literature on sample balance, with balance being roughly defined as similar covariate distributions across groups. As propensity scores are typically employed to approximate random assignment, much of the extant literature has focused on balance as a necessary quality, suggesting that not only is balance a desirable quality, but that "balance between the treatment groups is the ultimate goal of using the propensity score method. If balance is achieved, then the treatment groups are thought to be comparable in a similar way as if the study was a randomized trial" (Weitzen, Lapane, Toledano, Hume, & Mor, 2005, p. 234).

Even though measures and discussions of sample balance abound in the propensity literature, there is scant empirical evidence justifying this emphasis. In many applications the supposed relationship between balance and error in the estimate of the treatment effect is taken as given. It is assumed that models with better balance inherently produce better estimates, while those with poorer balance, by necessity, show greater divergence from the true population parameter.

The following section provides greater detail on the current conventional wisdom relating covariate balance to estimation error. The methods employed during the conduct of the simulation will then be discussed with the proceeding section detailing the results of the study. The final section will offer an interpretation of the obtained results before turning to implications for practice.

1.1 Balance and Estimation Error

Sample covariate balance is often referenced as an essential quality for propensity score models. Within the literature balance is discussed not only as a factor which can aid in the selection of the most accurate propensity score model, but also as a characteristic which (according to many) must be present in order to make causal inferences. The pivotal role accorded to balance is thus clear. One can use balance measures to ensure the appropriateness of the proposed model and must have adequate balance in order to proceed with estimation.

When creating and utilizing propensity scores researchers are faced with two key decisions. The first involves which covariates should be included when estimating the propensity scores and the second involves the selection of a conditioning method. Given that in a real-world application one is unlikely to know the true mechanism underlying the selection process, assessment of balance measures has been suggested as a way in which researchers can gain confidence in the accuracy (or at least in the usefulness) of their chosen criteria. It has therefore been argued that "balance diagnostics serve an important role in assessing whether the propensity score model has been correctly specified" (Austin, 2008, p. 1224).

When the treated and untreated samples do not exhibit adequate balance, the assumption is that something has been done incorrectly. Thus, "we know we have a consistent estimate of the propensity score when matching on the propensity score balances the raw covariates" (Ho, Imai, King, & Stuart, 2007, p. 219). This focus on the importance of balance measures is due not only to a perceived relationship between balance and correct specification of the PS model, but is also done with an eye towards the overall goal of estimating treatment effects. Therefore, when one analyzes balance measures, one can not only gain assurance that the correct model has been selected, but can also increase confidence in the accuracy of obtained estimates since it has been argued that balance diagnostics are "tools not only for measuring and reporting the amount of balance reached by a given PS model but also for selecting an optimal PS model in terms of bias and variance of the treatment effect" (Ali et al., 2014, p. 806).

Some have placed an even greater emphasis on the importance of balance, suggesting that a lack thereof not only increases bias and/or variance, but that "inferences about treatment effect made using propensity-score matching are valid only if, in the matched sample, treated and untreated subjects have similar distributions of measured baseline covariates" (Austin, 2009, p. 3083). Under this stricter assertion one is not even able to make causal inferences without first achieving an accepted level of balance, again reconfirming the conventional wisdom that "a critical part of using propensity scores is evaluating whether treatment and control groups show remaining differences (e.g. Cohen's *d*) on the propensity score covariates..." (Connelly, Sackett, & Waters, 2013, p. 419).

2. Method

2.1 Design of the Simulation Study

The present study used simulation to investigate the purported relationship between sample covariate balance and estimation error in estimates of the treatment effect. Data were generated using PROC IML in SAS 9.3 (SAS Institute, 2011) with values for explanatory variables being drawn from normal distributions. Several factors were manipulated: correlation between the covariates (r_{12} = 0, .2, .5), the strength of relationships between the covariates and both treatment assignment and outcome (βj = .1, .2, .4), the number of covariates (k = 3, 9, 15, 30), the reliability of covariates (r_{xx}= .4, .6, .8, 1.0), the magnitude of the population treatment effect (Δ = 0, .2, .5, .8), sample size (n = 250, 500, 1000), and accuracy of model specification (correct specification, omitted covariates, incorrect functional form). For each factor analyzed multiple conditioning methods were employed: (a) ignoring the covariates, (b) matching without a caliper, (c) matching with a caliper, (d) stratification, (e) weighting, and (f) ANCOVA. For each assessed condition 5,000 samples were analyzed and for each sample, balance diagnostics as well as error in the estimated treatment effect were calculated to assess the relationship between sample covariate balance and estimation error of the treatment effect.

Multiple models were analyzed to assess the impact of manipulating the factors of interest. However, the design utilized was not fully crossed. Rather, specific combinations of design factors were probed to investigate relationships between sample balance and estimation error across a variety of conditions.

Once the values of key factors were selected, participants' true propensity scores were calculated using the following equation where p indicates the number of covariates and x_i is a vector of covariate values for individual i.

$$\text{logit}\,(\text{Z} = 1) = \log\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right] = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

Each true propensity score was compared to a uniform random number. If the propensity score was greater than the random number, the simulated participant was placed in the treatment group; otherwise the placement was in the control group. Individual outcomes were then determined using the following equation where b_z is the magnitude of the treatment effect after conditioning on the covariates:

$$y = a + b_z Z + \sum_{i=1}^{p} b_i x_i + \varepsilon$$

The data from each simulated sample were assessed using several propensity score conditioning methods. Specifically, samples were matched both with and without a caliper, samples were stratified into quintiles based on the propensity scores, an ANCOVA was performed using the propensity score as a covariate, and observations were weighted by the propensity score. Finally, to provide a comparison condition, the treatment effect was estimated while ignoring the covariates.

Balance in the sample was measured by first calculating the standardized mean difference of included covariates across treatment and control groups.

$$d_i = \frac{\overline{X_1} - \overline{X_2}}{S_p}$$

Two different balance measures were then created using the standardized mean differences for all included covariates. The first indicated the average standardized mean difference across all covariates,

$$\frac{1}{k}\sum_{i=1}^{k}|d_i|$$

and the other provided a count of the total number of covariates exhibiting adequate balance. In this research, standardized mean differences less than .25 were selected as those representing adequate balance because this value corresponds to a small effect size.

$$\#(|d_i| < .25)$$

In addition to balance diagnostics, each sample was also assessed for the degree of error in the estimate of the treatment effect. Estimation error was calculated by finding the difference between the true treatment effect and the estimate obtained from the sample.

$$Error = \widehat{b_z} - b_z$$

Note that this estimation error is neither statistical bias nor RMSE; i.e., the expected value of the estimation error is bias and the square root of the expected value of the squared estimation error is RMSE.

To evaluate the claims in the literature regarding the relationship between sample balance and error in the estimation of the treatment effect these measures were compared within each of the six conditioning methods for each manipulated facet. For every conditioning method/facet combination 5,000 samples were analyzed and results are provided graphically in the following section.

3. Results

Before presenting the results obtained by manipulating the factors of interest, it is useful to provide an illustration of what we might expect to observe if the accepted view of the relationship between balance and estimation error holds. The graphs in Figure 1 show how each indicator of balance would be expected to relate to measures of the estimation error. In the graph on the left it is clear that as the average standardized mean difference between treatment and control groups becomes greater, the amount of estimation error has a tendency to increase. A similar pattern can be found in the graph on the right which plots the total number of balanced covariates against the amount of estimation error. Again this graph illustrates that as balance improves the estimates generally become closer to the true treatment effect.

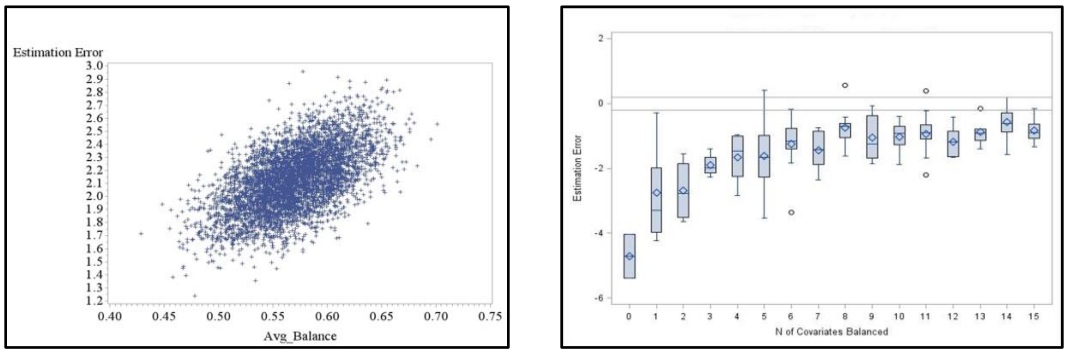

Figure 1: Expected relationship between sample balance and estimation error

In the remainder of this section graphical representations of the association between sample balance and estimation error will be provided for each manipulated factor. For the sake of brevity all manipulated conditions are not provided, rather key graphs representative of general findings are included.

Given that the first reference to balance in many propensity score studies relates to model specification, these components will be addressed first. As discussed above, the literature is replete with arguments relating measures of balance to model accuracy. Based on these assertions it would appear that many authors in this field would agree that a misspecified model will exhibit poor balance while a more accurate model will fare better in terms of balance diagnostics. In this study the impact of misspecification was assessed by either omitting covariates or modeling an incorrect functional form (not including a required polynomial term or omitting a necessary interaction). Results of these analyses are presented in Figures 2-5.

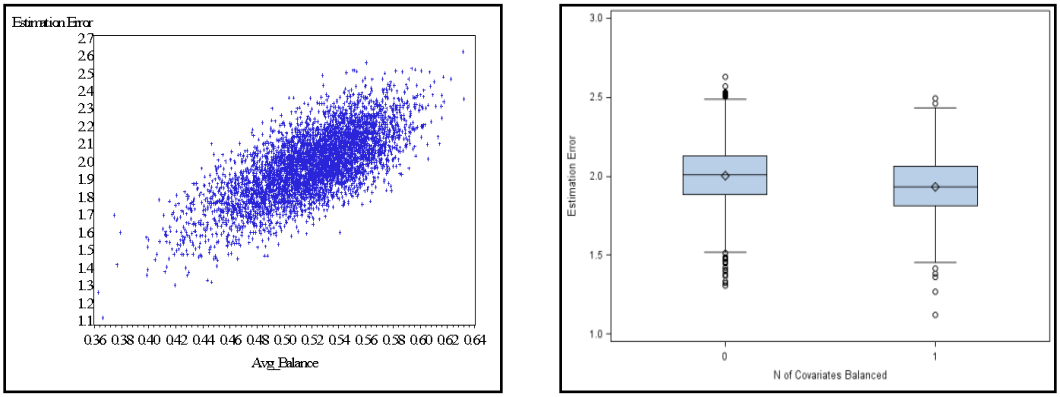

Figure 2: One-third of covariates omitted (Conditioning method = Non-caliper Matching)

Comparing the sets of results in Figures 2 and 3 it is clear that the expected relationship does not always obtain. In Figure 2, in which non-caliper matching is employed, the scatter plot resembles what we would expect to find and the large amount of estimation error may seem initially reasonable given the amount of misspecification in the current model. Yet even with one-third of the baseline covariates omitted, the samples represented in Figure 3 (with caliper matching being employed) largely achieved balance at satisfactory levels. Each of the graphs in Figure 3 also suggest that the relationship

between balance and estimation error may not be as clear as argued in the literature, as no clear association is present in the scatterplot and the boxplots show large amounts of variation in the estimation error even in samples where all included covariates have achieved adequate balance.

A similar pattern is evident with the other forms of misspecification examined. When non-linearity was omitted from the model a relationship is evident between estimation error and balance when ignoring covariates (Figure 4), yet the same relationship is not present when conditioning using caliper matching, as shown in Figure 5.

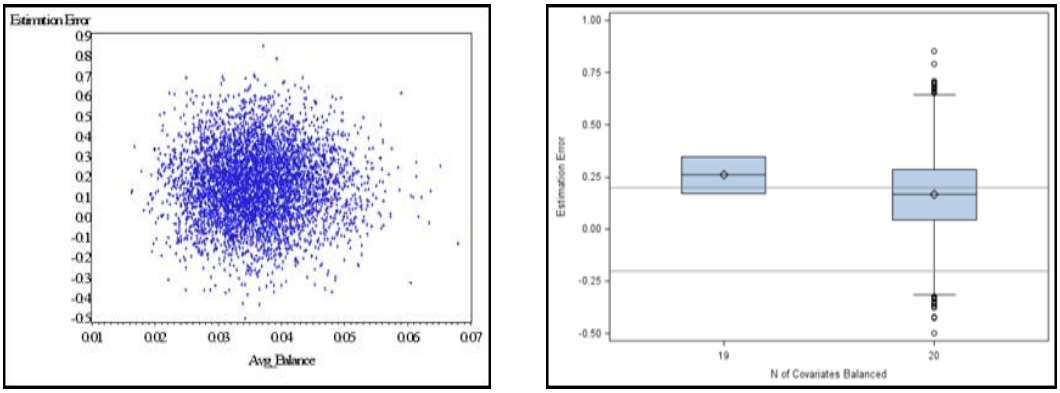

Figure 3: One-third covariates omitted (Conditioning method= Caliper Matching)

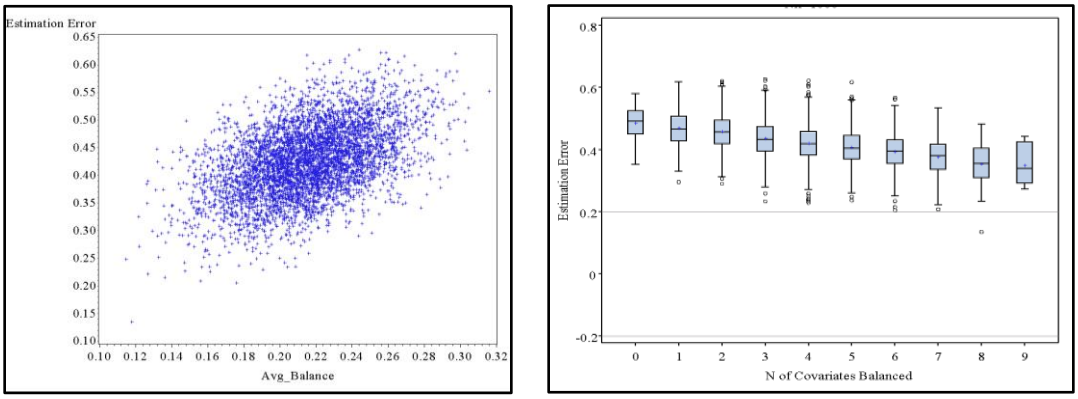

Figure 4: Non-linearity omitted (Conditioning method= Ignoring Covariates)

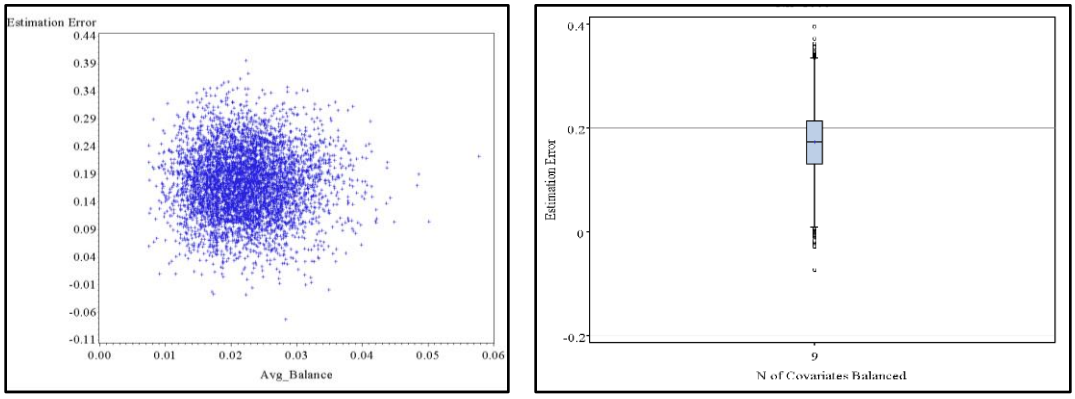

Figure 5: Non-linearity omitted (Conditioning method= Caliper Matching)

Although multiple factors were examined during the course of this study, the general pattern remained consistent. Figures 6 - 9 illustrate the relationship between estimation error and balance under different conditioning methods when the true effect size of the treatment is manipulated. When the population effect size is zero and covariates are ignored (Figure 6), the anticipated relationship between balance and estimation error is evident. However, when stratification is used for conditioning (Figure 7), no relationship is seen. The same difference is obtained when the population effect size is 0.8 (Figures 8 and 9).

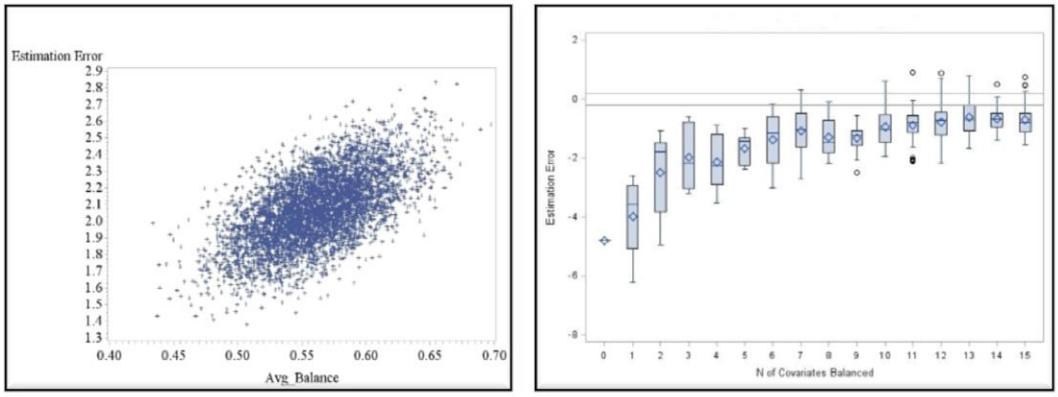

Figure 6: Effect size = 0.0 (Conditioning method = Ignoring Covariates)

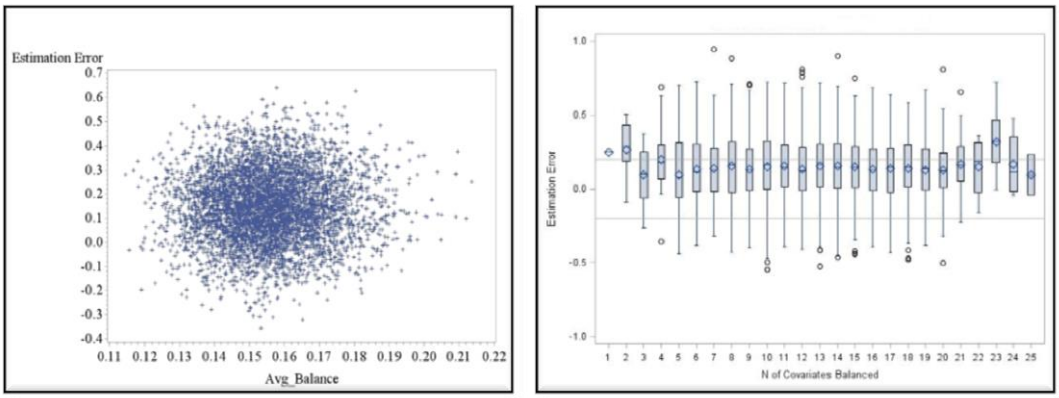

Figure 7: Effect size = 0.0 (Conditioning method= PS Stratification)

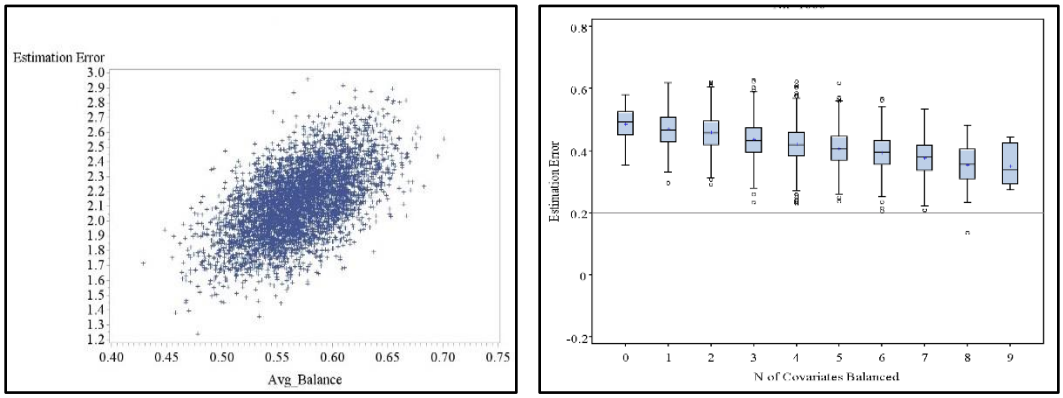

Figure 8: Effect size = 0.8 (conditioning method = Ignoring Covariates)

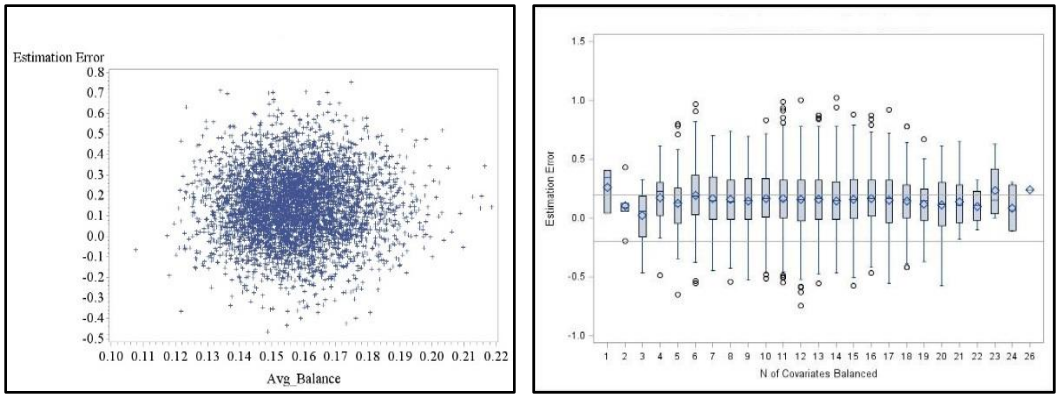

Figure 9: Effect size = 0.8 (Conditioning method = PS Stratification)

The impact of covariate intercorrelation is presented in Figures 10 - 13. With independent covariates (r_{12} = 0), the anticipated relationship between sample balance and estimation error is seen when non-caliper matching is employed for conditioning (Figure 10) but not when caliper matching is used (Figure 11).

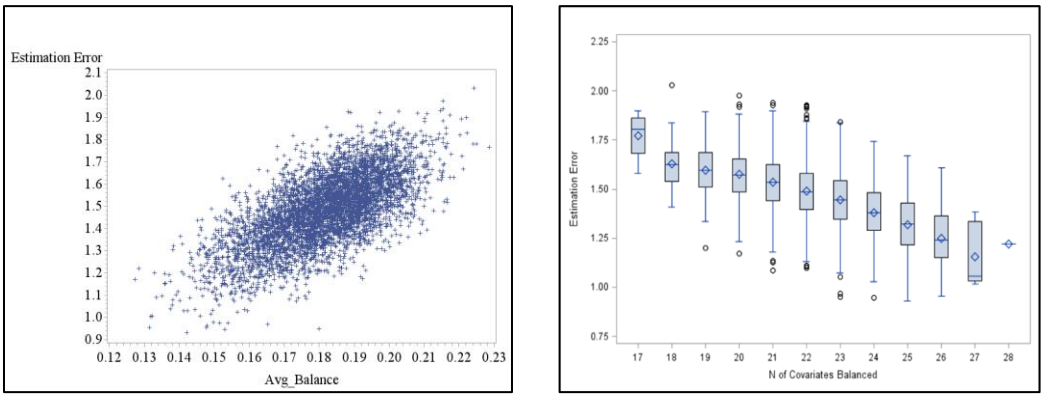

Figure 10: Correlation between covariates = 0 (Conditioning method = Non-caliper Matching)

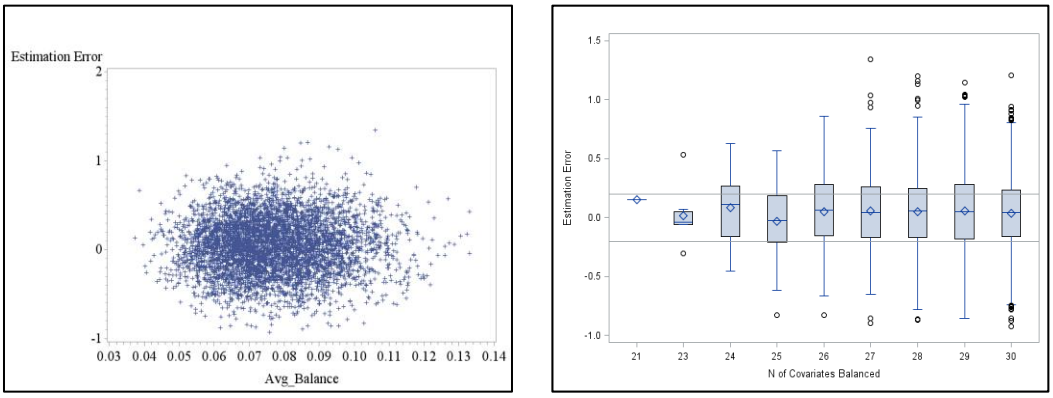

Figure 11: Correlation between covariates = 0 (Conditioning method = Caliper Matching)

The same pattern is found when the covariates are correlated with each other (r_{12} = .50). When non-caliper matching is used (Figure 12), the association between sample covariate balance and estimation error is evident. However, when caliper matching is used (Figure 13), no association is seen.

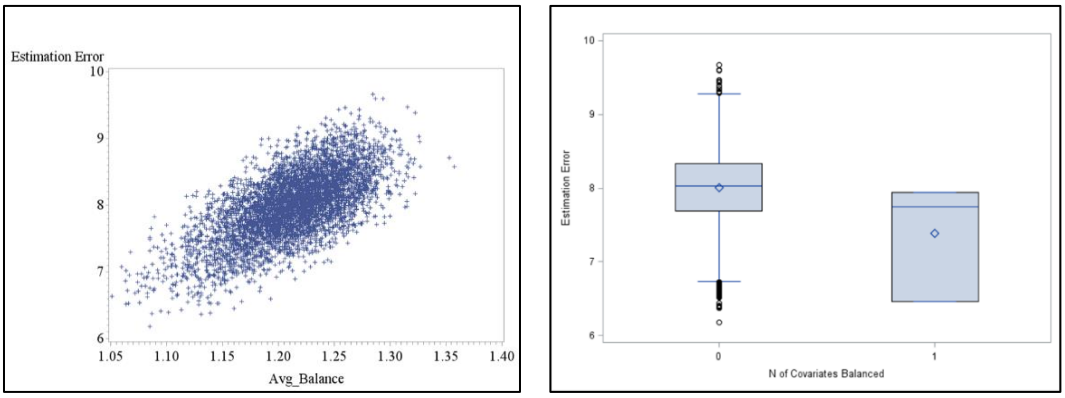

Figure 12: Correlation between covariates = .50 (Conditioning method = Non-caliper Matching)

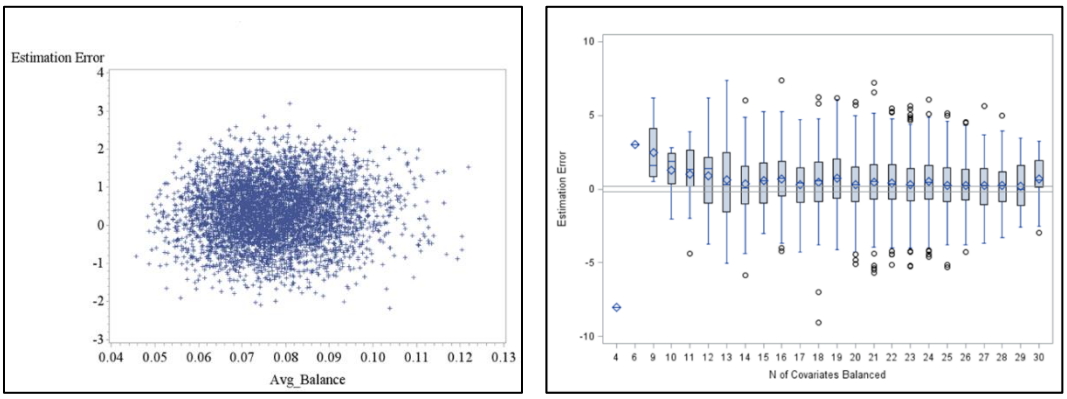

Figure 13: Correlation between covariates = .50 (conditioning method = Caliper Matching)

The final set of graphs (Figures 14 – 17) presents the impact of variation in the strength of relationship between the outcome variable and the covariates. With a modest relationship (β_i = .10), the familiar association between sample balance and estimation error is evident with non-caliper matching (Figure 14) but not with stratification on the propensity score (Figure 15). Similarly, with a strong relationship (β_i = .40), the same association is seen when covariates are ignored (Figure 16) but not when conditioning via stratification (Figure 17).

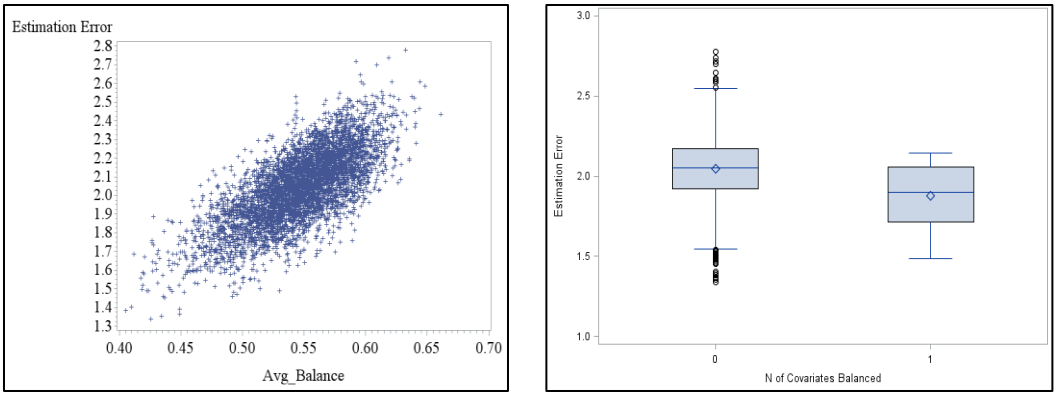

Figure 14: Covariate to outcome regression weight = .10 (Conditioning method = Non-caliper Matching)

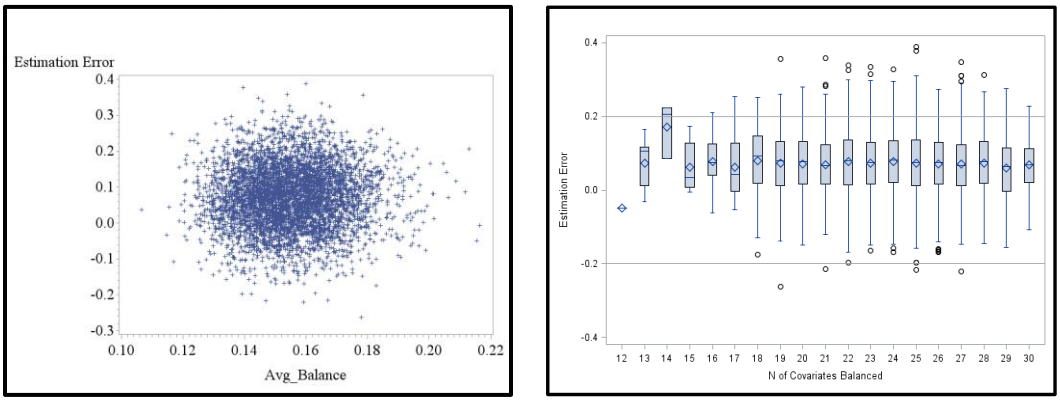

Figure 15: Covariate to outcome regression weight = .10 (Conditioning method = PS Stratification)

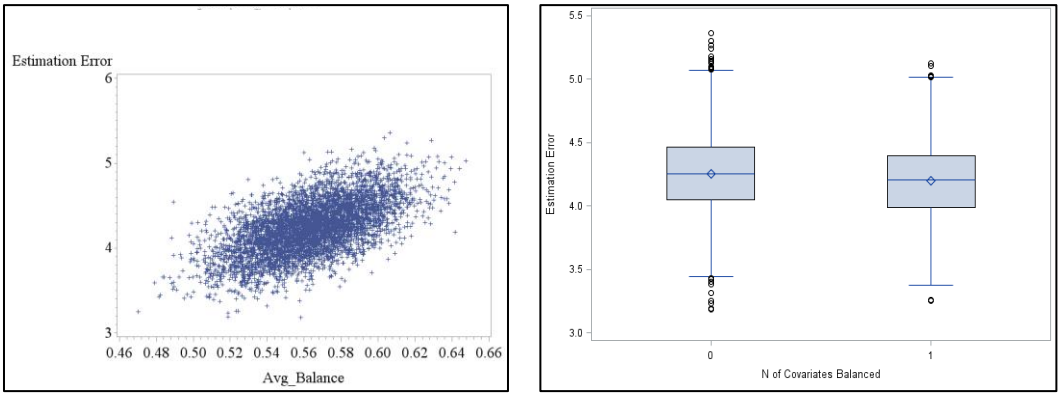

Figure 16: Covariate to outcome regression weight = .40 (Conditioning method = Ignoring Covariates)

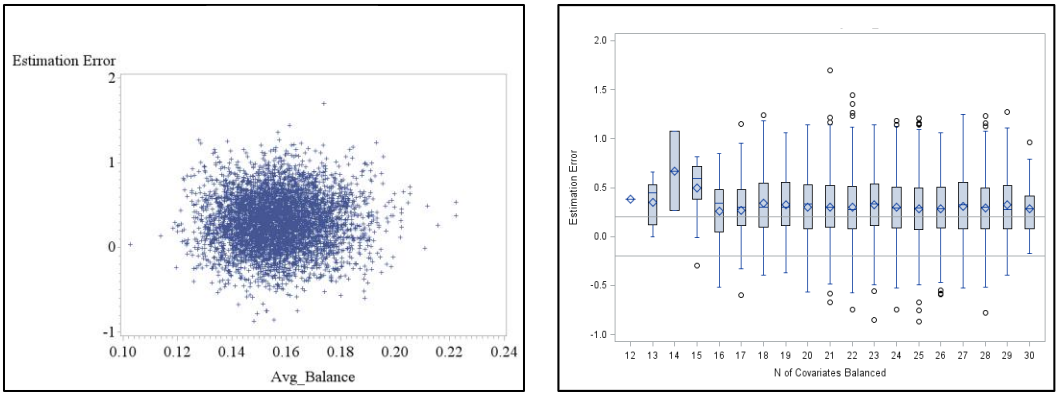

Figure 17: Covariate to outcome regression weight = .40 (Conditioning method = PS Stratification)

4. Discussion

The results of our simulation revealed two consistent patterns. First, the conditioning method selected, rather than the specific factor manipulated had a substantial impact on the resulting relationship between balance and estimation error. Second, certain conditioning methods consistently showed agreement with the conventional wisdom on this relationship, while other methods show little to no relationship at all.

When ignoring covariates or conditioning by matching without a caliper, greater mismatch in terms of sample balance was related to greater divergence between the estimated treatment effect and the true treatment effect. This relationship did not obtain when conditioning using caliper matching, PS stratification, weighting, or PS ANCOVA. On the whole, regardless of which factor was selected for manipulation the latter methods achieved better sample balance and level of balance was uncorrelated with the amount of estimation error calculated based on the given sample. This finding holds true when using either measure of balance employed by the study, be it the average standardized mean difference between treatment and control groups or the total count of covariates achieving adequate balance across groups.

In addition to examining the relationship between sample balance and estimation error this study also sought to determine if balance diagnostics could be used to detect misspecification in the propensity score model. While many have argued that poor balance signals misspecification and as a corollary good balance indicates an appropriate model, the results presented above do not provide support for such a conjecture. Rather, misspecified models tended to exhibit balance in acceptable ranges when certain conditioning methods were employed, but not when others were utilized. Accordingly poor balance does not necessarily entail a misspecified model and good balance does not imply correct specification.

This, however, is not to suggest that balance is not an important factor when using propensity score methods, but rather that the focus on sample balance may be a bit misplaced. Returning to the discussion above on Rubin's Causal Model, in order for researchers to make causal inferences regarding treatment effects we must have sufficient reason to believe that the assignment mechanism is strongly ignorable. This assumption

is met in randomized experiments due to the random assignment mechanism. As the samples grow larger it is less likely that there will be any difference between groups in terms of covariates that may impact the outcome. Yet we must bear in mind that the inferences we aim to make are not to the samples, but are instead to the populations that the samples represent. As such covariate balance in the population is important. If our populations differ significantly in terms of certain factors there is likely a selection process taking place that inhibits us from meeting the assumption of strongly ignorable assignment. But even in the absence of such a selection mechanism, some samples will nonetheless exhibit mismatch on certain characteristics.

Given this it may not be recommended to establish cut points in terms of balance diagnostics, where measures above a certain level are thought to yield accurate results while measures below said level lead to doubts over accurate estimation. As the results presented above show it is possible to derive an estimate near the true effect with poor sample balance and is also quite possible to reach an estimate that diverges greatly from the true effect with samples that are nearly perfectly balanced.

As the prevalence of studies employing propensity score methods continues to increase, this fallibility of sample balance indicators should be kept in mind. Researchers should continue to report balance measures, both before and after conditioning, to provide support for the implicit argument that strongly ignorable assignment has been achieved, but we should not be focused on a specified threshold. Instead of directing such a large amount of attention to balance, researchers should focus on other factors that have repeatedly proved vital to obtaining unbiased estimates in PS models, such as measuring the right covariates and measuring them reliably (Austin, Grootendorst, Normand, & Anderson, 2007; Cook, Steiner, & Pohl, 2009; Shadish, Clark, & Steiner, 2008), treating missing date reasonably (D'Augstino & Rubin, 2000), reporting effect sizes in addition to hypothesis test results (Zhang, Ni, & Xu, 2014), and clearly explicating limits to generalizations (Pearl, 2009).

References

Ali, M. S., Groenwold, R. H. H., Pestman, W. R., Belitser, S. V., Roes, K. C. B., Hoes, A. W., de Boer, A., & Klungel, O. H. (2014). Propensity score balance measures in pharmacoepidemiology: A simulation study. *Pharmacoepidemiology and Drug Safety, 23*, 802-811.

Austin, P. C. (2008). Assessing balance in measured baseline covariates when using many-to-one matching on the propensity score. *Pharmacoepidemiology and Drug Safety, 17*, 1218-1225.

Austin, P.C. (2009). Some methods of propensity score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal, 51*(1), 171-184.

Austin, P.C., Grootendorst, P., Normand, S.L., & Anderson, G.M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine, 26*, 754-768.

Connelly, G. S., Sackett, P. R., & Waters, S. D. (2013). Balancing treatment and control groups in quasi-experiments: An introduction to propensity scoring. *Personnel Psychology, 66*, 407-442.

Cook, T.D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research, 44*(6), 828-847.

D'Agostino, R. B., and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association, 95*(451), 749-759.

Rubin, D. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Ho, D. E., Imai, K.,, King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199-236.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-960.

Pearl, J. (2009). Remarks on the method of propensity score. *Statistics in Medicine, 28*, 1415-1424

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*, 1334-1343.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2005). Weaknesses of goodness-of-fit tests for evaluating propensity score models: The case of the omitted confounder. *Pharmacoepidemioogy and Drug Safety, 14*, 227-238.,

Zhang, Z., Ni, H., & Xu, X. (2014). Do observational studies using propensity score analysis agree with randomized controlled trials in the area of sepsis? *Critical Care, 29*(5), 886.e9-886.