DETERMINATION OF THE OPTIMAL NUMBER OF STRATA FOR BIAS REDUCTION IN PROPENSITY SCORE MATCHING

Allen Akers, B.S., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2010

APPROVED:

Darrell M. Hull, Major Professor
Jon I. Young, Committee Member
Robin K. Henson, Committee Member
Abbas Tashakkori, Chair, Department of Educational Psychology
Jerry R. Thomas, Dean of the College of Education
Michael Monticino, Dean of the Robert B. Toulouse School of Graduate Studies

Akers, Allen. Determination of the optimal number of strata for bias reduction in propensity score matching. Doctor of Philosophy (Educational Research), May 2010, 51 pp., 3 tables, 11 figures, references, 49 titles.

Previous research implementing stratification on the propensity score has generally relied on using five strata, based on prior theoretical groundwork and minimal empirical evidence as to the suitability of quintiles to adequately reduce bias in all cases and across all sample sizes. This study investigates bias reduction across varying number of strata and sample sizes via a large-scale simulation to determine the adequacy of quintiles for bias reduction under all conditions. Sample sizes ranged from 100 to 50,000 and strata from 3 to 20. Both the percentage of bias reduction and the standardized selection bias were examined. The results show that while the particular covariates in the simulation met certain criteria with five strata that greater bias reduction could be achieved by increasing the number of strata, especially with larger sample sizes. Simulation code written in R is included.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DETERMINATION OF THE OPTIMAL NUMBER OF STRATA FOR BIAS REDUCTION IN PROPENSITY SCORE MATCHING

1. Introduction

Fully randomized experimental designs are generally regarded as the gold standard for any research involving a treatment or intervention (Barth, Guo, & McCrae, 2008). This is due to the fact that the randomization process should remove any bias due to pre-existing factors among the participants by equally dispersing them amongst the treatment (or levels of treatment) and control groups. The rigorous controls needed for a fully randomized experiment are not always possible, due to practical or ethical reasons (Austin, Grootendorst, & Anderson, 2007). In medical research, people are generally treated based on their need for treatment rather than randomly assigned to treatments, for ethical reasons. In the fields of economics, social work, and education, data is mainly observational due to the breadth and scope of the subject. Trying to implement experimental designs in these areas would not only raise ethical questions, but also would likely greatly diminish the external validity, due to the circumscribed nature of the setting (Störmer et al., 2006).

1.1. Development of Propensity Score Analysis

For these reasons and others, a method was needed that would help adjust for the fact that participants were not randomly assigned. One of the most prominent methods (especially in the aforementioned fields of study) is propensity score analysis (PSA) (Störmer et al., 2006; D'Agostino & Rubin, 2000; Bhattacharya & Vogt, 2007). PSA is grounded in the counterfactual framework (Winship & Morgan, 1999), which is the assumption that all participants have a theoretical outcome as both treated and non-treated, regardless of the group within which they are placed. In non-experimental designs there is usually some bias for treatment group. This

bias is a measure of imbalance between the groups on covariates over which the researcher has no control. The aim of PSA and other adjustment methods is to reduce this bias as much as possible and to make a robust estimation of the counterfactual.

Prior to PSA, the primary method of accounting for bias was to match all treatment participants with control participants based on all of the measured factors that might influence treatment selection. This is an arduous and sometimes-impossible task to find matching pairs between groups of subjects on numerous individual factors where all must be in alignment simultaneously. In response, Rosenbaum and Rubin (1983) developed the PSA, which combines all of the individual factors into a single scalar that may be used for matching. The propensity score is the likelihood of a participant to be placed into the treatment group (or a particular treatment group) or control group. In a fully randomized experimental design with a single treatment group and a control group, each person would have a true propensity score of 0.50 (with any variation being due to sampling error), meaning that they are just as likely to be selected for the treatment group as the control group. In a research design without random assignment, the propensity scores for participants must be estimated and these estimates will likely not cluster around 0.50, but will be distributed across the range of possible values.

$$e(x_i) = p(Z_i = 1 | X_i = x_i) \qquad (1)$$

Where for participant i, $e(x_i)$, is the probability that participant i is assigned to one group (e.g., treatment), $Z_i = 1$, given his/her scores, x_i, on the set of covariates, X_i.

Luellen, Shadish, and Clark (2005) presented the following general introduction to propensity scoring:

1. "A propensity score is the conditional probability that a person will be in one condition rather than in another (e.g., get a treatment rather than be in the control group) given a set

of observed covariates used to predict the person's condition (Rosenbaum and Rubin 1983a)." [p. 531]

2. "With a quasi-experiment, the true propensity score function is not known and must be estimated. The probabilities of receiving treatment (i.e., propensity scores) are a function of individual characteristics and are likely to vary from 0.50. For instance, if the researcher dummy codes treatment as 1 and control as 0, then a propensity score above 0.50 would mean the person was more likely to select into treatment than control, and a score below 0.50 would mean the opposite." [p. 532]
3. "Because propensity scores are derived from observed covariates…a crucial step in designing a quasi-experiment is identifying potentially relevant covariates to measure. Potentially relevant covariates are those expected to affect treatment selection and outcomes." [p. 532]
4. "Researchers can use propensity scores to balance nonequivalent groups using matching, stratification, covariance adjustment, or weighting on the propensity score." [p. 532]

One of the basic assumptions of the application of propensity scores is that all (or at least the overwhelming majority of) variables related to treatment assignment are included in the selection model. Luellen, Shadish, and Clark (2005) also included variables related to treatment outcome, but there is other evidence (Bhattacharya & Vogt, 2007) suggesting that if variables are related more to outcome than treatment assignment they should not be included in the development of the propensity scores because of the likelihood of introducing further bias into treatment assignment. Therefore, variables that may be more highly correlated with outcomes than with treatment selection should be considered with great caution and possibly omitted from the calculation of propensity scores.

1.2. Estimation of Propensity Scores

Once all variables determined to be influential in selection assignment have been determined, the propensity scores may be estimated using regression (typically logistic) or discriminant analysis (less common). Using logistic regression as an example, the dependent variable is a dummy coded treatment assignment, with 0 being control and 1 being treatment. All contributing factors are then used as independent variables in the model. Once the model has been developed then propensity scores are calculated by running the individual values of independent variables for each participant through the resulting equation to obtain an estimate.

A major assumption in PSA is that all relevant differences between control and treatment groups can be captured by observable characteristics in the data. If there are missing factors then it will be impossible to obtain a good model fit and to be able to make meaningful inferences. It is also assumed that treatment and control groups have significant overlap in the covariates being used in the PSA model and that they are normally distributed. If the range of a treatment group for certain covariates lies entirely out of the range of the same covariates for the control group then judgments made based on the resulting propensity scores would be invalid (Caliendo & Kopeinig, 2006). Due to this fact, it is very important to find a control group that is representative of the same population from which the treatment group was drawn. Barring the ability to do this, other statistical controls and adjustments may need to be made in addition to the use of propensity scores.

2. Propensity Score Techniques

Once the propensity score model has been determined to be the best possible, given the available data, then the scores will need to be applied to the participants, which is usually in one of three ways: matching, stratification, or regression adjustment (D'Agostino & Kwan, 1995).

Both matching and stratification are ways of matching or grouping like individuals for the purpose of later statistical analysis, whereas regression adjustment is typically the calculation of the treatment effect itself. There are numerous matching methods and each technique is best suited to a particular set of data. Typically, several methods are applied to a given set of data to determine which method achieves the greatest level of balance among the covariates.

2.1. Nearest Available Neighbor Matching

There are numerous ways that matching has been implemented and several methods should generally be attempted to determine which results in the greatest bias reduction. The first type of matching is "nearest available neighbor" matching based entirely on the propensity score. This is the easiest method to employ and takes on several forms. There can be a one-to-one matching between treatment and control, with any unmatched participants being discarded. This is generally considered a waste of potential data and a method that allows matching of a single treated participant with up to four control participants (or vice versa) is recommended (Rosenbaum, 2002).

2.2. Matching Within Calipers

Another type of matching is matching within calipers. (Rosenbaum & Rubin, 1985a) With this type of matching, matches must be less than a predefined difference from the control propensity score value. There is a tradeoff between obtaining inexact matches or having incomplete matches (and therefore lost data). Grossly inexact matches may greatly overestimate resulting treatment effects, whereas incomplete matches will reduce the sample size and consequently reduce the power of any statistical method used to determine treatment effect. Therefore, it is extremely important to utilize an optimal caliper size based on previous work or calculations based on the precision/loss tradeoff.

2.3. Mahalanobis Metric Matching

Another type of matching is Mahalanobis metric matching (Rubin, 1980), which has been used both in place of PSA and as a supplement to PSA. As a replacement to PSA, the Mahalanobis distance for the covariance matrix of all covariates for a randomly chosen treatment participant and each of the control participants is calculated and the closest (smallest) is matched. Both individuals are removed from the pool and the next treatment participant is chosen. This process continues until all treatment participants have been matched. As a supplement to PSA, either the propensity score can be included in the covariance matrix or calipers are utilized and only those corresponding participants within the caliper range have their Mahalanobis distances calculated. Participants are matched and removed from the pool, as mentioned above.

2.4. Kernel Based Matching

With kernel based matching (Caliendo & Kopeinig, 2006), each person in the treatment group is matched to a weighted sum of individuals who have similar propensity scores with the greatest weight being given to people with closer scores. Some kernel based matching uses all people in non-treated group (e.g. Gaussian kernel) whereas others only use people within a certain probability user-specified bandwidth (e.g. Epanechnikov). The choice of bandwidth involves a trade-off of bias with precision.

2.5. Stratification

Also called subclassification, stratification separates control and treatment participants into strata where all have propensity scores within a certain range (Lunceford & Davidian, 2004). This is typically a separation into quintiles, but a larger number of strata may also be used. Based on a review of the literature, though, using any number of strata other than 5 appears to be a less common practice (Austin, 2008; D'Agostino, 1998; Leon & Hedeker, 2007; Lunceford &

Davidian, 2004; Störmer et al., 2006). This may be due to the fact that it limits the consequent statistical analysis to a multi-level general linear model, reducing the flexibility of the researcher to explore other options. Because having 5 strata has already imposed this restriction, it is more likely that it has just become the de facto standard and is used as a matter of convenience and to make results comparable to previous studies.

2.6. Regression Adjustment

With regression adjustment, the researcher immediately begins statistical analysis once propensity scores have been estimated and uses them, possibly alongside variables that could not be balanced, as predictors in a regression, survival or logistic model (D'Agostino, 1998). Separate regressions can be fitted by propensity score quintile, to estimate the treatment effect within quintiles, as well as the overall treatment effect.

2.7. Calculating Treatment Effect

If the estimated propensity scores are not directly used for regression estimation then there is still a need to run a statistical analysis to calculate the treatment effect. Common procedures following PSA are multiple regression, application of a general linear model, survival analysis, structural equation modeling, or hierarchical linear modeling (Luellen, Shadish, & Clark, 2005). To insure that the development of the propensity score model did not influence the treatment effects, it is important to use another sample of data in order to cross-validate the results. This may be done by either randomly holding back a considerable amount of participants from the initial analysis or by obtaining a new sample of participants gathered from the same population and for whom the exact same covariate values are known.

2.8. Determination of Bias Reduction and Robustness

There are sensitivity analyses that may be conducted to determine bias reduction, but the best estimate is to compare against the "gold standard" of a fully randomized experimental design. There are two ways to do this (or simulate it). The first involves obtaining actual data from an experiment that utilized randomized assignment to treatment and found a significant treatment effect. In addition to this, the data must also contain any measures that could be considered confounding measures in a purely observational study, including any that could lead to self-selection bias or other treatment selection bias. This would be difficult to do, since most of these measures would be unnecessary given the experimental controls being placed on the design, unless the purpose of the study was to provide later data for PSA validation. In most cases, such an approach is untenable.

The more feasible option is to simulate data that models an experimental study and fabricate covariates. In this manner, a very large dataset can be created and can be manipulated in many ways to accommodate different data conditions. Sampling conditions can be altered to simulate variations from the assumptions in order to determine the robustness of the PSA in these situations. The limitation to this approach is the very fact that the data is simulated and must be simulated based on certain assumptions and algorithms. Due to this, it would be easy to draw incorrect conclusions by simulating exactly the conditions necessary to demonstrate robustness which may not be relevant to the data in any particular real world situation.

3. Variable Selection in Propensity Score Matching

Generally, the original work of Rosenbaum and Rubin (1983a, 1983b, 1984, 1985a, 1985b) suggested that any covariate that could be considered to explain confounding should be included in the propensity score model. There have still been disagreements in the literature

about which variables should be included, whether those related to assignment or outcome or both. Bhattacharya and Vogt (2007) showed through simulation and empirical analysis that strong instrumental variables should not be used to develop propensity score estimations. They further caution that "since there is no statistical test to determine whether a particular variable is an instrument, the researcher must rely on knowledge about the problem to assess which variables are appropriate instruments and which variables are appropriate propensity score matching predictors" (p. 20).

Another recent simulation study suggested that variables related to the exposure, but not to the outcome will increase the variance of the estimated exposure effect without decreasing bias (Brookhart et al., 2006). This is somewhat counterintuitive, considering that exposure (or treatment selection) was the primary thing propensity scores were created to predict. Corroborating these findings is another simulation study by Austin, Grootendorst, and Anderson (2006) that found that "including a variable that is related to treatment, but not to outcome, does not improve balance and reduces the number of matched pairs available for analysis." Since it was previously shown that incomplete matching left much greater residual bias than inexact matching, this would be a poor trade-off in terms of the overall effectiveness of the analysis.

The majority of research involving variable selection has focused on individual matching techniques, but Leon and Hedeker (2006) investigated the effect of misspecified propensity scores in longitudinal treatment with a simulation and determined that the same type of variable inclusion is important with stratification matching as with individual matching methods. Given the conflicting viewpoints on which variables should be included based on individual matching methods, this is both a boon and a bane to the researcher employing stratification. Most research

seems to defer to the original guidelines of Rosenbaum and Rubin, to include any variable that may explain confounding.

4. Rationale for the Present Study

Although the majority of research in propensity score matching has focused on nearest available matching and its variants, there are certain instances when that type of exact matching is either not possible or is unreasonably prohibitive, such as smaller populations from which to construct matches or at the opposite end of the spectrum when there is a large population and the researcher wants to include as much control data as possible. Because of this, there is still significant interest in stratification matching.

Throughout the articles implementing stratification, the general consensus has been to use quintiles (Austin, 2008; D'Agostino, 1998; Leon & Hedeker, 2007; Lunceford & Davidian, 2004; Stürmer et al., 2006). Typically the reason for this has either been stated that 5 strata was already considered the de facto standard in previous studies or refer to a 1965 article by Cochrane and Chambers that determined that "five subclasses are often enough to remove 95% of the bias associated with a single covariate." Rosenbaum and Rubin (1984) suggested that separation into only 5 strata based on the propensity score was enough to remove 90% of the bias that could be removed by individually matching on all covariates. However, Lunceford and Davidian (2004) demonstrated that bias is increased with greater sample sizes if stratification is limited to quintiles, due to residual confounding as the datapoints within strata become more heterogeneous. The authors suggested that there is a trade-off point between bias and variability and that future research in this area should focus on finding the optimal number of strata given larger sample sizes. Recently, Rubin (2010) suggested grouping into 5 to 10 strata, but with no justification or guideline as to how this number should be determined. The percentage of bias

reduction is not an effective measure of overall bias, because of the fact that it is relative to the initial bias. Therefore, researchers typically use standard bias values as an absolute criterion with $d < 0.20$ or 0.25 as maximum allowed bias scores, but strive to get these numbers as close to zero as possible (Shadish & Steiner, 2010). Most attempts to reduce bias, though, have been through changes in matching strategy or included covariates and no systematic examination of the effect of number of strata on bias has been undertaken.

The purpose of the current study is to fill this gap with quantitative evidence for guidelines for choosing an optimal number of strata in order to sufficiently reduce bias while maintaining adequate within stratum variability. Are 5 strata really enough to remove 90% of the bias in the propensity score? How many strata are required to remove 90% of the bias in the individual confounding covariates? How many strata are required to reduce standard bias below 0.20?

5. Methodology

The purpose of the current study was to determine this optimal number of strata for reducing bias in larger sample sizes. The method by which this was accomplished was through a large-scale simulation encompassing a range of sample sizes broken into various number of strata and analyzing the resulting balance and bias. Treated and comparison groups were kept equal in size to avoid introducing imbalance to cells that might complicate interpretation.

The R statistical software package was utilized to create large datasets of simulated data as well as to perform the subsequent analyses. The MatchIt library provided the propensity score matching and the bias measures for the simulated datasets. The R function that performed the simulations is included in Appendix B. The parameters passed to the function are the number of iterations (1,000 were used in this study in order to provide stability) and the sample size. The

sample size is the size of the treated group which is generally the limiting factor in most studies. A correspondingly sized sample is used as controls for matching. The psmSim function was run for sample sizes of: 100, 300, 500, 1000, 3000, 5000, 10000, 30000, and 50000. Sample sizes of smaller than 100 did not lend themselves well to propensity score matching due to the large number of missing simulations for which matching could not be performed and were therefore left out of the study.

5.1. Description of the Simulation Function

The function begins by creating a matrix in which to store the summary data from all of the simulation iterations for a given sample size. It then sets the means and standard deviations to be used for all relevant variables. These values were chosen at reasonable values that could equate to score ranges on tests, biometric measures, or demographic data and used throughout all simulations. The X_1 variable is related to outcome only and is included based on recommendations (Rosenbaum & Rubin, 1984) that all variables that could possibly be confounding be included in the propensity score calculation. Since the end goal of this study is optimal matching of datasets rather than a full analysis, its relation to outcome is not realized, but is included for consistency and as a means for comparison with later work that may expand upon this simulation framework. The X_2 variable is related only to treatment selection and the X_3 variable is related to both treatment selection and (theoretically, though also not realized in the current study) outcome.

Next, normal distributions are created for each of the variables for the simulated participants based on the previously defined means and standard deviations. These distributions are created with twice the sample size that was passed into the function for reasons that are explained shortly. Creation of the variable distributions is followed by minimums and ranges

being calculated for the two variables related to treatment selection: X_2 and X_3. These are subsequently used as a scaling method by which 60% of the treatment selection is attributed to X_2 and 40% of the treatment selection is attributed to X_3. The top half of the resulting scores are selected for "treatment" and the bottom half discarded, which is why the initial samples were created at twice the requested number.

Now that the treatment group has been created, another group of controls are simulated using the same means and standard deviations for each of the variables. This insures that the assumption of shared support is upheld. As can be seen in Figs. 1, 2, and 3 (based on a sample size of 50,000) there is considerable overlap between the distributions of X_1, X_2, and X_3 in the treated and control populations. This is not at all surprising for X_1, since it played no role in treatment selection and therefore should present no selection bias.

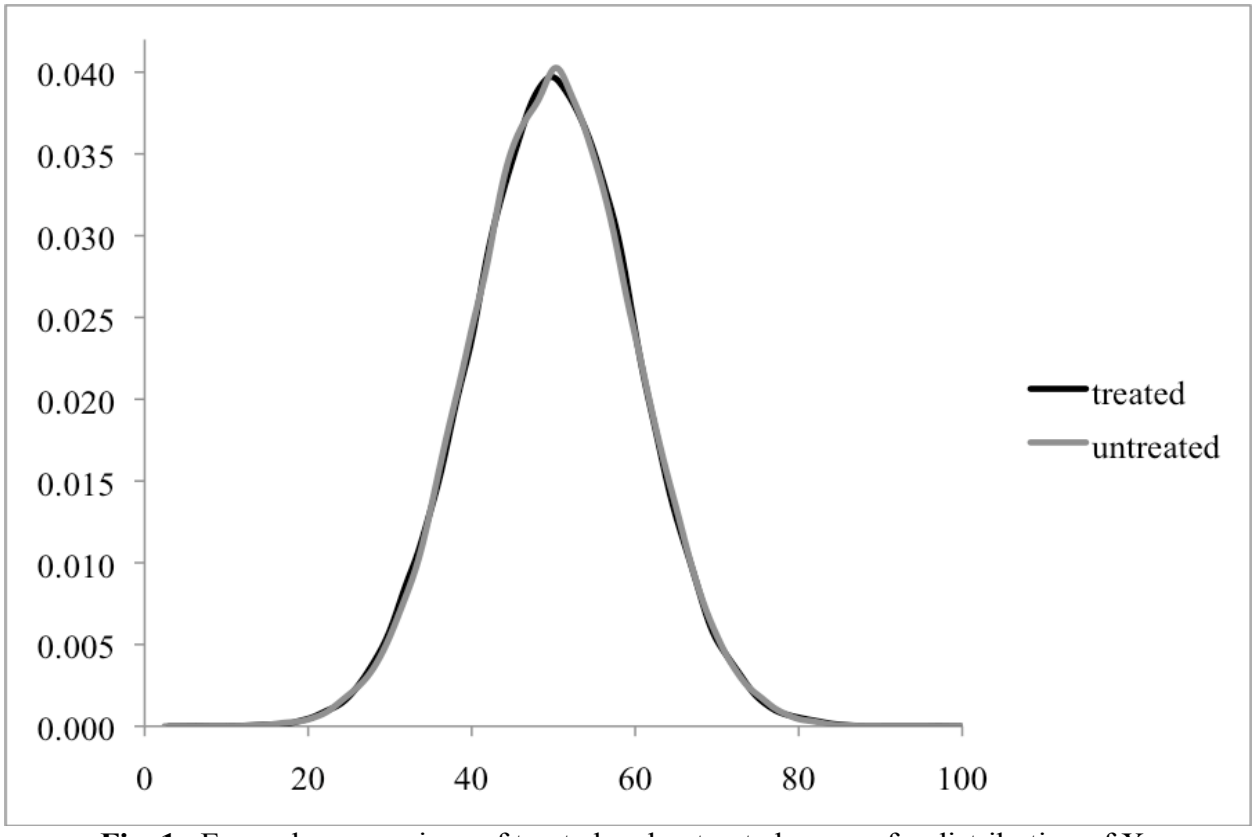

Fig. 1. Example comparison of treated and untreated groups for distribution of X_1 with $n = 50{,}000$.

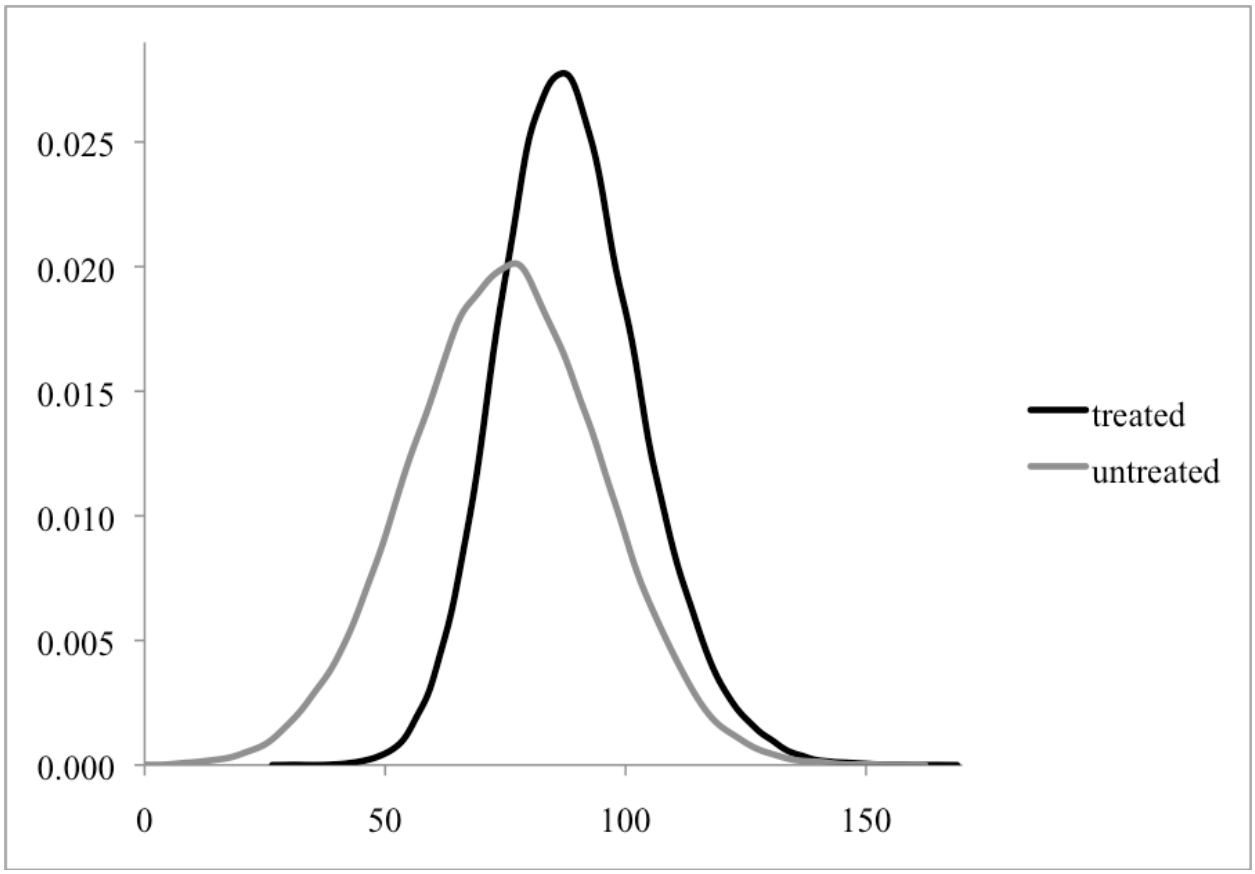

Fig. 2. Example comparison of treated and untreated groups for distribution of X_2 with n = 50,000.

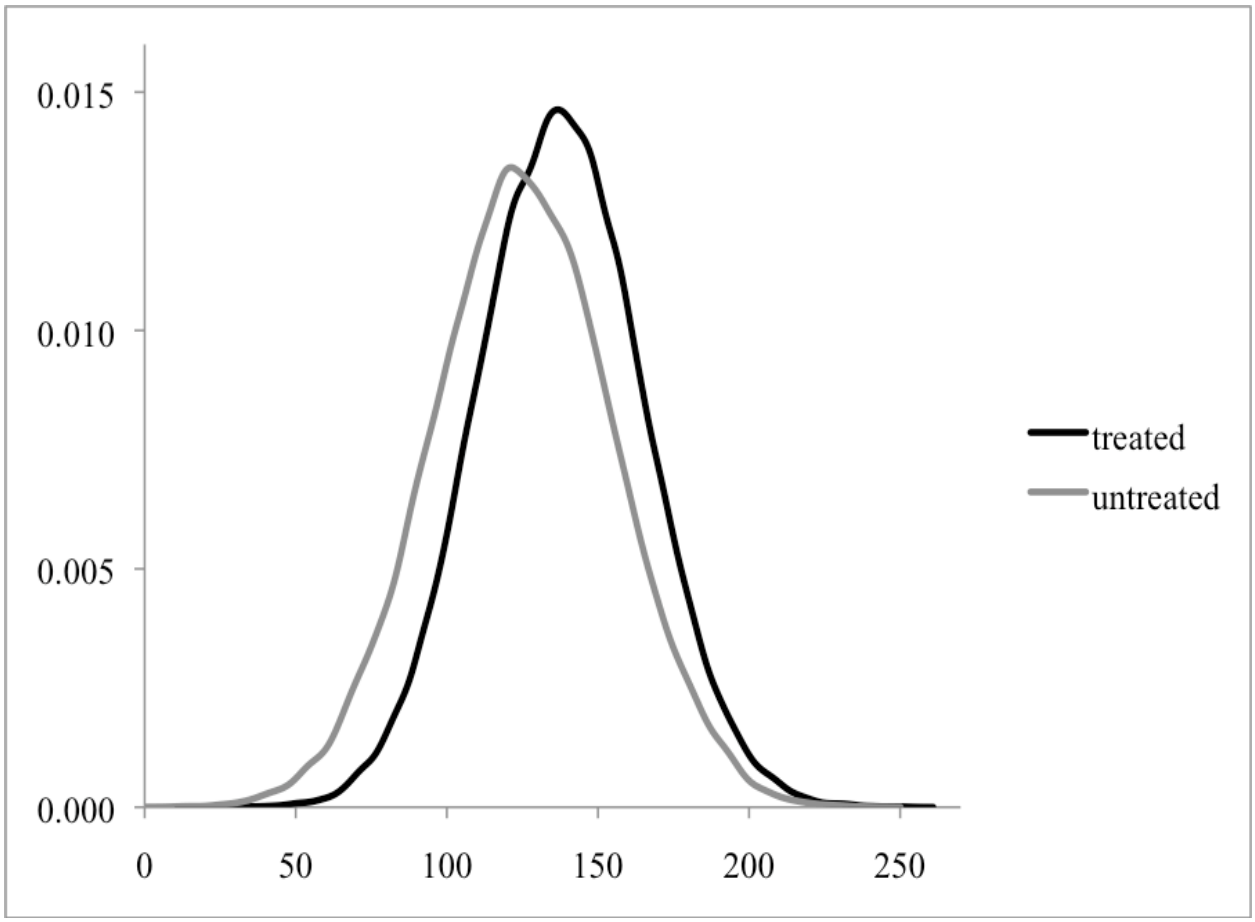

Fig. 3. Example comparison of treated and untreated groups for distribution of X_3 with n = 50,000.

Following creation of the control group, variable values are concatenated and then grouped into a dataframe object for analysis by the matchit function of the MatchIt library. This process is performed for all strata sizes being investigated, from 3 to 20. Although three strata are generally regarded as too few to significantly reduce bias (Drake, 1993), it was included in the simulations for completeness and to help visualize trends across strata sizes. The number of strata was also capped at 20, since even this number is considered beyond the range one might expect to continue to reap significant benefits in bias reduction in even the largest datasets.

The variables for the simulated participants are passed to the matchit function with all three variables (X_1, X_2, and X_3) as factors related to treatment selection. The matching method chosen is subclassification with the number of strata set according to the value of the loop

iterator (3 through 20) and since no calculation for propensity score is chosen, it defaults to logistic regression. A summary object is created from the outcome of the propensity score matching function with standardized difference scores calculated.

The next portions of the script simply compile the results obtained from the summary object. The first loop pulls out bias reduction scores along with a count of missing values. It also takes the current reduction value and places it into a weighted average of all iterations for each number of strata. This was done as a memory optimization strategy, as maintaining all iteration values for an entire simulation would quickly overrun the memory available of the computer being used for the simulations. If the current reduction score is not a valid value then the number of missing values is incremented rather than attempting to add it to the weighted average. This count is also used in the weighted average calculation itself to keep track of the number of previous valid values. The next loop extracts the standard bias numbers from the summary object in the same fashion, also placing them into the output table. Finally, the last loop of the function simply converts the missing values numbers into a percentage for easier direct interpretation. The full R script for the simulation function can be found in Appendix B.

6. Simulation Results

Results were very consistent across number of strata for all sample sizes with respect to bias reduction scores for the propensity score (see Fig. 4). At three strata, all sample sizes showed between a 70 to 75% reduction in bias as compared to before matching. Unlike previous literature stating that over 90% of bias should be removed with only 5 strata (Rosenbaum & Rubin, 1984), these simulations suggest that approximately nine strata were needed in order to reduce bias by 90% even at the smallest sample sizes used.

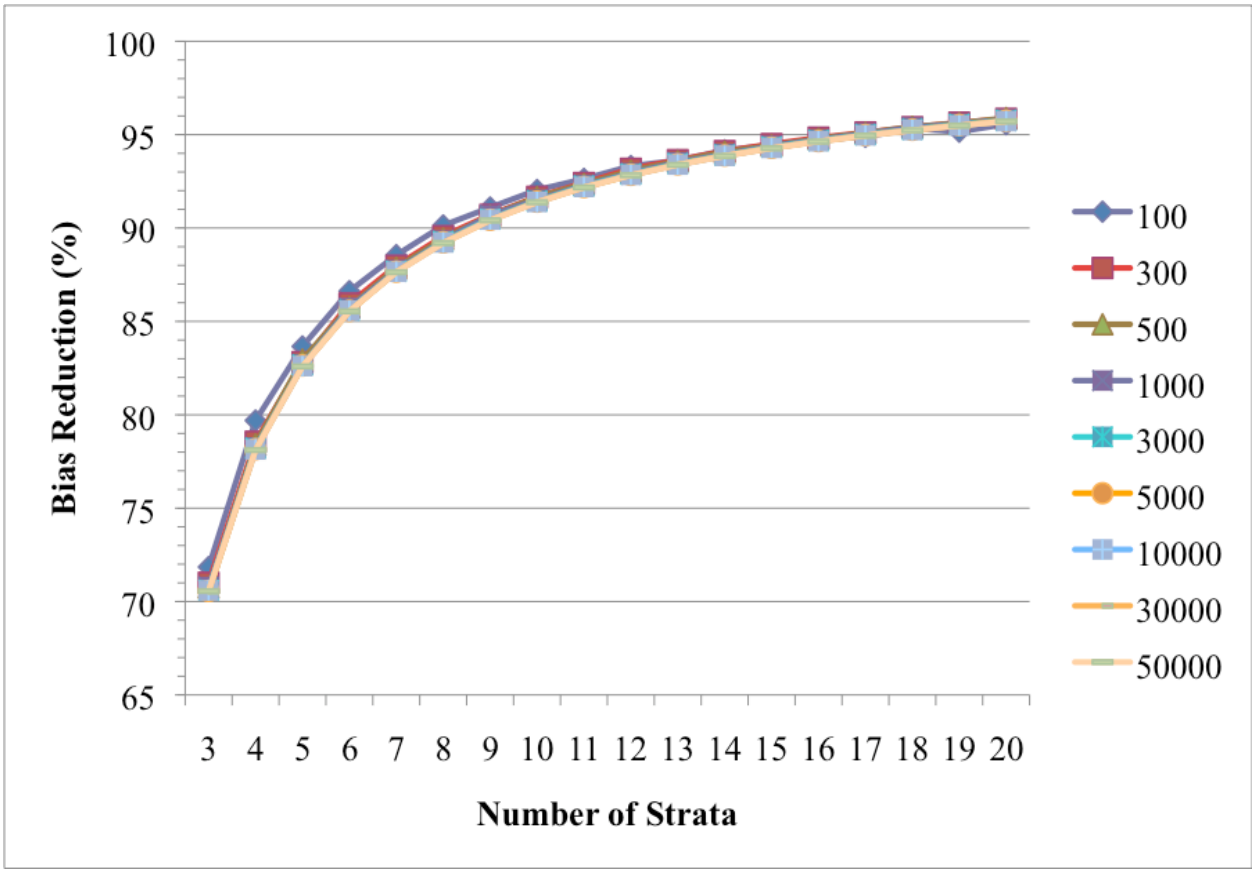

Fig. 4. Bias reduction for propensity score versus strata count.

As expected, the bias reduction scores for X_1 show no improvement (see Fig. 5). They are erratic due to the fact that X_1 played no role in treatment selection and should have therefore been randomly distributed simply by chance. The selection bias for X_1 should be minimal to begin with and any trend in improvement based on propensity score matching would be spurious.

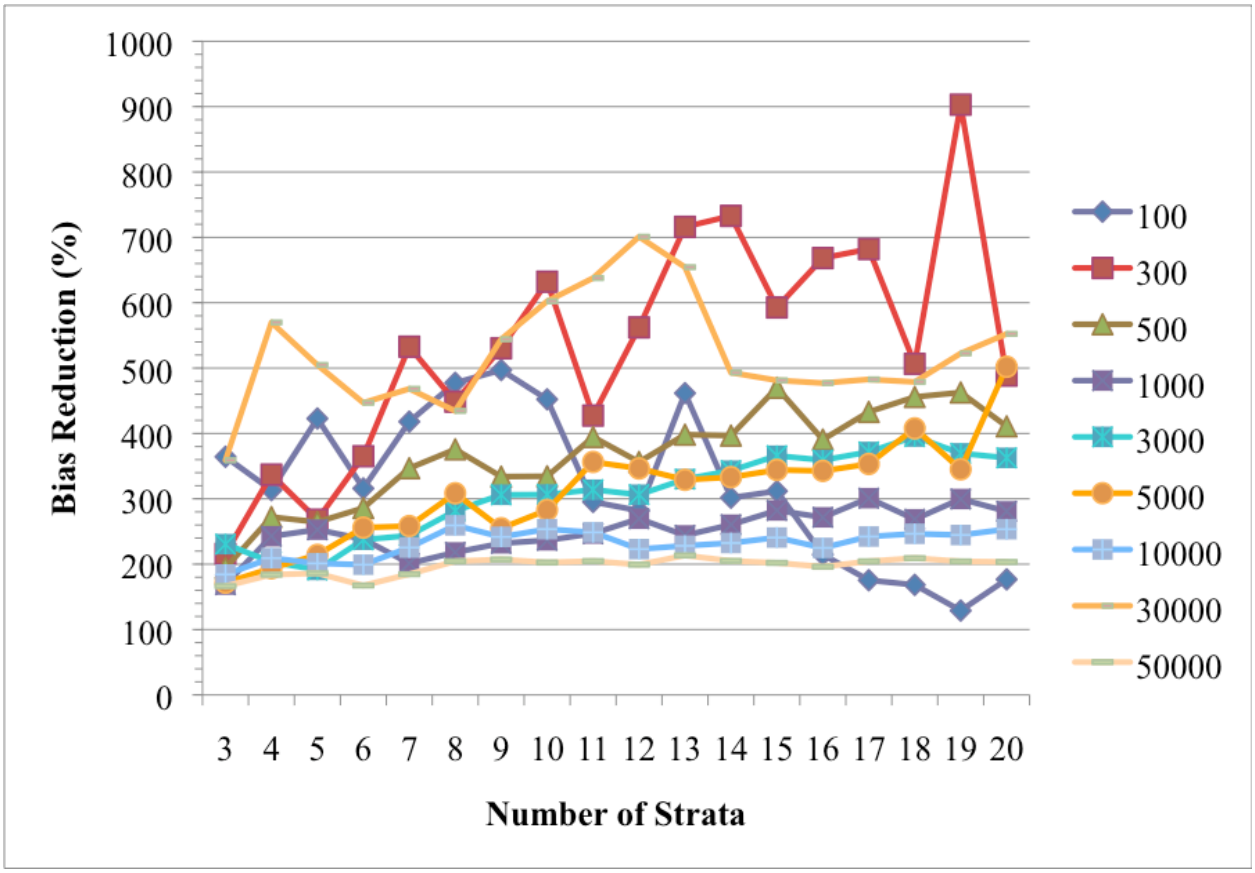

Fig. 5. Bias reduction for *X1* versus strata count.

In order to interpret the results of the bias reduction scores for both X_2 and X_3, it becomes necessary to take into account the missing data at the 100 participant sample size (see Figs. 6 and 7). Propensity score matching is not considered a small sample procedure, so it is not surprising that it is not entirely successful with samples as small as 100 treated participants. At six strata, there were 1.6% missing cases which increased to 17.6% by 10 strata and 60.5% at fifteen strata. Since these missing cases were omitted from the averages rather than being treated as zeros, the results were slightly erratic rather than being completely distorted. A trend can still be observed in X_2 that bias reduction improved more for the 100 and 300 sample sizes at lower numbers of strata and ceased to benefit as much with larger sample sizes at the higher number of strata.

Above n = 300, trend lines were almost identical and required 10 strata to break 90% bias reduction and showed tapering to 95% reduction by twenty strata.

In the case of bias reduction for X_3, the effect of the small sample size for n = 100 was much more apparent. Almost no trend can be determined for this set of simulations because of the erratic nature of the results. Moving past that sample size, the trend for sample sizes of 300 and above show that larger sample sizes again benefitted from a larger number of strata. Reduction for n = 300 plateaus at almost 89%, while n = 500 goes to 92% and n = 1000 to 94%. All higher sample sizes track almost identically across all number of strata and appear to still be slightly improving at twenty strata, where they show approximately 95% bias reduction.

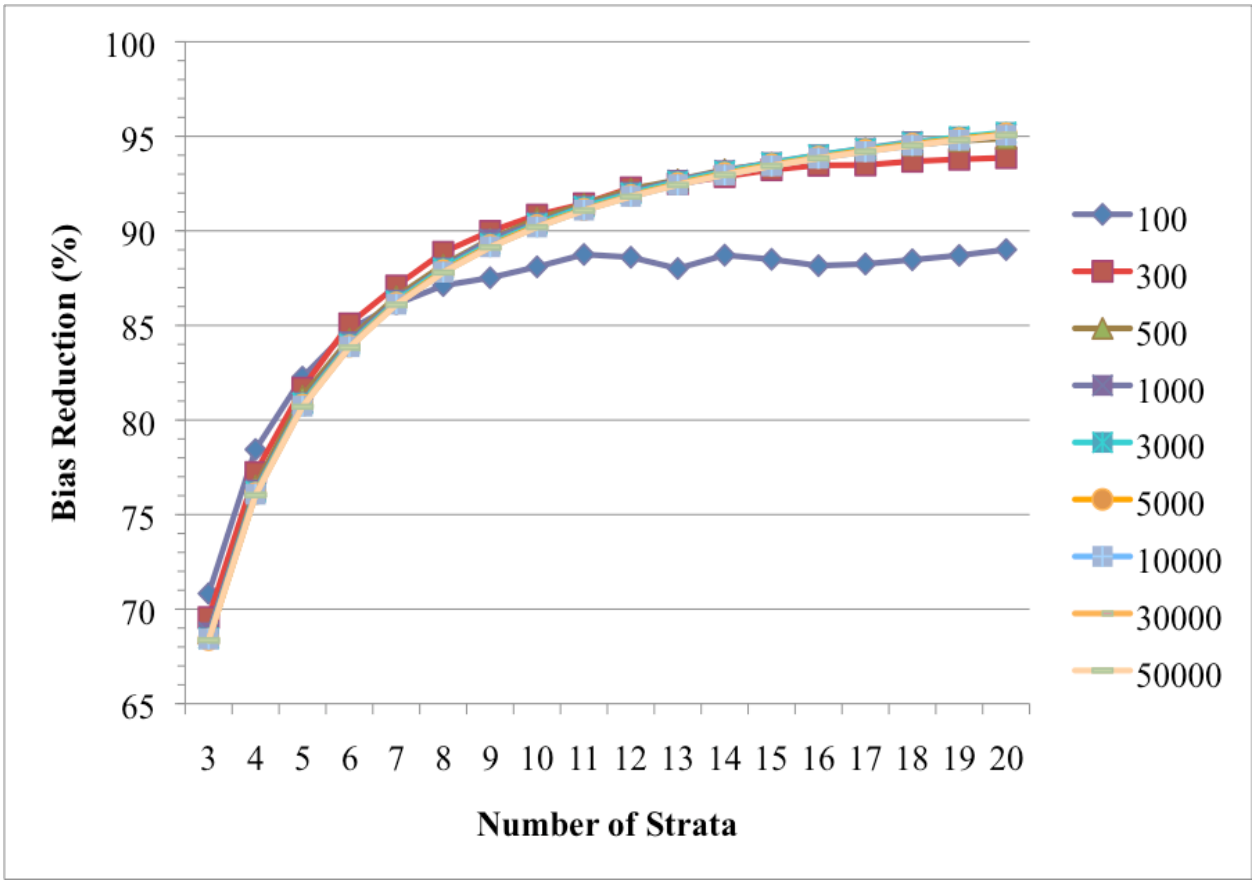

Fig. 6. Bias reduction for *X2* versus strata count.

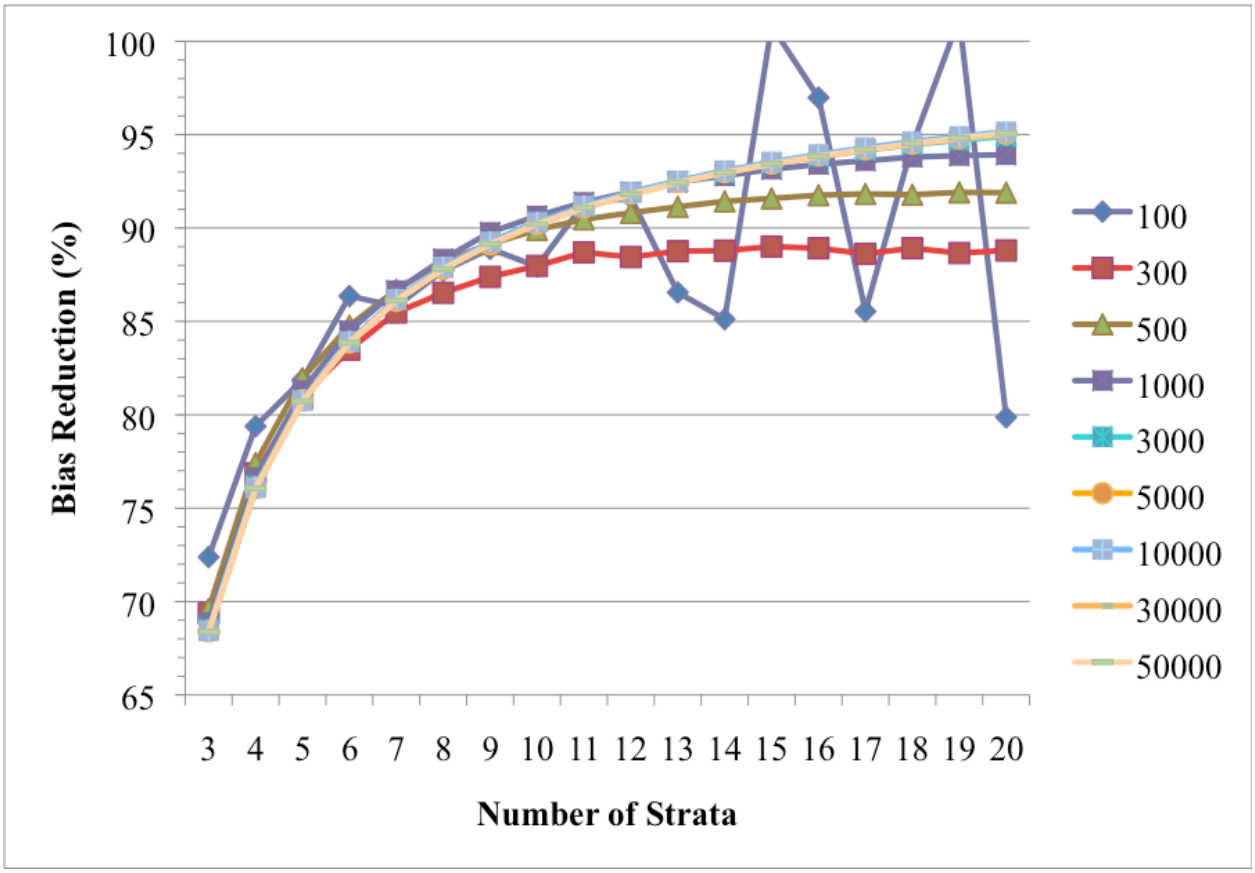

Fig. 7. Bias reduction for *X3* versus strata count.

Bias reduction scores are helpful for showing improvement with different matching procedures and with differing parameters, but are of little help in determining whether the bias for a given procedure is acceptable, since they rely on the initial unmatched bias. For determining if a given procedure yields a passable level of selection bias, the standardized bias scores are needed. These scores are also less prone to being erratic, especially with variables that are not highly related to treatment selection. The standardized bias scores in the case of the MatchIt package are the standardized mean differences between treated and untreated within each strata or subclassification. The raw differences are divided by the standard deviation of the treated strata, therefore the within group variability must remain high with regard to differences between treated and untreated to make these values sufficiently small. Researchers typically use

$d < 0.20$ or 0.25 as maximum allowed bias scores, but strive to get these numbers as close to zero as possible (Shadish & Steiner, 2010).

The standardized bias for the propensity score for all sample sizes was between 0.367 to 0.384 at three strata (see Fig. 8). All values of n except for $n = 100$ were similar across strata and decreased to values between 0.052 and 0.056 by twenty strata. The 100 participant sample followed the same general trend, but had slightly lower bias across all strata.

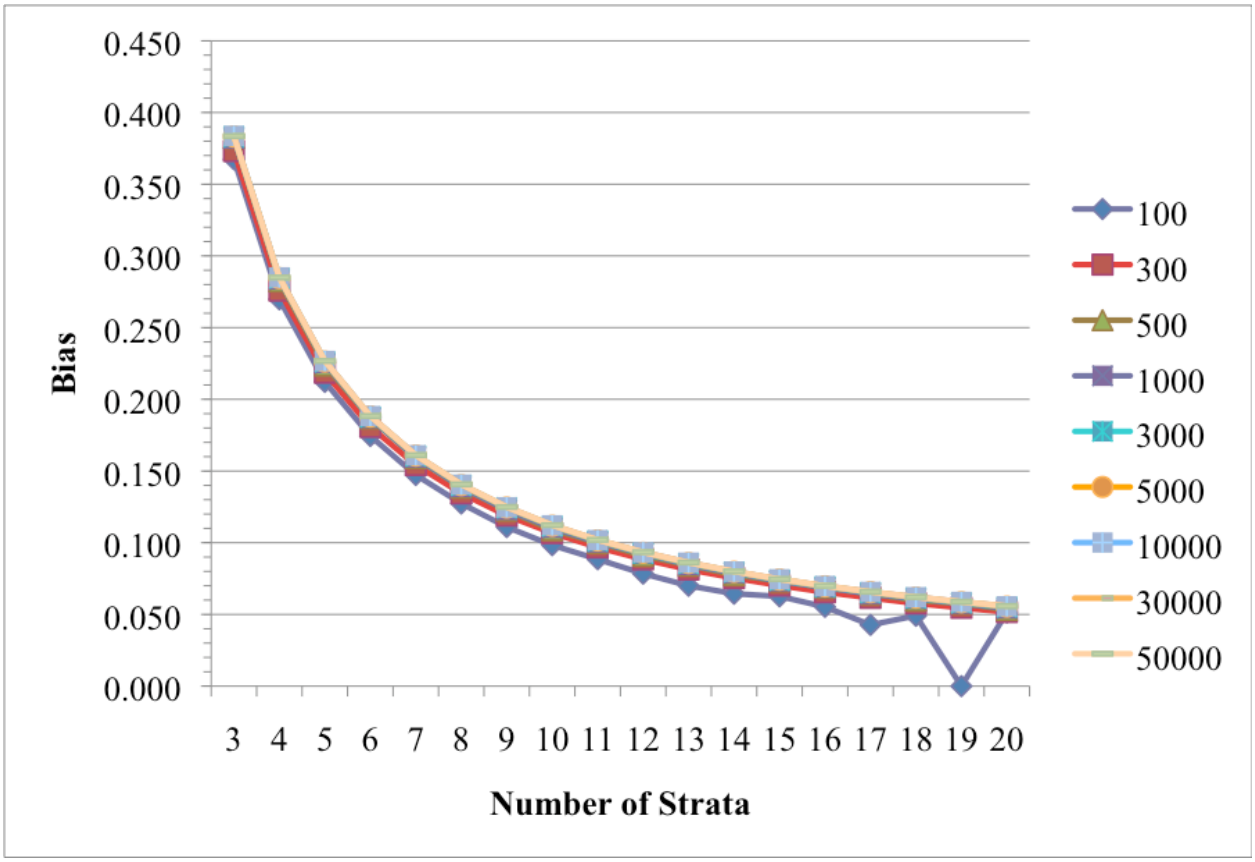

Fig. 8. Bias for propensity score versus strata count.

As expected, the bias scores for X_1 were largely unaffected by number of strata and were lower with larger sample sizes (see Fig. 9). The standardized bias scores increased slightly with increasing number of strata, not because the difference between treated and untreated groups increased, but because the variance within the groups became smaller while the difference

remained the same. As with the bias reduction scores, missing data played a large role in skewing the results at n = 100 due to the fact that by the 10 strata simulations over half of the iterations did not return values from the bias calculation. The missing values increased to over 99% at the higher levels with the nineteen strata simulations having no successful iterations at all. Other than these small sample limitations, the results are exactly what would be expected from a variable that played no role in treatment selection.

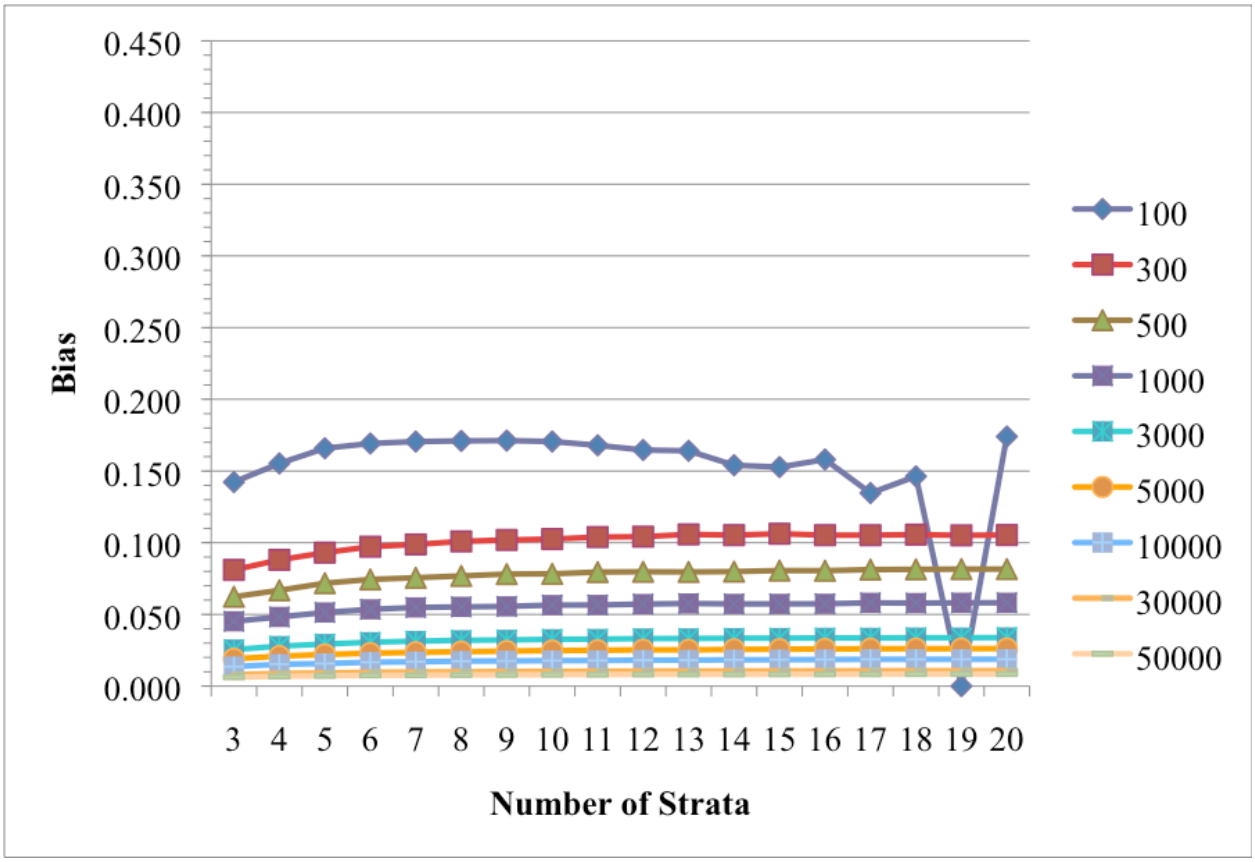

Fig. 9. Bias for *X1* versus strata count.

The bias score results for X_2 and X_3 were very similar, with the main difference being that X_2 had higher bias scores at lower numbers of strata but benefitted more as the number of strata increased (see Figs. 10 and 11). Some of the smaller sample sizes (up through n = 500) actually had lower standardized bias scores by the twenty strata level for the X_2 variable. Since X_2 played

a greater role in treatment selection, it would have had greater selection bias before matching, but also benefitted more from the matching since it is based on the propensity score to which it is more highly correlated.

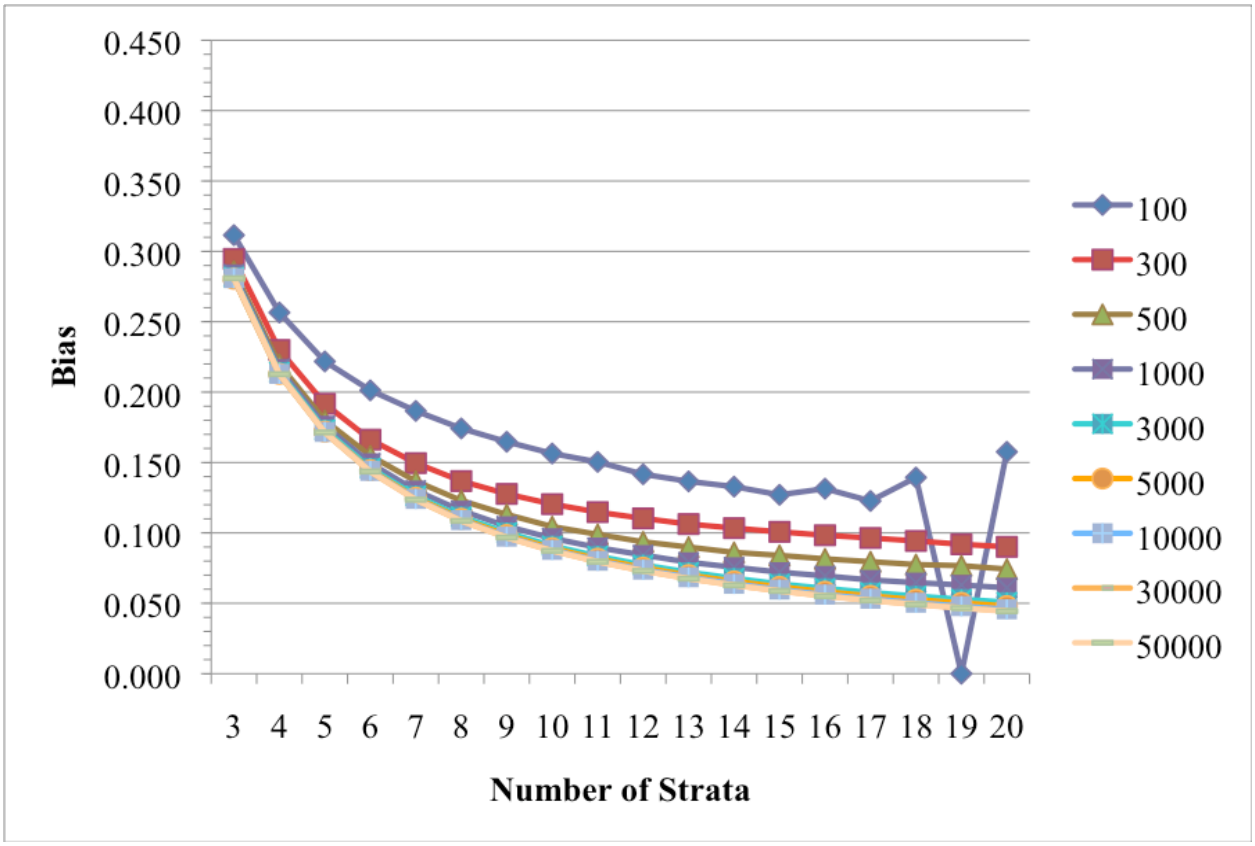

Fig. 10. Bias for *X2* versus strata count.

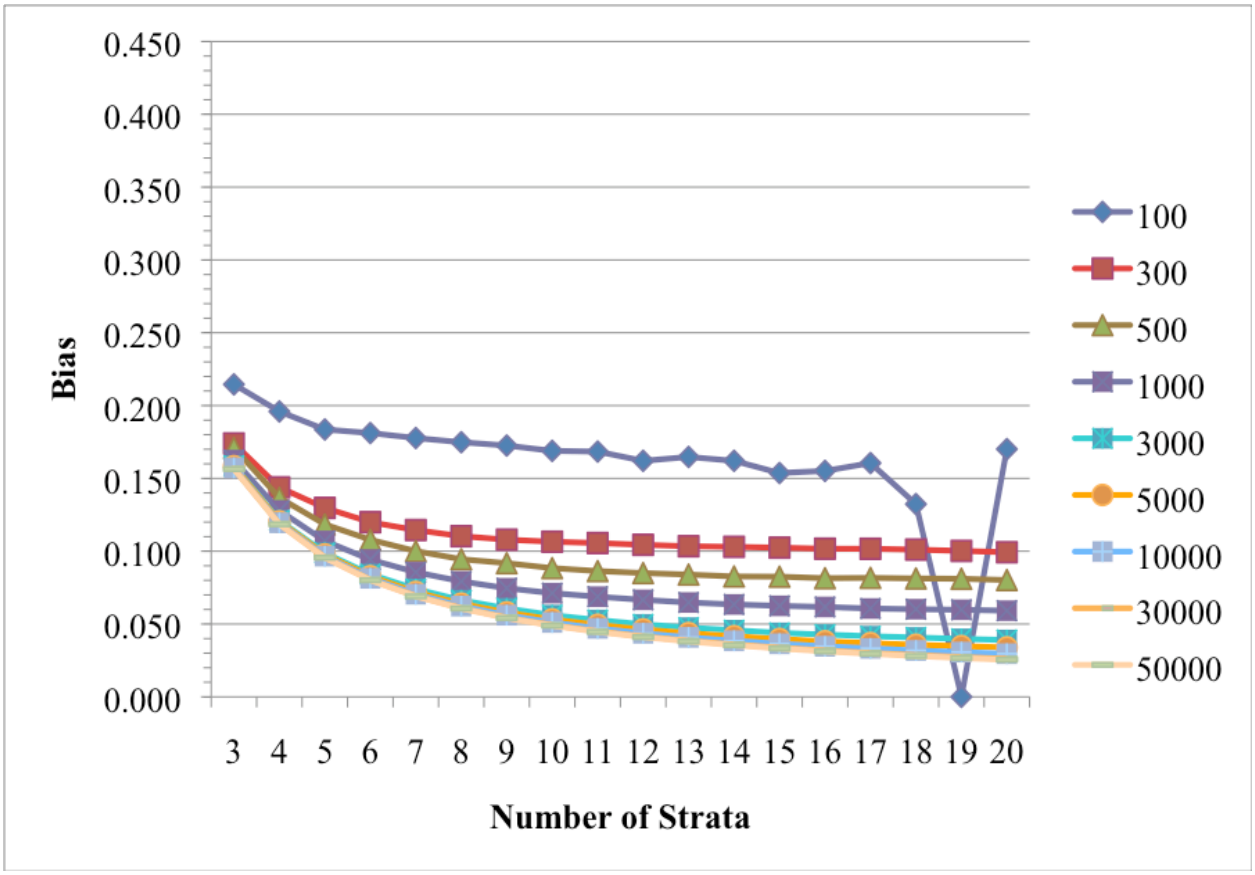

Fig. 11. Bias for *X3* versus strata count.

7. Discussion

Based on these simulations, it appears that sample sizes larger than somewhere between 1,000 and 3,000 participants in the treated group behave very similarly across varying numbers of strata. Full results for all sample sizes in table format can be found in Appendix C. Contrary to previous suggestions, based on these simulations and the calculations of the MatchIt package, it appears that even with smaller sample sizes that it requires at least eight strata to reduce 90% of the selection bias in a single variable (the propensity score) and that it may take even more to also reduce bias to this level in the component variables (X_2 and X_3).

It also appears that the addition of variables that are not related to treatment selection in the calculation of the propensity score simply adds unnecessary variance and may reduce the

amount of bias that may be removed from variables that actually are a factor in treatment selection. The bias level for these unrelated variables (which should be near zero to begin with) is affected by sample size, but not by the matching process itself in any meaningful way. The standardized bias actually increases with increasing number of strata due to the fact that raw bias is maintained while within group variability decreases. It would be more useful to remove these types of variables from the propensity score calculation, but still verify the bias related to them both before and after matching to insure that the matching process did not adversely affect what should already be a minimal selection bias. With the MatchIt package, this is accomplished automatically by adding the variable names in calls to the summary function in the parameter addlvariables. By using this parameter, any variable not used in the propensity score calculation will be included in the summary output with all corresponding balance measures just as if they were used in the PS model.

The functioning of variables that are significantly related to treatment selection varies based on the amount of influence. In the case of the present simulations, X_2 contributed 60% and X_3 contributed 40% to the treatment selection. In real life, it is generally not known exactly how much individual variables contribute to treatment selection. As an empirical guide, it is evident from the results that the more highly a variable affects treatment selection the greater bias it will have initially, but also the more it can be improved by the matching process and especially with greater number of strata in the case of stratification or subclassification.

In examining the bias results in terms of the standard bias, it took only a few strata to bring the propensity score as well as all variables below the $d < 0.20$ or 0.25 often used as determination for acceptable selection bias (Shadish & Steiner, 2010). The propensity score as well as both contributing variables continued to receive benefits in bias reduction up through

eight strata, at which point the bias improvement gains tended to taper, so the cutoff point appears to be based on the amount of bias reduction that is considered adequate for a given study. What is made clear from the results of this study is that beyond sample sizes of approximately 3,000 that increasing numbers of participants do not receive substantial benefit from larger numbers of strata.

8. Conclusion

As suspected in previous research (Lunceford & Davidian, 2004; Caliendo & Kopeinig, 2008; Rubin, 2010), the 5 strata standard that has been used in most studies (Austin, 2008; D'Agostino, 1998; Leon & Hedeker, 2007; Lunceford & Davidian, 2004; Störmer et al., 2006) may not be sufficient to optimally remove bias in all cases. The current study shows that there is definitely a trend for improvement in both bias reduction scores and standardized bias with larger strata sizes for both the propensity score itself and its component covariates, but that those differences are fairly similar at sample sizes of 3,000 and beyond. There are different standards by which acceptable bias has been judged, including greater than 90% bias reduction and $d < 0.20$ (Shadish & Steiner, 2010). This leads to three primary questions regarding selection of the optimal number of strata to be employed in propensity score stratification, based on these standards:

1. Are 5 strata really enough to remove 90% of the bias in the propensity score and covariates?
2. How many strata are required to reduce standard bias below 0.20?
3. Is the ultimate goal of stratification matching to achieve sufficient bias reduction or optimal bias reduction?

The following tables summarize the findings regarding strata requirements as they relate to sample size and criteria for adequate bias:

Table 1
Number of strata to reach criterion for propensity score.

	n								
Criterion	100	300	500	1000	3000	5000	10000	30000	50000
Reduction > 90%	9	9	9	9	9	9	9	9	9
$d < 0.20$	6	6	6	6	6	6	6	6	6

Table 2
Number of strata to reach criterion for covariates.

	n								
Criterion	100	300	500	1000	3000	5000	10000	30000	50000
Reduction > 90%	N/A	N/A	11	10	10	10	10	10	10
$d < 0.20$	7	5	5	5	5	5	5	5	5

8.1. Number of Strata Required to Remove 90% of the Bias

Consistently across all sample sizes, nine strata were required to remove 90% of the bias from the propensity scores. This is much larger than the 5 strata previously cited to be adequate for removing 90% of the bias in a single covariate. Even though the propensity score was calculated with the component covariates of X_1, X_2, and X_3, the process reduces those covariates to a single scalar which therefore constitutes a single covariate in the matching procedure. It is apparent and very consistent that almost double the number of strata than have been previously used in the literature are required to meet the criterion of 90% bias reduction for only the propensity score measure itself.

The number of strata required to reduce both of the functioning covariates comprising the propensity score (X_2 and X_3) greater than 90% was not available for sample sizes below 500, because the mean reduction never exceeded this mark for one of the covariates. At $n = 500$, it required 11 strata to achieve 90% reduction in bias in both variables. This may be a little misleading, because if the percentages were rounded to the nearest percent then X_3 would have

been 90% (89.900) reduction at 10 strata the same as X_2, which is consistent across all subsequent sample sizes. So, in summary, to reduce bias in the propensity score and all covariates at least 90% required approximately 10 strata.

8.2. Number of Strata Required for $d < 0.20$

The more practical criterion based on the standard mean difference or standardized bias (d) was much more easily attained for both the propensity score and the individual covariates. For the propensity score, six strata reduced standardized bias below the 0.20 criterion across all sample sizes. Five strata were required to reduce the functioning covariates (X_2 and X_3) for all sample sizes other than 100, which likely had greater initial bias due to the smaller number of participants. This type of result is probably what has led to propagation of the use of quintiles in research that has used $d < 0.20$ as a criterion for balancing covariates. This would be adequate if it weren't for the fact that the true goal is to reduce the bias as much as possible (Shadish & Steiner, 2010), with 0.20 being a mark of minimum sufficiency. Due to the fact that in the present simulation all confounding was known (because it was systematically programmed into the treatment selection), the bias reduction of the propensity score is likely more effective than it would be in a real world situation where all confounders cannot be fully reflected in the propensity score. Given that the researcher performing propensity score matching wants to reduce bias as much as possible, it seems valuable to note that beyond sample sizes of approximately 1,000 the sufficient standardized bias of 0.20 can be reduced in the functioning covariates by as much as 50% or more by increasing to 10 strata.

8.4. Meeting the Ultimate Goal

The following table shows the further bias reduction that was achieved with an increase from 5 to 10 strata. On average, bias declined by up to 0.115 for sample sizes greater than 300.

Although a difference of 0.115 may not necessarily affect the outcome of most research, it must give pause when examining any marginally significant findings for which only quintile stratification was used. This 0.115 standardized bias would likely be attributed to a standardized treatment effect, giving it over $1/10^{th}$ of a standard deviation of unwarranted treatment effect and possibly granting statistical significance especially with larger sample sizes. It would certainly seem remiss to neglect this amount of residual treatment selection bias given the ease by which it can be removed, simply by the addition of a few more strata in the matching procedure.

Table 3
Bias improvements with 10 strata.

	Propensity Score							
	n							
	300	500	1000	3000	5000	10000	30000	50000
5 strata	0.219	0.223	0.225	0.226	0.226	0.227	0.227	0.227
10 strata	0.107	0.109	0.11	0.112	0.112	0.112	0.112	0.112
difference	0.112	0.114	0.115	0.114	0.114	0.115	0.115	0.115
% improved	51	51	51	50	50	51	51	51

8.3. Recommendations for Future Research

Based on the findings of this study, I would recommend that in the same manner that alternate individual matching strategies are investigated on a particular set of data, that the same method be applied to stratification matching before selecting the number of strata. Trends across strata at any given sample size are very apparent and using a generalized R function to make a table summary for bias on a particular set of data would be much more useful in determining the optimal number of strata than using a particular number of strata just because it has been previously used on disparate sets of data in the literature. An interesting follow-up to this study would be to find large datasets with a rich and complete set of covariates and verify that the bias trend across strata is consistent with the simulation results. Also, assuming the dataset is

sufficiently large, it would be possible to randomly select subsamples of varying sizes to see if the sample size trends of the current simulation also hold against real world data.

Appendix 1. R code.

```
library(MatchIt)

psmSim <- function(iterations, n)
{
	rdxVals <- matrix(0,nrow=18, ncol=10)

	for (thisIteration in 1:iterations)
	{
		x1mean <- 50
		x1std <- 10
		x2mean <- 75
		x2std <- 20
		x3mean <- 125
		x3std <- 30

		x1 <- rnorm(n*2,mean=x1mean,sd=x1std)
		x2 <- rnorm(n*2,mean=x2mean,sd=x2std)
		x3 <- rnorm(n*2,mean=x3mean,sd=x3std)
		minX2 <- min(x2)
		rangeX2 <- max(x2) - minX2
		minX3 <- min(x3)
		rangeX3 <- max(x3) - minX3

		Tr <- ((x2-minX2)/rangeX2)*0.60 + ((x3-minX3)/rangeX3)*0.40
		TrMedian <- median(Tr)
		x1treated <- x1[Tr>=TrMedian]
		x2treated <- x2[Tr>=TrMedian]
		x3treated <- x3[Tr>=TrMedian]

		x1control <- rnorm(n,mean=x1mean,sd=x1std)
		x2control <- rnorm(n,mean=x2mean,sd=x2std)
		x3control <- rnorm(n,mean=x3mean,sd=x3std)

		x1 <- c(x1treated,x1control)
		x2 <- c(x2treated,x2control)
		x3 <- c(x3treated,x3control)
		Tr <- c(rep(1,length(x1treated)),rep(0,length(x1control)))

		myDF <- data.frame(x1,x2,x3,Tr)

		for (strataNum in 3:20)
		{
			myMatch <- matchit(Tr ~ x1 + x2 + x3, myDF, method="subclass",
subclass=strataNum)
```

```
                mySummary <- summary(myMatch, standardize=TRUE)

                for (impType in 1:4)
                {
                        thisReduction <- abs(mySummary$reduction[impType,1])
                        if (!is.na(thisReduction))
                        {
                                rdxVals[strataNum-2,impType] <- (rdxVals[strataNum-
2,impType]*(thisIteration-rdxVals[strataNum-2,5]-1) + thisReduction)/(thisIteration-
rdxVals[strataNum-2,5])
                        }
                        else if (impType == 1)
                        {
                                rdxVals[strataNum-2,5] <- rdxVals[strataNum-2,5] + 1
                        }
                }

                for (biasType in 6:9)
                {
                        thisBias <- abs(mySummary$sum.subclass[biasType-5,3])
                        if (!is.na(thisBias))
                        {
                                rdxVals[strataNum-2,biasType] <- (rdxVals[strataNum-
2,biasType]*(thisIteration-rdxVals[strataNum-2,10]-1) + thisBias)/(thisIteration-
rdxVals[strataNum-2,10])
                        }
                        else if (biasType == 6)
                        {
                                rdxVals[strataNum-2,10] <- rdxVals[strataNum-2,10] + 1
                        }
                }

        }

    }

    for (rowNum in 1:18)
    {
        rdxVals[rowNum,5] <- 100*(rdxVals[rowNum,5])/iterations
        rdxVals[rowNum,10] <- 100*(rdxVals[rowNum,10])/iterations
    }
    return(rdxVals)
}
```

Appendix 2. Tables of results.

100 Treated
Bias Reduction

Strata	PS	X_1	X2	X_3	% Missing
3	71.848	364.351	70.826	72.383	0.100
4	79.694	313.330	78.441	79.383	0.100
5	83.660	422.825	82.268	81.876	0.500
6	86.620	315.927	84.777	86.354	1.600
7	88.550	418.092	86.133	85.811	3.600
8	90.138	477.401	87.110	87.707	5.600
9	91.099	496.989	87.520	88.888	11.600
10	92.036	452.213	88.099	87.942	17.600
11	92.629	295.328	88.754	91.299	24.700
12	93.319	281.851	88.607	91.625	34.200
13	93.622	461.652	88.009	86.553	42.700
14	94.158	301.712	88.717	85.118	52.700
15	94.453	311.791	88.497	100.820	60.500
16	94.732	214.719	88.160	96.983	70.500
17	94.884	175.458	88.249	85.541	76.900
18	95.333	168.374	88.472	94.452	81.000
19	95.157	129.014	88.703	101.219	87.400
20	95.562	176.756	89.010	79.866	90.200

100 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.367	0.142	0.312	0.215	0.100
4	0.270	0.155	0.257	0.196	0.400
5	0.212	0.166	0.222	0.184	2.100
6	0.174	0.169	0.201	0.181	6.000
7	0.147	0.171	0.187	0.178	16.000
8	0.127	0.171	0.174	0.175	25.200
9	0.111	0.171	0.165	0.173	38.900
10	0.098	0.171	0.156	0.169	52.300
11	0.088	0.168	0.150	0.168	65.000
12	0.078	0.165	0.142	0.162	77.000
13	0.070	0.164	0.137	0.165	83.900
14	0.064	0.154	0.133	0.162	92.200
15	0.063	0.153	0.127	0.154	95.100
16	0.055	0.158	0.131	0.155	96.600
17	0.043	0.135	0.123	0.161	99.000
18	0.049	0.146	0.139	0.132	99.300
19	0.000	0.000	0.000	0.000	100.000
20	0.051	0.174	0.158	0.170	99.900

300 Treated
Bias Reduction

Strata	PS	X_1	X_2	X_3	% Missing
3	71.006	215.854	69.555	69.381	0.000
4	78.559	337.061	77.228	76.862	0.000
5	82.821	268.072	81.688	80.964	0.000
6	85.961	365.173	85.086	83.485	0.000
7	87.978	532.746	87.088	85.479	0.000
8	89.563	448.679	88.871	86.530	0.000
9	90.705	530.279	89.975	87.382	0.100
10	91.669	632.158	90.837	87.971	0.000
11	92.410	427.316	91.436	88.696	0.400
12	93.171	562.267	92.276	88.446	0.000
13	93.637	716.037	92.489	88.762	0.400
14	94.120	732.908	92.870	88.787	0.700
15	94.497	592.765	93.232	89.007	0.700
16	94.839	668.352	93.469	88.918	1.800
17	95.116	682.084	93.482	88.621	2.500
18	95.401	506.428	93.678	88.919	4.200
19	95.644	903.008	93.784	88.661	5.700
20	95.848	488.457	93.867	88.799	7.200

300 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.373	0.081	0.295	0.174	0.000
4	0.276	0.088	0.230	0.144	0.000
5	0.219	0.093	0.192	0.130	0.000
6	0.181	0.097	0.166	0.120	0.000
7	0.154	0.099	0.150	0.115	0.000
8	0.134	0.101	0.137	0.110	0.300
9	0.119	0.102	0.128	0.108	0.300
10	0.107	0.103	0.120	0.107	0.500
11	0.097	0.104	0.115	0.106	0.900
12	0.088	0.104	0.110	0.104	1.600
13	0.081	0.106	0.106	0.103	2.200
14	0.075	0.105	0.103	0.103	3.700
15	0.070	0.106	0.101	0.102	5.800
16	0.065	0.105	0.098	0.102	8.500
17	0.062	0.105	0.096	0.102	11.100
18	0.058	0.106	0.094	0.101	17.900
19	0.055	0.105	0.092	0.100	21.900
20	0.052	0.105	0.090	0.099	27.300

500 Treated
Bias Reduction

Strata	PS	X_1	X_2	X_3	% Missing
3	70.754	194.274	68.721	69.600	0.000
4	78.423	272.168	76.512	77.411	0.000
5	82.955	264.499	81.264	81.958	0.000
6	85.771	285.937	84.280	84.768	0.000
7	87.872	346.839	86.525	86.693	0.000
8	89.437	375.734	88.197	88.218	0.000
9	90.653	333.991	89.565	89.141	0.000
10	91.653	334.499	90.650	89.900	0.000
11	92.353	394.608	91.403	90.470	0.000
12	93.070	356.438	92.206	90.813	0.000
13	93.578	398.167	92.700	91.136	0.000
14	94.081	396.747	93.234	91.428	0.000
15	94.421	469.933	93.562	91.585	0.000
16	94.785	390.614	93.936	91.760	0.000
17	95.097	432.987	94.335	91.828	0.100
18	95.420	455.875	94.551	91.786	0.000
19	95.619	462.702	94.764	91.909	0.000
20	95.892	410.930	94.886	91.895	0.300

500 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.382	0.062	0.286	0.171	0.000
4	0.282	0.067	0.220	0.137	0.000
5	0.223	0.072	0.180	0.118	0.000
6	0.187	0.074	0.155	0.108	0.000
7	0.159	0.076	0.137	0.100	0.000
8	0.138	0.077	0.123	0.095	0.000
9	0.122	0.078	0.113	0.092	0.000
10	0.109	0.078	0.104	0.088	0.000
11	0.100	0.080	0.099	0.086	0.000
12	0.091	0.080	0.094	0.085	0.000
13	0.084	0.080	0.090	0.084	0.000
14	0.077	0.080	0.086	0.083	0.000
15	0.073	0.081	0.084	0.083	0.100
16	0.069	0.081	0.082	0.081	0.300
17	0.064	0.081	0.080	0.082	0.100
18	0.060	0.081	0.078	0.081	0.200
19	0.058	0.082	0.077	0.081	0.300
20	0.053	0.082	0.074	0.080	1.100

1,000 Treated
Bias Reduction

Strata	PS	X_1	X_2	X_3	% Missing
3	70.779	168.763	68.754	68.887	0.000
4	78.249	242.690	76.301	76.643	0.000
5	82.682	252.806	80.925	81.299	0.000
6	85.712	238.309	84.142	84.497	0.000
7	87.743	201.258	86.344	86.652	0.000
8	89.323	218.044	88.049	88.338	0.000
9	90.677	231.909	89.543	89.768	0.000
10	91.510	236.579	90.429	90.647	0.000
11	92.286	246.844	91.291	91.373	0.000
12	92.937	269.583	92.040	91.932	0.000
13	93.517	244.051	92.677	92.463	0.000
14	93.954	260.452	93.185	92.796	0.000
15	94.394	282.235	93.617	93.148	0.000
16	94.731	271.694	94.014	93.406	0.000
17	95.056	300.980	94.365	93.602	0.000
18	95.373	268.702	94.717	93.799	0.000
19	95.576	299.709	94.898	93.878	0.000
20	95.804	280.850	95.112	93.923	0.000

1,000 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.382	0.045	0.284	0.163	0.000
4	0.284	0.048	0.217	0.128	0.000
5	0.225	0.051	0.176	0.107	0.000
6	0.187	0.054	0.150	0.094	0.000
7	0.159	0.055	0.130	0.086	0.000
8	0.139	0.055	0.116	0.079	0.000
9	0.123	0.056	0.105	0.075	0.000
10	0.110	0.057	0.096	0.071	0.000
11	0.100	0.057	0.090	0.069	0.000
12	0.092	0.057	0.084	0.067	0.000
13	0.084	0.058	0.079	0.065	0.000
14	0.079	0.057	0.076	0.063	0.000
15	0.073	0.057	0.072	0.062	0.000
16	0.069	0.057	0.069	0.062	0.000
17	0.064	0.058	0.067	0.061	0.000
18	0.061	0.058	0.065	0.060	0.000
19	0.058	0.058	0.063	0.060	0.000
20	0.054	0.058	0.061	0.059	0.000

3,000 Treated
Bias Reduction

Strata	PS	X_1	X_2	X_3	% Missing
3	70.609	230.669	68.458	68.541	0.000
4	78.204	203.417	76.203	76.194	0.000
5	82.660	191.313	80.886	80.794	0.000
6	85.592	237.508	84.011	83.897	0.000
7	87.680	243.817	86.266	86.150	0.000
8	89.275	281.634	88.007	87.879	0.000
9	90.481	306.177	89.313	89.252	0.000
10	91.439	306.528	90.366	90.307	0.000
11	92.245	313.914	91.275	91.172	0.000
12	92.882	306.324	91.973	91.902	0.000
13	93.444	330.099	92.602	92.519	0.000
14	93.906	343.082	93.141	92.977	0.000
15	94.328	365.918	93.585	93.474	0.000
16	94.684	359.235	94.006	93.842	0.000
17	94.993	371.677	94.350	94.167	0.000
18	95.280	395.253	94.662	94.472	0.000
19	95.545	369.560	94.974	94.699	0.000
20	95.755	362.692	95.208	94.915	0.000

3,000 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.383	0.025	0.282	0.159	0.000
4	0.285	0.028	0.214	0.121	0.000
5	0.226	0.029	0.173	0.099	0.000
6	0.188	0.031	0.146	0.084	0.000
7	0.161	0.031	0.126	0.074	0.000
8	0.140	0.032	0.112	0.066	0.000
9	0.124	0.032	0.100	0.061	0.000
10	0.112	0.033	0.091	0.056	0.000
11	0.101	0.033	0.084	0.053	0.000
12	0.093	0.033	0.077	0.050	0.000
13	0.086	0.033	0.072	0.048	0.000
14	0.080	0.033	0.068	0.046	0.000
15	0.074	0.033	0.064	0.044	0.000
16	0.069	0.034	0.061	0.043	0.000
17	0.065	0.034	0.058	0.042	0.000
18	0.062	0.034	0.055	0.041	0.000
19	0.058	0.034	0.053	0.040	0.000
20	0.055	0.034	0.051	0.039	0.000

5,000 Treated
Bias Reduction

Strata	PS	X_1	X_2	X_3	% Missing
3	70.567	171.811	68.388	68.410	0.000
4	78.159	193.733	76.115	76.082	0.000
5	82.641	214.288	80.789	80.772	0.000
6	85.568	256.023	83.909	83.880	0.000
7	87.662	258.320	86.182	86.106	0.000
8	89.238	308.551	87.892	87.815	0.000
9	90.440	254.963	89.219	89.140	0.000
10	91.414	282.927	90.278	90.226	0.000
11	92.194	356.437	91.137	91.118	0.000
12	92.850	346.385	91.868	91.843	0.000
13	93.413	329.075	92.499	92.468	0.000
14	93.885	333.116	93.037	93.004	0.000
15	94.294	344.214	93.495	93.463	0.000
16	94.656	342.701	93.888	93.901	0.000
17	94.973	353.294	94.262	94.234	0.000
18	95.262	407.761	94.585	94.540	0.000
19	95.512	345.137	94.870	94.826	0.000
20	95.745	502.015	95.134	95.061	0.000

5,000 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.383	0.019	0.281	0.158	0.000
4	0.284	0.021	0.213	0.120	0.000
5	0.226	0.022	0.172	0.098	0.000
6	0.188	0.023	0.145	0.083	0.000
7	0.161	0.024	0.125	0.072	0.000
8	0.140	0.024	0.110	0.064	0.000
9	0.125	0.024	0.099	0.058	0.000
10	0.112	0.025	0.089	0.053	0.000
11	0.102	0.025	0.082	0.049	0.000
12	0.093	0.025	0.075	0.046	0.000
13	0.086	0.025	0.070	0.044	0.000
14	0.080	0.026	0.066	0.042	0.000
15	0.074	0.026	0.062	0.040	0.000
16	0.070	0.026	0.058	0.038	0.000
17	0.066	0.026	0.055	0.037	0.000
18	0.062	0.026	0.053	0.036	0.000
19	0.059	0.026	0.050	0.035	0.000
20	0.055	0.026	0.048	0.034	0.000

10,000 Treated
Bias Reduction

Strata	PS	X_1	X_2	X_3	% Missing
3	70.599	184.260	68.418	68.419	0.000
4	78.138	209.335	76.076	76.073	0.000
5	82.607	201.298	80.729	80.752	0.000
6	85.574	199.207	83.877	83.928	0.000
7	87.661	225.452	86.124	86.191	0.000
8	89.226	259.341	87.820	87.912	0.000
9	90.450	241.749	89.159	89.260	0.000
10	91.396	253.614	90.204	90.295	0.000
11	92.200	248.770	91.095	91.191	0.000
12	92.848	223.075	91.824	91.902	0.000
13	93.396	228.230	92.431	92.529	0.000
14	93.870	232.419	92.957	93.069	0.000
15	94.291	240.750	93.433	93.539	0.000
16	94.650	224.933	93.838	93.936	0.000
17	94.965	242.319	94.191	94.308	0.000
18	95.253	246.554	94.522	94.626	0.000
19	95.494	244.615	94.795	94.905	0.000
20	95.724	253.487	95.057	95.155	0.000

10,000 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.384	0.014	0.281	0.157	0.000
4	0.285	0.015	0.213	0.120	0.000
5	0.227	0.016	0.172	0.097	0.000
6	0.188	0.017	0.144	0.081	0.000
7	0.161	0.017	0.124	0.070	0.000
8	0.141	0.017	0.109	0.062	0.000
9	0.125	0.018	0.097	0.056	0.000
10	0.112	0.018	0.088	0.051	0.000
11	0.102	0.018	0.080	0.047	0.000
12	0.093	0.018	0.074	0.044	0.000
13	0.086	0.018	0.068	0.041	0.000
14	0.080	0.018	0.064	0.038	0.000
15	0.075	0.018	0.060	0.036	0.000
16	0.070	0.019	0.056	0.035	0.000
17	0.066	0.019	0.053	0.033	0.000
18	0.062	0.019	0.050	0.032	0.000
19	0.059	0.019	0.048	0.031	0.000
20	0.056	0.019	0.046	0.030	0.000

30,000 Treated
Bias Reduction

Strata	PS	X_1	X_2	X_3	% Missing
3	70.540	359.059	68.353	68.333	0.000
4	78.121	569.894	76.048	76.036	0.000
5	82.590	504.814	80.707	80.699	0.000
6	85.534	446.924	83.842	83.833	0.000
7	87.648	468.294	86.116	86.113	0.000
8	89.205	434.451	87.814	87.800	0.000
9	90.411	544.115	89.141	89.119	0.000
10	91.384	602.253	90.215	90.190	0.000
11	92.174	637.763	91.095	91.061	0.000
12	92.829	700.944	91.826	91.792	0.000
13	93.391	654.394	92.455	92.417	0.000
14	93.867	493.086	92.985	92.957	0.000
15	94.275	481.278	93.447	93.410	0.000
16	94.634	477.065	93.854	93.809	0.000
17	94.951	482.710	94.217	94.165	0.000
18	95.233	478.840	94.537	94.481	0.000
19	95.488	522.483	94.823	94.773	0.000
20	95.713	552.369	95.080	95.027	0.000

30,000 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.384	0.008	0.281	0.157	0.000
4	0.285	0.009	0.213	0.119	0.000
5	0.227	0.009	0.171	0.096	0.000
6	0.188	0.010	0.144	0.080	0.000
7	0.161	0.010	0.124	0.069	0.000
8	0.141	0.010	0.109	0.061	0.000
9	0.125	0.010	0.097	0.054	0.000
10	0.112	0.010	0.087	0.049	0.000
11	0.102	0.010	0.080	0.045	0.000
12	0.093	0.010	0.073	0.041	0.000
13	0.086	0.011	0.068	0.039	0.000
14	0.080	0.011	0.063	0.036	0.000
15	0.075	0.011	0.059	0.034	0.000
16	0.070	0.011	0.055	0.032	0.000
17	0.066	0.011	0.052	0.030	0.000
18	0.062	0.011	0.049	0.029	0.000
19	0.059	0.011	0.047	0.027	0.000
20	0.056	0.011	0.045	0.026	0.000

50,000 Treated
Bias Reduction

Strata	PS	X_1	X_2	X_3	% Missing
3	70.562	166.402	68.358	68.393	0.000
4	78.113	183.960	76.027	76.055	0.000
5	82.590	186.020	80.701	80.717	0.000
6	85.541	167.425	83.837	83.859	0.000
7	87.637	184.697	86.096	86.112	0.000
8	89.197	204.496	87.797	87.806	0.000
9	90.415	207.200	89.133	89.141	0.000
10	91.381	202.606	90.197	90.212	0.000
11	92.173	204.741	91.076	91.089	0.000
12	92.828	199.306	91.809	91.816	0.000
13	93.384	213.353	92.430	92.442	0.000
14	93.861	205.036	92.964	92.976	0.000
15	94.274	202.028	93.430	93.440	0.000
16	94.632	196.040	93.833	93.846	0.000
17	94.950	204.425	94.193	94.205	0.000
18	95.232	209.274	94.512	94.524	0.000
19	95.485	204.239	94.797	94.816	0.000
20	95.712	203.413	95.055	95.072	0.000

50,000 Treated
Standard Bias

Strata	PS	X_1	X_2	X_3	% Missing
3	0.384	0.006	0.281	0.156	0.000
4	0.285	0.007	0.213	0.118	0.000
5	0.227	0.007	0.171	0.096	0.000
6	0.188	0.008	0.144	0.080	0.000
7	0.161	0.008	0.124	0.069	0.000
8	0.141	0.008	0.109	0.061	0.000
9	0.125	0.008	0.097	0.054	0.000
10	0.112	0.008	0.087	0.049	0.000
11	0.102	0.008	0.080	0.045	0.000
12	0.093	0.008	0.073	0.041	0.000
13	0.086	0.008	0.068	0.038	0.000
14	0.080	0.008	0.063	0.035	0.000
15	0.075	0.008	0.059	0.033	0.000
16	0.070	0.008	0.055	0.031	0.000
17	0.066	0.008	0.052	0.030	0.000
18	0.062	0.008	0.049	0.028	0.000
19	0.059	0.008	0.047	0.027	0.000
20	0.056	0.008	0.044	0.025	0.000

APPENDIX

EXTENDED LITERATURE REVIEW

Research on Propensity Scores

There are numerous articles implementing propensity scores throughout the recent literature in the fields of medicine, social work, economics, epidemiology, education and psychology. Propensity scores were first introduced through a series of seminal articles on the topic by Rosenbaum and Rubin (1983, 1984). The propensity score relies on the assumption that differential assignment to treatment is strongly ignorable, meaning that no significant bias exists between treated groups and non-treated groups. As a formal definition, Rosenbaum and Rubin (1983) state that "Generally, we shall say treatment assignment is strongly ignorable given a vector of covariates v if

$$(r_1, r_0) \perp\!\!\!\perp z \mid v, \quad 0 < \mathrm{pr}(z = 1 \mid v) < 1$$

for all v."

Subsequent articles by Rosenbaum and Rubin expanded on the techniques surrounding the use of propensity scores, such as assessing sensitivity (1983) and removal of bias by subclassification on the propensity score (1984). The subclassification article stated that when trying to subclassify on individual covariates that "as the number of covariates increases, the number of subclasses grows exponentially." Because the propensity score reduces multiple covariates into a single scalar, it was determined that 90% of the bias could be removed with only 5 subclasses.

Rosenbaum and Rubin (1985) demonstrated the construction of a control group based on matching by the propensity score. Using the propensity score for matching greatly reduces the complexity of the matching procedure and also insures that fewer cases are lost due to incomplete matching. Direct matching also tends to decrease bias more than comparing subgroups, usually even with regression adjustments within groups. This article also introduced

the matching methods of nearest available matching, Mahalanobis metric matching, and nearest available Mahalanobis metric matching within calipers. The latter technique was found to be superior to the preceding techniques in bias reduction (given an appropriate caliper selection) and was less computationally intensive than the Mahalanobis metric matching alone, which required distances to be computed among all pairs in the data which in that particular study "required the computation of about 1.5 million Mahalanobis distances."

Rosenbaum and Rubin (1985b) elucidated the importance of proper matching methods and described the bias that can be introduced by either incomplete matching or inexact matching, which are generally diametrically opposed shortcomings based on matching procedures. The conclusion was that reducing (or, more preferably, completely removing) incomplete matching was more important and left "only a small residual bias due to inexact matching."

Propensity Scores Gain Acceptance

Other researchers were not aggressive in adopting the fairly new concept of propensity scores and there were few articles on the topic outside of further clarifications from Rosenbaum and Rubin until the early 1990s. Rubin and Thomas (1992) presented several theorems and calculations surrounding matching using linear propensity score methods that helped to further clarify the mathematical bases of the fledgling technique. Following publication of this article, there was a steady increase in articles on the propensity scores from a wider range of authors. One such article discussed the limitations of propensity score subclassification with small sample sizes (Drake, 1993), which concluded that there can be considerable residual bias if sample sizes are too small to allow for at least dividing data into quintiles based on the propensity score.

D'Agostino and Kwan (1995) brought propensity scores to the forefront of the medical research community in a conference presentation on propensity score matching and stratification

as viable alternatives to fully randomized experimental design. Application of matching and subclassification based on propensity scores were the focus of articles for the next several years (Rubin & Thomas, 1996; Smith, 1997; Rubin, 1997; Dehejia & Wahba, 1998). Smith (1997) built upon basic matching by showing how treated "participants" can be matched to multiple control "participants" in order to reduce bias when the nature of the investigation focuses on treatment effects in a small and rare minority that is difficult to exactly match, but for which there are a large number of controls. The matching is performed in the same manner as one to one matching (generally within calipers), except that multiple controls may possibly be matched to each treatment case, with or without replacement. Imbens (1999) expanded the propensity score methodology to allow for estimation of causal effects for multi-valued treatments rather than the two value (treated or untreated) model that had been used previously.

At this point in the history of the propensity score literature, the basis of propensity score use had been fairly fully developed. Researchers then turned to introducing the technique to new fields of study, creating specialized (or more generalized) applications of the techniques, or in many cases either defending or disparaging the results and inferences made using propensity scores.

Building on the Basics

The application of propensity scores to longitudinal data was investigated by Winship and Morgan (1999) (along with other methods relevant to estimating causal effects), especially as they apply to sociological studies. Greater bias reduction can be realized through "blocking" by matching on prognostic variables in addition to regression adjustment (Rubin & Thomas, 2000). The authors emphasized that the greatest results are achieved by using a full set of covariates rather than selecting only the covariates that account for the most variation.

Dehejia and Wahba (2002) demonstrated that propensity score matching can be used to match a relatively rare group of treated individuals without regard to outcome measures to a large and diverse population, thereby allowing outcomes to only be measured for the matching participants at a considerable cost savings to the researcher. The same idea was applied to the medical field when there are rare outcomes to a common treatment (Braitman & Rosenbaum, 2002). In this instance, other multivariate methods could not be used because the number of covariates greatly exceeds the number of rare outcomes, making model convergence impossible.

Several other articles utilized propensity score matching and showed their effectiveness by use of real data, simulations and through theoretical comparison to competing causal inference methods (Hirano, Imben, & Ridder, 2003; Zhao, 2004; Caliendo & Kopeinig, 2006; John, Wright, Duku, & Willms, 2008; Barth, Guo, & McCrae, 2008). Caliendo & Kopeinig (2006) noted that one area that still requires study is a guideline for the optimal number of strata when performing propensity score stratification matching. Rubin (2004) advocated teaching statistical inference for causal effects to all graduate and undergraduate statistics students and provided a framework for developing a curriculum and Luellen, Shadish, and Clark (2005) presented an introduction to propensity scores and compared results from propensity score analysis to a randomized experiment. They also detailed how to use classification tree analysis and bagging for "researchers interested in computing propensity scores using more complex classification algorithms known as ensemble methods."

Dose-Response Extension

Another popular topic was an extension to the binary treatment levels inherent in the original concept of propensity score analysis to include multiple levels of treatment and to quantify differential levels of dose-reponse (Imbens, 2000; Foster, 2003; Imai & Van Dyk,

2004). Imbens (2000) first introduced the method to allow dose-response modeling by expanding on the generalizations of Joffe and Rosenbaum (1999). Imai and Van Dyk then used these methods to demonstrate that with this technique "bias and error reduction is relatively robust to model misspecification."

REFERENCES

Austin, P. (2008). The performance of different propensity-score methods for estimating relative risks. Journal of Epidemiology 61, 537-545.

Austin, P., Grootendorst, P., Normand, S. T., & Anderson, G. M. (2007a). 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, 734-753.

Austin, P., Grootendorst, P., Normand, S. T., & Anderson, G. M. (2007b). 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. doi: 10.1002/sim.2618

Barth, R. P., Guo, S., & McCrae, J. S. (2008). Propensity score matching strategies for evaluating the success of child and family service programs. Research on Social Work Practice 18 (3), 212-222.

Braitman, L. E., & Rosenbaum, P. R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores. Annals of Internal Medicine 137 (8), 693-695.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. American Journal of Epidemiology 163, 1149-1156.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys 22 (1), 31-72.

D'Agostino Jr., R. B. (2007). Discussion of : statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. Journal of Biopharmaceutical Statistics 17, 29-33. doi: 10.1080/10543400601050276

D'Agostino Jr., R. B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine 17, 2265-2281.

D'Agostino Jr., R. B., & Kwan, H. (1995). Measuring effectiveness: What to expect without a randomized control group. Medical Care 33 (4), AS95-AS105.

D'Agostino Jr., R. B., Lang, W., Walkup, M., & Morgan, T. (2001). Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood (SMBG). Health Services & Outcomes Research Methodology 2, 291-315.

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

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. Review of Economics and Statistics 84 (1), 151-161.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 49 (4), 1231-1236.

Foster, E. M. (2003). Propensity score matching: an illustrative analysis of dose response. Medical Care,41 (10), 1183-1192.

Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. Evaluation Review 32 (4), 392-409. doi: 10.1177/0193841X08317586

Haviland, A., Nagin, D. S., Rosenbaum, P. R., & Tremblay, R. E., (2008). Combining group-based trajectory modeling and propensity score matching for casual inferences in nonexperimental longitudinal data. Developmental Psychology 44 (2), 422-436. doi: 10. 1037/0012-1649.44.2.422

Hirano, K., Imbens, G. W., & Ridder G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161-1189.

Imai, K., & Van Dyke, D. A. (2004). Causal inference with general treatment regimes: generalizing the propensity score. Journal of the American Statistical Association 99 (467), 854-866.

Imbens, G. W. (2000). The role of propensity score in estimating dose-response functions. Biometrika 87 (3), 706-710.

John, L., Wright, R., Duku, E. K., & Willms, J. D. (2008). The use of propensity scores as a matching strategy. Research on Social Work Practice 18 (1), 20-26.

Jung, S., Chow, S., & Chi, E. M. (2007). A note on sample size calculation based on propensity analysis in nonrandomized trials. Journal of Biopharmaceutical Statistics 17, 35-41.

Leon, A. C., & Hedeker, D. (2007). Quintile stratification based on a misspecified propensity score in longitudinal treatment effectiveness analyses of ordinal doses. Computational Statistics & Data Analysis 51, 6114-6122.

Luellen, J. K., Shandish, W. R., & Clark, M. H. (2005). Propensity scores: an introduction and experimental test. Evaluation Review 29, 530-558.

Lunceford, J.K., & Davidian M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine 23, 2937-2960.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. Statistical Science 17 (3), 286-304.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society 45 (2), 212-218.

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

Rosenbaum, P. R., & Rubin, D. B. (1985a). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. American Statistician 39 (1), 33-38.

Rosenbaum, P. R., & Rubin, D. B. (1985b). The bias due to incomplete matching. Biometrics 41 (1), 103-116.

Rubin, D. B. (1980). Bias reduction using Mahalanobis metric matching. Biometrics 36, 293-298.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics 47 (4), 1213-1234.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine 127 (8), 757-764.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. Journal of Educational and Behavioral Statistics 29 (3), 343-367.

Rubin, D. B. (2010). Propensity score methods. American Journal of Ophthalmology 149 (1), 7-9.

Rubin, D. B., & Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. Biometrika 79 (4), 797-809.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. Biometrics 52 (1), 249-264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association 95 (450), 573-585.

Schafer, J. L., & Kang, J. (2008). Average casual effects from nonrandomized studies: A practical guide and simulated example. Psychological Methods 13 (4), 279-313. doi: 10.1037/a0014268

Seeger, J. D., Kurth, T., & Walker, A. M. (2007). Use of propensity score technique to account for exposure-related covariates. Medical Care 45 (10) Suppl 2, S143-S148.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. Journal of the American Statistical Association 103 (484), 1334-1356. doi: 10.1198/016214508000000733

Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn & Infant Nursing Reviews* 10 (1). 19-26.

Shah, B. R., Laupacis, A., Hux, J. E., & Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. Journal of Clinical Epidemiology 58, 550-559. doi: 10.1016/j.jclinepi.2004.10.016

Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. Sociological Methodology 27, 325-353.

Smith, J. A., & Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. American Economic Review 91 (2), 112-118.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. Journal of Clinical Epidemiology 59 (5), 437-447. doi: 10.1016/j.jclinepi.2005.07.004

Stürmer, T., Schneeweiss, S., Rothman, K. J., Avorn, J., & Glynn, R. J. (2007). Performance of propensity score calibration- a simulation study. American Journal of Epidemiology 165 (10):1110-1118; doi:10.1093/aje/kwm074

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. Annual Review of Sociology 25, 659-706.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. Review of Economics and Statistics 86 (1), 91-107.