ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2017-58633

Reducing bias in Observational Studies: An Empirical Comparison of Propensity Score Matching Methods

Gözlemsel Çalışmalarda Yanlılığı Azaltma: Eğilim Skoru Eşleştirme Yöntemlerinin Deneysel Karşılaştırılması

Lateef Babatunde AMUSA[a]

[a]Department of Statistics,
University of Ilorin,
Niger

Received: 26.10.2017
Received in revised form: 30.01.2018
Accepted: 30.01.2018
Available online: 05.04.2018

Correspondence:
Lateef Babatunde AMUSA
University of Ilorin,
Department of Statistics,
NIGER/NİJER
amusasuxes@gmail.com

ABSTRACT Objective: This paper aims to compare the performance of four widely used propensity score matching (PSM) methods, namely; Nearest neighbor matching, Caliper matching, Mahalanobis metric matching including the propensity score, and Stratification matching, in terms of bias reduction on observational data from which the treatment effects are intended to be assessed. **Material and Methods:** The selection bias, standardized bias and percent bias reduction are evaluated for each of the PSM methods using empirical data drawn from the Nigeria Demographic Health Survey of 2013. Factors that are associated with Ideal family size determination were extracted. The women were then divided into two groups: those who have at least a secondary school education, subsequently regarded as 'treated' group, and those who have no form of formal education, regarded as 'control group. **Results:** The balance metrics adopted showed a high level of imbalance between the two groups of interest for the unmatched data. Caliper matching was shown to have outperformed the other three methods in the task of bias reduction and achieving balance between the two treatment groups. **Conclusion:** Results from this study can help medical and health researchers to choose appropriate propensity matching methods to estimate treatment effect in the presence of confounding variables.

Keywords: Propensity score matching; imbalance; bias reduction; observational data

ÖZET Amaç: Bu makalenin amacı yaygın olarak kullanılan dört eğilim skoru eşleştirme (ESE) yöntemi olan En yakın komşu eşleştirmesi, Caliper eşleştirmesi, eğilim skorunu içeren Mahalanobis metrik eşleştirmesi ve Tabakalı eşleştirme yöntemlerinin performansını deneysel etkilerin değerlendirilmesi amaçlanan gözlemsel verideki yanlılığı azaltma bakımından karşılaştırmaktır. **Gereç ve Yöntemler:** 2013'te yapılan Nijerya Demografik Sağlık Araştırması'ndan elde edilen deneysel veriler kullanılarak her bir ESE yöntemi için seçim yanlılığı, standartlaştırılmış yanlılık ve yüzde yanlılık azaltma değerleri elde edilmiştir. İdeal aile büyüklüğünün belirlenmesi ile ilgili faktörler elde edilmiştir. Daha sonra kadınlar iki gruba ayrılmıştır: en az ortaöğretim mezunu olanlar tedavi grubunda, hiç resmi eğitim almayanlar ise kontrol grubunda yer almıştır. **Bulgular:** Kabul edilen denge metrikleri eşleştirilmemiş iki grup arasında dengesizlik olduğunu göstermiştir. Caliper eşleştirmesi, iki tedavi grubu arasında dengeyi sağlama ve yanlılığı azaltma bakımından diğer üç yönteme göre çok daha iyi sonuç vermiştir. **Sonuç:** Bu çalışmadan elde edilen sonuçlar karıştırıcı değişkenlerin varlığında tedavi etkisini tahmin etmek için uygun eğilim skoru eşleştirme yönteminin seçiminde tıp ve sağlık alanındaki araştırmacılara yardım edebilir.

Anahtar Kelimeler: Eğilim skoru eşleştirme; dengesizlik; yanlılık azaltma; gözlemsel veri

Observational investigations are progressively being used to evaluate the causal impacts of treatments and interventions on health and general wellbeing of individuals. In randomized experiments, randomization gives the expectation that subjects in the treatment and control groups are similar in both measured and unmeasured baseline attributes. However, in observational studies, treatment assignment is affected by subject characteristics. As a result, treated subjects often differ systematically from the untreated subjects. This concept has been referred to as treatment-selection bias in the literature.[1] Scientific researchers regularly need to utilize observational data to assess treatment effects in light of the fact that experimental designs or randomized control trials are frequently infeasible.[2]

As mentioned above, discrepancies emanating from observed and unobserved covariates between groups to be compared in observational studies produce biased results. The notion of bias here shows the systematic differences between treatment and control groups with respect to one or more covariates. There are two analytical ways to remove this bias: They are; (1) those that model the response variable, that is, they focus on the relationship between covariates and outcomes through Regression models; and (2) those that model the treatment assignment with respect to observed covariates using Propensity scores which mimic random assignment of experimental designs. The propensity score is defined as the probability of receiving the treatment (compared with the control exposure) conditional on a subject's observed baseline covariates.[3,4] Rosenbaum and Rubin demonstrated that the propensity score is a balancing score: conditional on the propensity score, treatment received is independent of measured baseline characteristics.[3] Thus, treated and untreated subjects with the same propensity score will have the same distribution of measured baseline covariates.

Although logistic regression is the most commonly used method for estimating the propensity score, probit models, discriminant analysis, and more recently, bagging or boosting, recursive partitioning or tree-based methods, random forests (Lee et al., 2010), and neural networks for estimating the propensity score have been adopted.[5-7]

Matching has generally become a popular approach to estimating causal effects, the most frequently used are the matching methods based on propensity score. Propensity score matching (PSM) methods have been widely used in medical and health researches to reduce the effect of confounding when estimating treatment effects using observational data. The availability of quite a number of PSM methods makes it difficult for researchers to choose an appropriate PSM method among the seemingly similar but different approaches.[8] The need for more information related to the use and systematic comparison of PSM in medical and health research motivates this current study.

With regards to the aforementioned concerns, this study aims to guide medical and health researchers on the adoption of PSM methods for improving the validity of the assessment of treatment effects. Rosenbaum and Rubin (1985) suggested that empirical results of matching might provide further information in practice; empirically comparing the various PSM methods and evaluating their effectiveness in reducing the selection bias would aid researchers in understanding and adopting PSM in practice.[9] Four (4) matching methods that depend on the propensity score, in terms of their bias reduction ability are compared. The four matching methods are, nearest neighbor matching, caliper matching, Mahalanobis metric matching including the Propensity score, and Stratification matching. Although, there are much more than these four Propensity score matching methods, these are the basic, most commonly used. Specifically, the selection bias, standardized bias and percent bias reduction are evaluated for each of the PSM methods using empirical data drawn from the Nigeria Demographic Health Survey of 2013 (NDHS, 2013). It is worthy of note that as a common practice of most PSM applications, this study used a large data set from a national database to increase representativeness, and ascertain the reliability of the methodological comparison results.[10]

MATERIAL AND METHODS

The data used in this study were extracted from the Nigeria Demographic Health Survey of 2013. The Demographic and Health Survey (DHS) is conducted in many developing countries by Measure DHS (www.measuredhs.com) to provide cross-sectional information on demographic and health indicators, including information on fertility and family planning, knowledge and current use of contraception methods, as well as sexually transmitted diseases.[11] The survey is designed to provide this information at national, regional, and state or district levels, for both urban and rural areas. The 2013 NDHS data comprises of 38948 respondents, with female respondents being within the age range of 15-49 years.[11] In line with the study's aim, 26403 of them who are married were selected.

For the purpose of this study, it is of interest to examine the effect of education on Ideal family size in the presence of confounding variables. As a result, a host of factors such as social, economic, cultural, demographic and environmental factors (later referred to as covariates) that associate with Ideal family size determination were extracted from the data. Covariates extracted were: age (years), type of residence (rural= 1, urban = 0), sex of household head (male= 1, female= 0), age at first birth (years), number of siblings, husband educational status (educated= 1, uneducated= 0), working status (working= 1, not working= 0), husband's age (years), and interval of marriage to first birth (months). Also, the data were divided into 2 groups of women who have at least a secondary school education, subsequently regarded as 'treated' group, and those who have no form of formal education, regarded as 'control' group. After missing data analysis, by exclusion of observations with incomplete cases on the covariates, the sample size reduced to 24222. Because we have a sufficiently large sample size, this study adopts sampling without replacement in implementing each of the four PSM methods.

ESTIMATION OF PROPENSITY SCORES

The propensity score was defined by Rosenbaum and Rubin (1983a) to be the probability of treatment assignment given the observed baseline covariates:[3]

Let Z be treatment assignment indicator, X be the observed baseline covariates, and e be the propensity scores.

$$e(x_i) = P(Z_i = 1|X_i),\ i = 1, \ldots, n \qquad (1)$$

Where it is assumed that, given the X's, the Z_is are independent:

$$P(Z_1 = z_1 \ldots Z_N = z_n) = \prod_{i=1}^{N} e(x_i)^{z_i} \{1 - e(x_i)\}^{1-z_i} \qquad (2)$$

The propensity score, specifically for this study, is the predicted probability of being 'educated', estimated from a logistic regression using respondent's educational status (Educated vs. Uneducated) as the dependent variable and the 9 covariates as the predictors.

PSM METHODS

The following four (4) commonly used PSM methods were compared in this current study using procedures from R package "MatchIt' developed by Ho et al (2011).[12]

Let π_i and π_j be the propensity scores for educated and uneducated respondents respectively, I be the set of educated respondents, and J is the set of uneducated respondents.

Nearest Neighbour Matching

A neighborhood C (π_j) contains the group of uneducated respondents, j (i.e., j ∈ J) as a match for educated respondents i (i.e., i ∈ I), if the absolute difference of propensity scores is the smallest among all possible pairs of propensity scores between i and j, as:

$$C\,(\pi_i) = \min \|\pi_i - \pi_j\|, \, j \in J \tag{3}$$

Once a j is found to match to i, j is removed from J without replacement. If for each i there is only a single j found to fall into C (π_i), then the matching is nearest neighbor pair matching or 1-to-1 matching.

Caliper Matching

As opposed to nearest neighbor matching, Caliper matching provides a restriction imposed on the distance between π_i and π_j, and as such, j is selected as a match for i, only if the absolute difference of propensity scores between the two group of respondents meets the following condition:

$$\|\pi_i - \pi_j\| < \xi, \, j \in J \tag{4}$$

Where ξ is a caliper or a pre-specified tolerance for matching. Rosenbaum and Rubin (1985) suggested using a caliper size of a quarter of propensity scores standard deviation[9].

Mahalanobis Metric Matching Including the Propensity Score

A common matching technique is Mahalanobis metric matching using several background covariates. It matches each case i in the treated (educated) group with a case j in the control (uneducated) group with the closest Mahalanobis distance, defined by:

$$d(i,j) = (x - y)^T \, C^{-1} \, (x - y) \tag{5}$$

where x and y are values of the matching variables for educated respondents i and uneducated respondents j, and C is the covariance matrix of the matching variables from the full set of the controls. The control participant j, with the minimum distance $d(i,j)$ is chosen as the match for educated respondents i, and the matched pair is then removed from the pool. This process is continued until matches are found for all educated respondents.

The Mahalanobis metric matching including the propensity score, which is one of the techniques that Rosenbaum and Rubin outlined for constructing a matched sample, is exactly based on the procedure described above, with the singular addition of including the estimated propensity score e(x) as a covariate together with other covariates in the calculation of the Mahalanobis distance.[9,13]

Stratification Matching

Stratification on the propensity score involves stratifying subjects into mutually exclusive subgroups based on their estimated propensity scores. Subjects are ranked by their estimated propensity score. Subjects are then stratified into subgroups based on previously defined cut-offs of the estimated propensity score. The cases in the educated group are matched with the cases in the uneducated group within each of the strata. A typical approach is to separate subjects into five equal-size groups utilizing the quintiles of the estimated propensity score. Cochran (1968) exhibited that stratifying on the quintiles of a continuous confounding variable removed approximately 90% of the bias due to that variable.[14] Rosenbaum and Rubin (1984) stretched out this result to stratification on the propensity score, expressing that stratifying on the quintiles of the propensity score removes approximately 90% of the bias owing to measured confounders when estimating a linear treatment effect.[4] Following Cochran's (1968) suggestion, five strata were classified for stratification.

BALANCE DIAGNOSTICS

The selection bias B_k for covariate X_k , k=1, ..., K is defined as the mean difference in the covariate between the treatment conditions. That is:

$$B_k = M_i(X_k) - M_j(X_k) \tag{6}$$

Where $M_i(X_k)$ and $M_j(X_k)$ are the averages (which can be a mean or proportion) of covariate k for the treated cases and control cases respectively. To evaluate the selection bias, the standardized bias (SB) defined by Rosenbaum and Rubin (1985) was adopted, and defined as follows:[9]

$$SB_k = \frac{B_k}{\sqrt{\frac{V_i(X_k) + V_j(X_k)}{2}}} * 100\% \tag{7}$$

Where $V_i(X_k)$ and $V_j(X_k)$ are the variances of the covariate for all the treated cases and all the non-treated cases, respectively.

$$SB_k = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\frac{s^2 i + s^2 j}{2}}} * 100\%\text{, for continuous covariates} \tag{8}$$

$$SB_k = \frac{p_i - \bar{p}_j}{\sqrt{\frac{p_i(1-p_i) + p_j(1-p_j)}{2}}} * 100\%\text{, for dichotomous covariates} \tag{9}$$

It has been suggested that a standardized bias of at most 10% is quite sufficient at balancing a given covariate between treatment groups.[15]

Following Cochran and Rubin (1973), the percent bias reduction (*PBR*) on the covariate to assess the effectiveness of matching was also utilized.[16] They suggested a PBR value of at least 80% as being acceptable in judging the bias reduction effectiveness of a matching method. The percent bias reduction is defined as follows:

$$PBR_k = \frac{|B_{k,before\ matching}| - |B_{k,after\ matching}|}{|B_{k,before\ matching}|} * 100\% \tag{10}$$

RESULTS

The description of background characteristics of the 24222 respondents employed in this study is shown in table 1 below. Out of the 24222 respondents, 8194 were classified as educated (treated group) while the remaining 16028 were uneducated (control group).

TABLE 1: Summary statistics of covariates for the treated and control groups.

Characteristic	Educated N = 8194	Uneducated N = 16028
Respondent Age (years)	31.70 ± 7.97	31.49 ± 9.24
Age at first birth (years)	21.46 ± 4.49	18.27 ± 3.66
Number of siblings	5.41 ± 2.44	5.40 ± 2.81
Husband's age (years)	40.01 ± 9.92	43.13 ± 12.14
Marriage to first birth interval (months)	26.5 ± 26.84	33.12 ± 29.73
Residential area		
Rural	3241 (39.6%)	12461 (77.7%)
Urban	4953 (60.4%)	3567 (22.3%)
Sex of household head		
Male	7000 (85.4%)	14998 (93.6%)
Female	1194 (14.6%)	1030 (6.4%)
Husband educational status		
Educated	6774 (82.7%)	3550 (22.1%)
Uneducated	1420 (17.3%)	12478 (77.9%)
Respondent currently working?		
Yes	6444 (78.6%)	10785 (67.3%)
No	1750 (21.4%)	5243 (32.7%)

Continuous variables are reported as mean ± standard deviation, while dichotomous variables are reported as number (percent).

TABLE 2: The selection bias and the standardized bias on the raw data (before matching).

Characteristic	B_k	SB_k
Respondent Age (years)	0.2111	0.0265
Residential area	0.3820	0.7810*
Sex of household head	-0.0814	-0.2308*
Age at first birth (years)	3.1965	0.7105*
Number of siblings	0.0083	0.0034
Husband educational status	0.6052	1.5989*
Respondent currently working?	0.1135	0.2770*
Husband's age (years)	-3.3027	-0.3121*
Marriage to first birth interval (months)	-6.6234	-0.2468*

** indicates substantial amount of bias (>10%)*

Table 2 shows the selection bias and the standardized bias on the raw data. Quite a number of covariates had substantial standardized biases with values larger than 10% threshold as given by Normand et al (2001).[15] This necessitates the application of bias reduction methods.

Summary Statistics of the estimated propensity score are shown in table 3. The overall propensity score averaged at 0.3383 with quite a large spread of almost the same value of the mean. Results also show, with respect to group comparisons, that the mean propensity scores of the two groups differ greatly, and for the five number summary statistics, there were evidence of huge differences in the propensity scores first quartile, median, and third quartile.

TABLE 3: Summary statistics of the propensity score by exposure groups.

Characteristic	Educated N = 8194	Uneducated N = 16028	Overall N=24222
Mean	**0.6250**	**0.3274**	**0.3383**
Standard Deviation	0.2625	0.2202	0.3122
Minimum	0.0106	0.0059	0.0059
Maximum	0.9958	0.9849	0.9958
First quartile	0.4516	0.0494	0.0610
Median	0.6902	0.0805	0.1869
Third quartile	0.8436	0.2593	0.6254

Figure 1 also supports the claims made above the evidence of imbalance between the treatment and control groups, and that treatment assignment was confounded with observed covariates: The quantile-quantile plot, on the left panel of the figure, is far from being a straight line, thus, there is no evidence of similarity in terms of the quantiles of the propensity scores. Also, the boxplot (on the right panel) shows evidence of difference in the distribution of the propensity scores between the two groups.

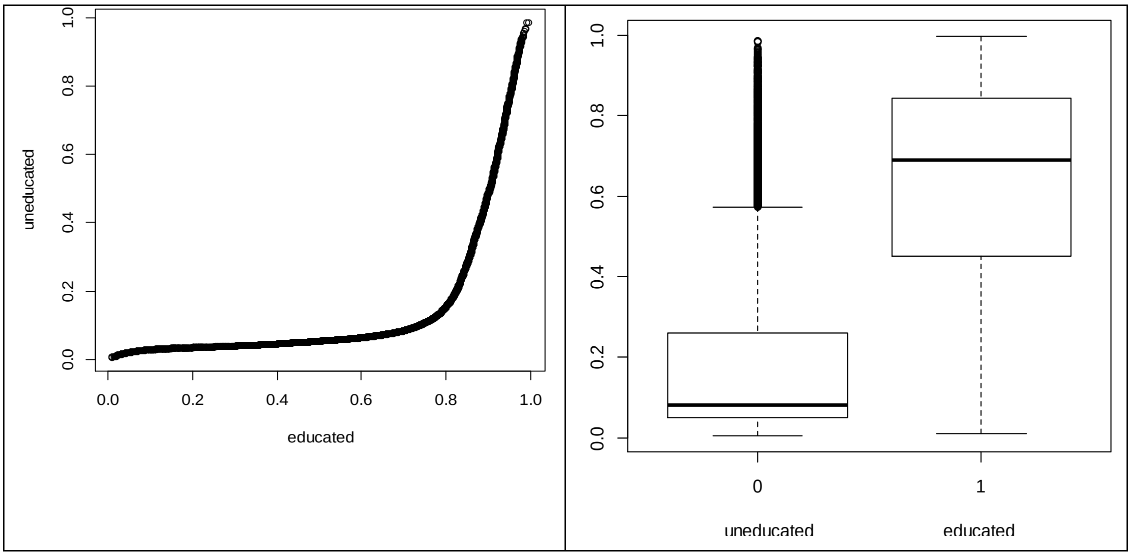

FIGURE 1: QQplot of the propensity score by educational status (left) and its Boxplot (right).

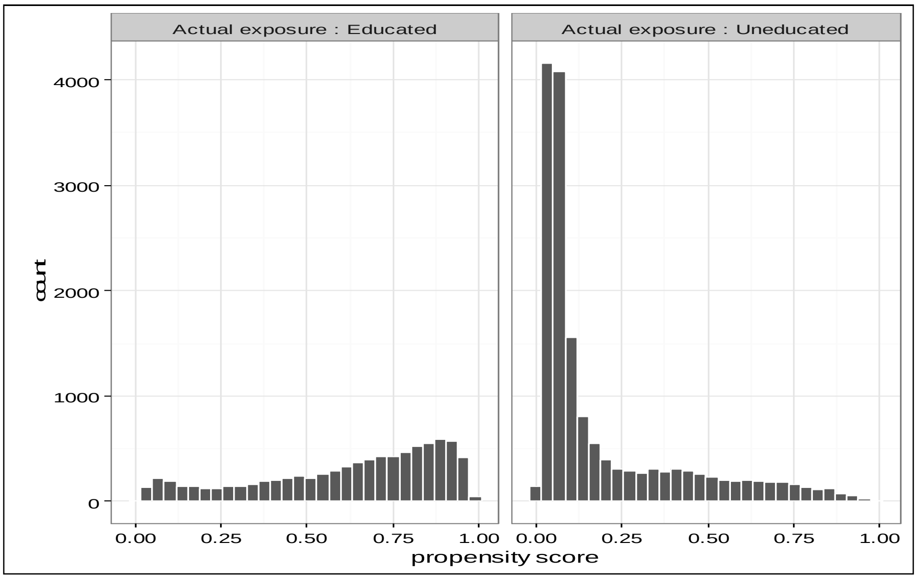

FIGURE 2: Histogram of the propensity score by educational status.

The distributions of the propensity score by the exposure conditions (i.e. educated vs. uneducated) are displayed in Figure 2, shows a sufficient common support, which in other words, according to Stuart, 2010, means there is substantial overlap of the propensity score distributions in the two groups, but potentially density differences.[17]

TABLE 4: Balance assessment after nearest neighbor matching (n = 8194).

Characteristic	B_k	SB_k (%)	PBR_k (%)
Respondent Age	-0.0531	-0.0067	74.8533
Residential area	0.1985	0.4058*	48.0418
Sex of household head	-0.0401	-0.1138*	50.7066
Age at first birth	1.8403	0.4091*	42.4278
Number of siblings	-0.0841	-0.0344	-910.1915
Husband educational status	0.3936	1.0398*	34.9685
Respondent currently working?	0.0207	0.0506	81.7279
Husband's age	-1.9562	-0.1848*	40.7703
Marriage to first birth interval	-4.0974	-0.1527*	38.1375

* indicates substantial amount of bias (>10%) (n=8194) indicates the number of matched pairs

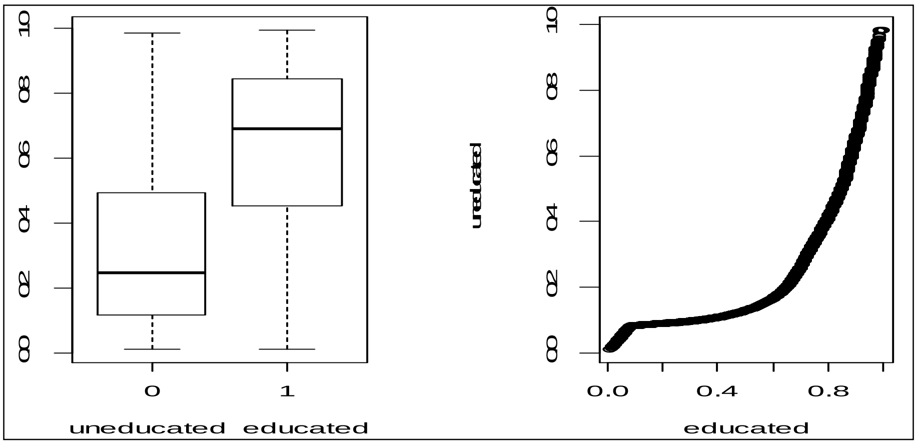

FIGURE 3: Boxplot of the propensity score by educational status (left) and its QQplot (right) after nearest neighbor matching.

In terms of bias reduction, after nearest neighbor matching, standardized biases were larger than 10% for six out of the nine covariates (Table 4). As for the percent bias reduction, all except one of the covariates had values less than 80%. Also, an examination of the quantile-quantile plot and boxplot shows that the two groups greatly differ in their empirical distributions (Figure 3). This all points to the fact that nearest neighbor matching did not do much of a good job in bias reduction.

TABLE 5: Balance assessment after Caliper matching (n = 4530).

Characteristic	B_k	SB_k (%)	PBR_k (%)
Respondent Age	-0.0905	-0.0114	57.128
Residential area	0.0064	0.0131	98.3238
Sex of household head	0.0016	0.0044	98.1029
Age at first birth	0.2808	0.0624	91.2153
Number of siblings	-0.0362	-0.0148	34.9367
Husband educational status	0.0089	0.0233	98.541
Respondent currently working?	-0.0106	-0.0259	90.6679
Husband's age	-0.3298	-0.0312	90.0142
Marriage to first birth interval	0.3995	0.0149	93.9675

(n=4530) indicates the number of matched pairs.

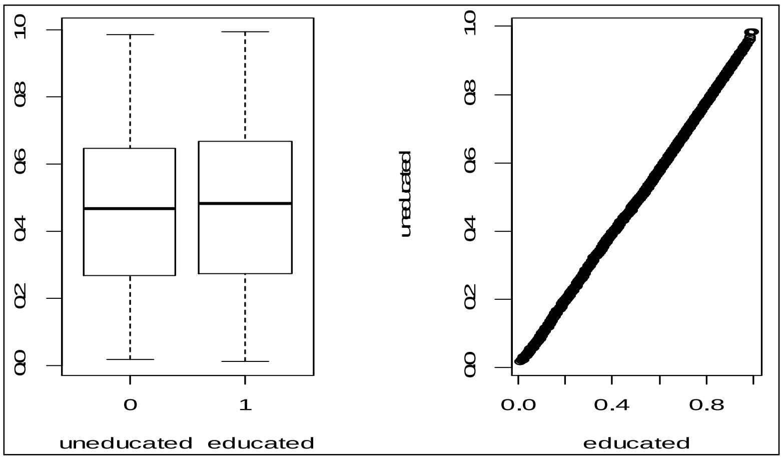

FIGURE 4: Boxplot of the propensity score by educational status (left) and its QQplot (right) after Caliper matching.

Table 5 shows the balance metrics after caliper matching. As typical of Caliper matching method, almost 45% of the data were lost; which indicates that almost half of the dataset does not meet the size of the caliper requirement. After caliper matching, standardized biases (with the highest value being 6%) were less than 10%, for all the nine covariates. Except for just two covariates (age, and number of siblings), the percent bias reduction had values at least 90%. Also, an examination of the quantile-quantile plot and boxplot shows that the two groups have identical distributions (Figure 4). Thus, it is evident that Caliper matching did an excellent job in bias reduction.

TABLE 6: Balance assessment after Mahalanobis metric matching including propensity score (*n* = 8194).

Characteristic	B_k	SB_k (%)	PBR_k (%)
Respondent Age	0.1926	0.3938*	-145.6286
Residential area	-0.0212	-0.0602	49.5757
Sex of household head	2.2375	0.4974*	73.93
Age at first birth	-0.0618	-0.0253	30.0001
Number of siblings	0.3998	1.0562*	-641.8823
Husband educational status	0.0028	0.0068	33.9401
Respondent currently working?	-2.8336	-0.2677*	97.5279
Husband's age	-2.2126	-0.0824	14.2059
Marriage to first birth interval	0.1926	0.3938*	66.5942

* indicates substantial amount of bias (>10%) (n=8194) indicates the number of matched pairs.

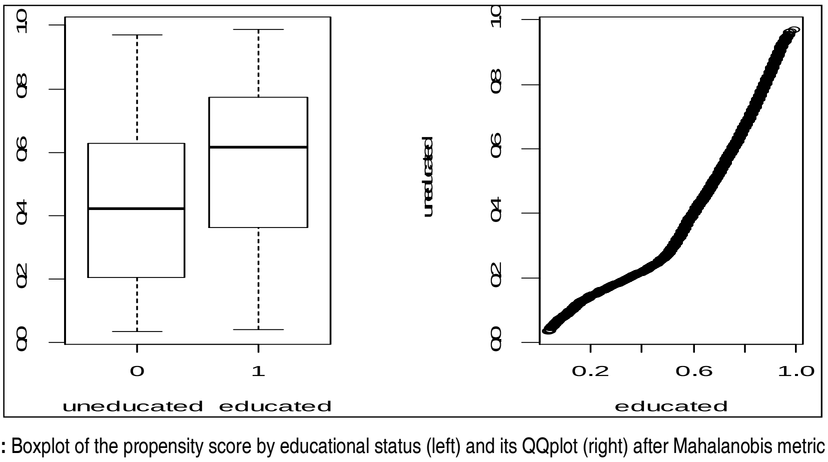

FIGURE 5: Boxplot of the propensity score by educational status (left) and its QQplot (right) after Mahalanobis metric matching.

Table 6 shows the balance metrics after Mahalanobis metric including the propensity score matching. It shows that standardized biases were larger than 10% for five out of the nine covariates. As for the percent bias reduction, Mahalanobis metric including the propensity score matching performance was not too good, as only one covariate (respondent working status) had the Cochran and Rubin, 1973's acceptable threshold value of at least 80% PBR value.[16] Also, an examination of the quantile-quantile plot and boxplot show that the two groups greatly differ in their empirical distributions (Figure 5). All these also points to the fact that Mahalanobis metric matching including the propensity score did not do much of a good job in bias reduction.

Figure 6 and 7 show the quantile-quantile plot and boxplot of propensity scores across the 'treatment' conditions in each stratum or quintile. From the boxplot, we can see that, except for quintile 1, the empirical distributions for the two treatment groups were similar in Quintiles 2, 3, 4 and 5. The quantile-quantile plots also provide us with the same information.

The mean biases (mean B_k s), the mean standard biases (SB_k s), in each quintile, and the mean percent bias reductions (PBR_k s) over the quintiles are listed in Table 7. Although some of the standard biases for each quintile were substantial, their average values across the 5 quintiles were all smaller than 10% except for just one covariate. As for the percent bias reductions, out of seven of the average PBR_k s, five of the covariates had values at least 80%.

TABLE 7: Selection bias, Standardized bias, and Percent bias reduction after Stratification over quintiles.

		SB_k						
Characteristic	**Q1 (n=1639)**	**Q2 (n=1639)**	**Q3 (n=1638)**	**Q4 (n=1639)**	**Q5 (n=1639)**	**Mean SB_k**	**Mean B_k**	**Mean PBR_k**
Respondent Age	0.1469	-0.041	-0.1041	-0.1223	-0.2109	-0.07	-0.53	-150.17
Residential area	0.3934	-0.0348	-0.0525	-0.0501	0.1419	0.08	0.04	89.81
Sex of household head	-0.1587	-0.0459	0.0791	0.0982	0.3238	0.06	0.02	74.31
Age at first birth	0.2141	0.0475	0.0783	0.0503	0.0295	0.08	0.38	88.19
Number of siblings	0.1455	-0.1076	-0.0586	-0.085	-0.1237	-0.05	-0.11	-1247.53
Husband educational status	0.4832	0.0532	0.0326	0.0000	0.0000	0.11*	0.04	92.88
Respondent currently working?	0.2809	-0.0462	-0.1629	-0.1297	-0.0716	-0.03	-0.01	90.66
Husband's age	0.0692	-0.1155	-0.1681	-0.1464	-0.1562	-0.10	-1.09	66.87
Marriage to first birth interval	-0.1840	0.0653	0.2431	-0.093	-0.2213	-0.04	-1.02	84.60

Each cell value is the standardized bias, and * indicates substantial amount of bias (>10%).

The final step is to select the best amongst the four adopted PSM methods. To summarize the systematic comparison of the PSM methods, the average$|SB_k|$, and average PBR_k for all the four PSM methods are listed in Table 8 below. The average$|SB_k|$, and average PBR_k is the mean standardized bias, and mean percent bias reduction respectively across all the covariates.

Considering the above balance metrics, it was shown that the Caliper matching, with mean standardized bias, and mean percent bias reduction values of 0.022, and 83.655 respectively, performed best in bias reduction. Stratification performed second best, while Mahalanobis metric including the propensity score, and nearest neighbor matching had the worst performances, with no clear distinction in the superiority of both.

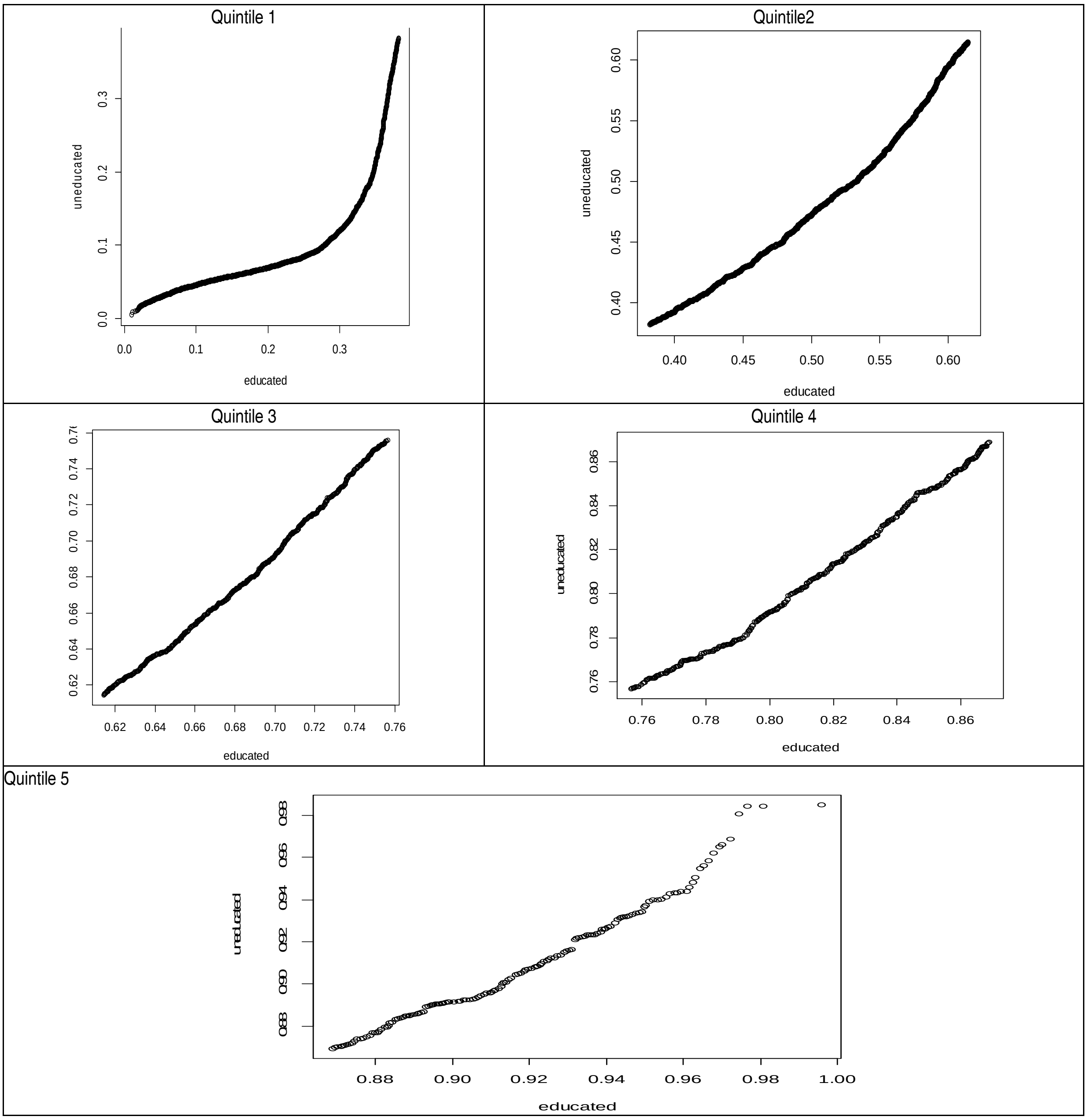

FIGURE 6: QQplots of propensity scores by educational status in each quintile after Stratification

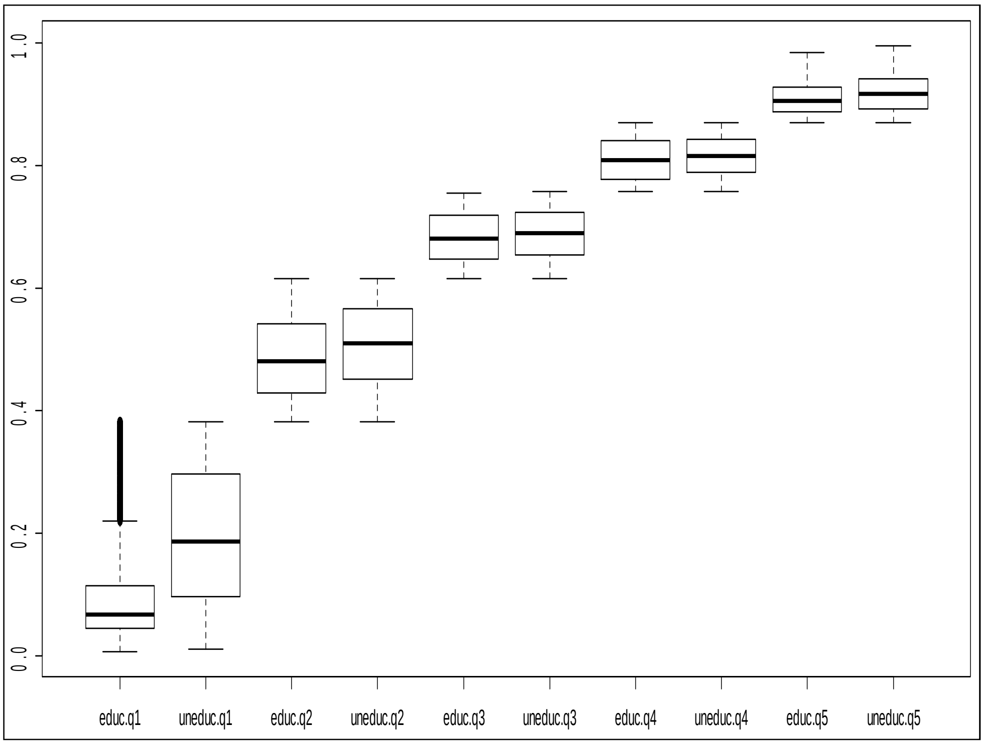

FIGURE 7: Boxplots of propensity scores by educational status in each quintile after Stratification

Note that "educ" is short for educated and "uneduc" is short for uneducated, and the q prefix is short for quintile

TABLE 8: Overall comparison of methods' bias reduction performance.

PSM method	Average $\lvert SB_k \rvert$	Average PBR_k
Before matching	0.465	—
Nearest neighbor matching	0.266	-55.395
Caliper Matching	0.022*	83.655*
Mahalanobis metric including the propensity score matching	0.309	-46.860
Stratification	0.069	-90.042

** indicates the best value.*

DISCUSSION

This study empirically compared and evaluated four (4) commonly used Propensity score matching methods: nearest neighbor, caliper, Mahalanobis metric including propensity score, and stratification matching, in terms of selection bias, standardized bias and percent bias reduction, using a national health survey data. The study results presented empirical advice for medical and health researchers to advance their knowledge on PSM methods and effectively choosing PSM for their research based on survey data. Because the data utilized is nation-wide based, insights as to how PSM methods perform in bias reduction on a typical survey data can be provided and generalized.

This study demonstrated that caliper matching is the most effective in bias reduction. This finding is consistent with previous research findings[8] that caliper matching was the best PSM technique. However, it is noteworthy that when the sample size is small or violates the statistical assumptions, caliper matching will possibly become problematic because it usually excludes the cases when they do not have matched pairs or do not meet the criterion for the caliper. Thus, when researchers use large datasets such as data from national databases (as we have used in this study), caliper matching is strongly advocated for; otherwise, researchers should be cautious about the potential limitation of caliper matching.

The results of the current study also show that Mahalanobis metric matching did not perform too well. This finding is consistent with the prior research, it was stated that Mahalanobis metric matching including the propensity score produced smaller standardized differences for individual variables but left a substantial difference along the propensity score.[9] These study results further confirmed the argument made by Guo, Barth and Gibbons (2006) that Mahalanobis matching with propensity scores need not be used in the PSM procedures.[8]

Nearest neighbour matching has been considered the most straightforward, relatively convenient matching method; however, its performance in this study was not impressive. One of the many possible reasons for the relatively poor performance of nearest neighbor matching is that it does not generally minimize the overall distance within pairs, and the theoretical arguments and simple examples have proven that its algorithm's distance can be much larger than the minimum attainable.[2,18] Therefore, nearest neighbor matching is recommended only when the sample has a relatively large comparison group as a sufficient storage for obtaining efficient matching pairs.

Although not the best, Stratification produced comparable results on bias reductions with caliper matching; it could be recommended as an effective matching method. The results of Stratification in this study confirms Cochran (1968) assertion that stratifying the propensity scores on five quintiles could remove over 90% of bias associated with the covariates.[14]

It is worthy of note that some researchers may not treat stratification as one of the PSM methods because it does not match case by case. Although, stratification is still included in this study as one of the matching methods for two reasons: (1) Stratification matches treatment and comparison groups for each stratum-that is, in stratification, the treated cases are matched with the non-treated cases by strata; and (2) stratification is an important technique often utilized by applied researchers when employing the PSM method. Therefore, it is necessary to include stratification in the systematic comparison of the commonly used PSM methods to provide essential information for researchers in PSM selections.[19]

Although common, significance tests that includes information on the sample size (e.g. t-test and Chi-square test for significant differences in the means and proportions respectively of the two groups) were not used as balance measures as advised by Austin (2007), and Imai, King, and Stuart (2008). They gave two reasons: First, balance is inherently an in-sample property, without reference to any broader population. Second, hypothesis tests can be misleading as balance measures, because they often confuse changes in balance with changes in statistical power.[20,21]

This study has a few limitations: Firstly, not all the available PSM methods were compared for bias reduction. These other PSM methods were not discussed and utilized in this study because they are neither significantly better nor commonly used. Secondly, a Monte Carlo simulation study is expected to be carried out to investigate the performance of these PSM methods under different data structures.

CONCLUSION

In conclusion, this research work can help medical and health researchers to choose appropriate propensity matching methods to estimate treatment effect in the presence of confounding variables.

This research work shall immensely help in exposing the strength and weakness of Propensity score matching methods for Observational data, which in the absence of experimental data, can be used to make causal inferences. Nevertheless, validation studies with different data structures might be desirable in future to give more credence to results obtained in this study.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

This study is entirely author's own work and no other author contribution.

REFERENCES

1. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med 2006;25(12):2084-106.
2. Rosenbaum PR. Observational Studies. 2nd ed. New York: Springer-Verlag; 2002. p.375.
3. Rosenbaum PR. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55.
4. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79(387):516-24.
5. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med 2010;29(3):337-46.
6. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004;9(4):403-25.
7. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf 2008;17(6):546-55.
8. Guo S, Barth R, Gibbons C. Propensity score matching strategies for evaluating substance abuse services for child welfare clients. Child Youth Serv Rev 2006;28(4):357-83.
9. Rosenbaum PR, Rubin DP. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Statist 1985;39(1):33-8.
10. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127(8 Pt 2):757-63.
11. NDHS. Nigeria Demographic Health Survey. Rockville, Maryland, USA: Published by National Population Commission; 2013. p.411.
12. Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42(8).
13. Guo S, Fraser MW. Propensity Score Analysis; Statistical Methods and Applications. 1st ed. Thousand Oaks, Calif: SAGE Publications; 2010. p.392.
14. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968;24(2):295-313.
15. Normand SL, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, et al. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. J Clin Epidemiol 2001;54(4):387-98.
16. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. Sankhya Serial A 1973;35(4):417-46.
17. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010;25(1):1-21.
18. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat 1993;2(4):405-20.
19. Bai H. A comparison of propensity score matching methods for reducing selection bias. Int J Res Meth Educ 2011;34(1):81-107.
20. Austin PC. The performance of different propensity score methods for estimating marginal odd ratios. Stat Med 2007;26(16):3078-94.
21. Imai K, King G, Stuart EA. Misunderstandings among experimentalists and observationalists in causal inference. J R Statist Soc A 2008;171(2):481-502.