Contemporary Clinical Trials 47 (2016) 85–92

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Propensity score and proximity matching using random forest

Peng Zhao [a], Xiaogang Su [b], Tingting Ge [c], Juanjuan Fan [d,*]

[a] Computational Science Research Center, San Diego State University, San Diego, CA, USA
[b] Department of Mathematical Sciences, University of Texas, El Paso, TX, USA
[c] Janssen Research and Development, San Diego, CA, USA
[d] Department of Mathematics and Statistics, San Diego State University, San Diego, CA, USA

ARTICLE INFO

Article history:
Received 1 March 2015
Received in revised form 7 December 2015
Accepted 14 December 2015
Available online 17 December 2015

Keywords:
Observational study
Matching
Propensity score
Proximity
Random forest

ABSTRACT

In order to derive unbiased inference from observational data, matching methods are often applied to produce balanced treatment and control groups in terms of all background variables. Propensity score has been a key component in this research area. However, propensity score based matching methods in the literature have several limitations, such as model mis-specifications, categorical variables with more than two levels, difficulties in handling missing data, and nonlinear relationships. Random forest, averaging outcomes from many decision trees, is nonparametric in nature, straightforward to use, and capable of solving these issues. More importantly, the precision afforded by random forest (Caruana et al., 2008) may provide us with a more accurate and less model dependent estimate of the propensity score. In addition, the proximity matrix, a by-product of the random forest, may naturally serve as a distance measure between observations that can be used in matching. The proposed random forest based matching methods are applied to data from the National Health and Nutrition Examination Survey (NHANES). Our results show that the proposed methods can produce well balanced treatment and control groups. An illustration is also provided that the methods can effectively deal with missing data in covariates.

1. Introduction

In an experimental study, treatments are assigned randomly to study subjects. This ensures that the treated and control groups have similar distributions in terms of confounding and other risk factors, both of which we will refer to as covariates. For ethical or practical reasons however, the random assignment of treatment is not always possible, in which case one may resort to observational studies for the effect of treatment on the disease. Because the distributions of covariates are often unbalanced between the treated and control groups in an observational study, these differences can lead to biased estimates of treatment effects. Common adjustments include matching, stratification, covariance adjustment, and weighted analysis. Note that adjustment of covariates through regression, or covariance adjustment, may be inadequate to eliminate the bias in observational studies [22].

As one popular method for drawing unbiased inferences from observational studies, the idea behind matching is to find control subjects that are as similar to the treated subjects as possible in terms of covariates, hence creating balanced treatment groups just as in a randomized study. Refs. [8,23] discuss three matching methods based on Mahalanobis distance [18] and propensity score [21], namely, nearest available matching on the estimated propensity score, Mahalanobis metric matching including the propensity score, and nearest available metric matching within calipers defined by the propensity score. More details on these matching methods will be provided in Section 3.

The Mahalanobis distance is often used to quantify the similarity between two subjects i and j based on their covariate values and may be written as

$$d(i,j) = (X_i - X_j)' C^{-1} (X_i - X_j) \quad (1)$$

where X_i and X_j are covariate vectors for subjects i and j respectively, and C is the sample covariance matrix of all available control subjects in the data. The propensity score is defined as the conditional probability of treatment given covariates,

$$P(Z = 1|X) = E(Z|X) \quad (2)$$

where Z is the treatment indicator while X denotes all covariates excluding the treatment indicator. The propensity score provides a scalar summary of all the covariates: under the assumption of strong ignoreability, the distribution of X given the propensity score is balanced between the treated and control groups [21]. If there are no missing data on covariates, the propensity score may be estimated for each subject using a logistic regression or discriminant analysis.

Though matching by Mahalanobis distance and propensity score may be successful in constructing balanced treated and control groups

* Corresponding author.
E-mail address: jjfan@mail.sdsu.edu (J. Fan).

http://dx.doi.org/10.1016/j.cct.2015.12.012
1551-7144/

in certain situations, as in the data examples used by Refs. [8,23], there are a number of shortcomings associated with the calculation of Mahalanobis distance as well as the estimation of propensity score using logistic regression or discriminant analysis. First of all, the parametric logistic model is prone to model mis-specification, which results in inaccurate or imprecise estimation of propensity score. Secondly, the Mahalanobis distance is only well-defined for continuous variables, but not for categorical variables. The same problem exists for many other distance measures in the literature. Thus a distance measure that handles both continuous and categorical variables is desirable. Another problem is missing data. With ever increasing number of covariates being collected, missing data become inevitable. In order to compute Mahalanobis distance and perform logistic regression or discriminant analysis, either missing data have to be imputed beforehand or a complete case analysis has to be performed. When the number of covariates is large, even a small percentage of missing data on each covariate may result in a large proportion of observations with missing data on certain covariates and hence become unusable (without imputation) in a multivariate analysis such as the logistic regression.

Fortunately, all the issues discussed above may be solved by using random forests (RF; [1]). RF is a model ensemble method built on the regression and classification trees (CART, [2]). RF excels in predictive modeling and provides a unified way of defining distance for data with a mixture of continuous and categorical variables. Besides, RF has many other merits that are practically appealing and useful for matching. These include missing data handling, automatic variable selection, no need for monotonic variable transformation, and efficient handling of categorical variables. The main shortcoming of random forests and other tree based methods is that they tend to work well with larger and higher dimensional data, which is usually not a problem in this era of big data especially because the main motivation of propensity score is to serve as a one-dimensional balancing score, overcoming the high dimensionality of covariates. In this paper, we propose matching methods based on propensity score and proximity matrix from random forest. As detailed in Section 3, both the propensity score and proximity matrix can be easily calculated once a random forest has been constructed.

There is a growing literature on using tree based methods to calculate the propensity score, see, e.g., Refs. [5,15], and references therein. In particular, a few articles have found that logistic regression gives subpar performance compared to tree ensembles or other machine learning methods [15,24,27]. To our best knowledge, no tree or random forest based methods have been proposed for matching.

As for missing data handling in propensity score estimation, the most typical way for treating missing data in the logistic regression based approach is multiple imputation, though a few variants are available. See, e.g., Refs. [14,11,19]. In the recursive partitioning framework, there are several ways of dealing with missing data, including surrogate splits and treating missing values as a separate category. Comparison studies among these strategies can also be found in the recent literature. See, e.g., Refs. [9,10,13,20]. The surrogate splits in tree based methods work well for data missing completely at random (MCAR), though simulation studies conducted by Refs. [20] found no substantial overall difference in the performance of surrogate split and multiple imputation approaches with missing data generated under both MCAR and missing at random (MAR). Surrogate splits will be used in this paper because they are more naturally integrated in the random forest algorithm while multiple imputation has to be performed beforehand.

The rest of the paper is organized as follows. Section 2 provides a description of the National Health and Nutrition Examination Survey (NHANES) data. Section 3 details the random forest based matching methods, including missing data handling as well as propensity score and proximity matrix calculations. Section 4.1 presents matching results of the NHANES data using the proposed matching methods. In Section 4.2, we demonstrate how the artificial effect of treatment due to unbalanced samples may be removed after matching. In Section 4.3, we illustrate that the proposed matching methods can produce balanced samples even when data include covariates with rather high missing rates. Section 5 concludes the paper with a brief summary.

2. Data

We use the National Health and Nutrition Examination Survey (NHANES) collected in 1999–2000 and 2001–2002 to describe the application of this method. The data can be downloaded from the Centers for Disease Control and Prevention (CDC) website. We are interested in studying the effect of smoking on body mass index (BMI) in this observational study dataset. Though there is a popular belief that smoking keeps a person thin, there is also evidence that smoking may increase central (or abdominal) obesity. Among the current epidemic of obesity and growing concerns of the adverse effect of smoking on various diseases, it is fair to say that the relationship between smoking and body mass index is not nearly resolved [6,17].

After excluding all pregnant subjects, the total sample size of this dataset is 8288 subjects, including 4067 smokers (treated subjects) and 4221 nonsmokers (control subjects). Smoker was designed as a binary variable (Yes/No) and defined by (1) smoked at least 100 cigarettes in life; (2) smoke cigarettes currently every day or some days; or (3) number of cigarettes smoked per day now is greater than 0. Available background covariates include four continuous variables: age, family poverty income ratio, systolic blood pressure (SBP), and diastolic blood pressure (DBP); nine binary variables: birth place, marital status, alcohol use, drug use, health insurance, private health insurance, home ownership, vigorous activity over past 30 days, and moderate activity over past 30 days; five ordered categorical variables: education, household size (number of people in the household), daily physical activity, compare activity with others same age, and compare activity with 10 years ago; and one unordered categorical variable: race. Employment status is also a background covariate in this dataset, but since all subjects were employed, this variable was not considered for matching. As seen in Table 1, most of the covariates contain missing values. The

Table 1
Variable description and number (percent) of missing values among smokers and nonsmokers.

Covariates		Nonsmoker	Smoker
Name	Description	($N_0 = 4221$)	($N_1 = 4067$)
Age	Age in years at examination	0 (0.00%)	0 (0.00%)
Race	Race/ethnicity	0 (0.00%)	0 (0.00%)
BirthPlace	Birth place (US or not)	2 (0.05%)	1 (0.02%)
Education	Highest grade or level	5 (0.12%)	6 (0.15%)
Marital	Marital status	219 (5.19%)	196 (4.82%)
HouseholdSize	# of people in the household	0 (0.00%)	0 (0.00%)
FMPIR	Family poverty income ratio	440 (10.42%)	412 (10.13%)
Alcohol	Alcohol use (yes or no)	304 (7.20%)	183 (4.50%)
Drug	Drug use (yes or no)	1586 (37.57%)	1643 (40.40%)
HIC	Health insurance coverage?	69 (1.63%)	50 (1.23%)
PHIC	Private health insurance coverage?	909 (21.54%)	965 (23.73%)
Owned	Home owned or not?	67 (1.59%)	44 (1.08%)
PHSAVG	Average level of physical activity	7 (0.17%)	2 (0.05%)
PHSVIG	Vigorous activity over past 30 days	2 (0.05%)	2 (0.05%)
PHSMOD	Moderate activity over past 30 days	3 (0.07%)	3 (0.07%)
PHSCOM	Compare activity with others same age	62 (1.47%)	62 (1.52%)
PHS10YR	Compare activity with 10 years ago	845 (20.02%)	542 (13.33%)
SBP	Systolic blood pressure	145 (3.44%)	137 (3.37%)
DBP	Diastolic blood pressure	145 (3.44%)	137 (3.37%)

only covariates with no missing values are age, race and household size. Drug use has 40% missing values among smokers and 38% missing values among nonsmokers. The mean BMI was 28.5 kg/m^2 in full nonsmoker dataset and 28.1 kg/m^2 in full smoker dataset.

In order for the matching to work well, the sample size of treated subjects (smokers) needs to be much smaller than that of controls (nonsmokers). That is, we will need to find the most similar controls from a large control reservoir to match a relatively small group of treated subjects. To this end, a random sample of 500 smokers was selected and the full nonsmoker dataset was used as the control reservoir for matching. In this way, this random subset of 500 smokers could find its matching nonsmoker group from the full nonsmoker pool. This random sample of 500 smokers remained the same for all matching methods described below.

3. Method

In this section, we propose matching methods for observational data using propensity scores and proximity measures based on random forest. Recall that we are interested in investigating the effect of smoking (treatment, denoted by Z) on the body mass index (BMI, response Y) in the NHANES data. In order to compute the propensity score (Eq. (2)), random forests are constructed using the treatment indicator Z as the output and all other covariates, or X, as inputs. A proximity matrix is also computed as the by-product of the constructed random forest.

3.1. Random forest

Random forest (RF) is an ensemble method based on decision (classification or regression) trees. Classification and Regression Trees (CART, [2]) are built by recursively partitioning the data based on covariates, so that the observations in a node become increasingly pure (in terms of outcome) as data move from the root node to terminal nodes. Each terminal node is often summarized by the average outcome value of all the observations that end up there. There are two main differences between a tree in random forest and a decision tree. As each tree being constructed in random forest, at each internal node, only a random subset of covariates are evaluated to reach the best split. The trees in random forest are usually fully grown and not pruned. The averaging process of random forest greatly improves the prediction accuracy over a single tree [12].

In standard random forest [1], each tree is constructed based on a bootstrap sample of the training data. To use the data more efficiently, we use all of the data to build each tree so that propensity score and proximity measure can be calculated for each observation and between any pair of observations respectively, based on each tree. We have explored from 100 to 5000 trees in the random forest and found that the matching results become stable after 500 trees. Therefore, all the results reported in this paper are based on random forests of 500 trees. There is no pruning in our application of random forest. All trees were grown as large as possible until the node is pure (there are only treated or control subjects, or all subjects share the same values on all covariates) or until the node reaches a pre-determined minimum size of 25.

To split an internal node in a tree, a pre-defined number of covariates, which is by convention the square root of the total number of covariates in the dataset, are randomly selected. If a selected covariate is a continuous or an ordered categorical variable, every possible cut-point is considered. If a selected covariate is an unordered categorical variable, such as race, any possible subset of its categories is considered. For each proposed split, we form a 2×2 table based on the split (left versus right child nodes) and the treatment indicator (smoker versus nonsmoker in our application), and then calculate the chi-square or Fisher's exact test for the 2×2 table. The covariate and the corresponding cut-point producing the smallest p-value are chosen as the primary split.

All random forest and related codes were developed using statistical freeware R (www.r-project.org). One main output of RF is the estimated propensity score. Since RF is a model ensemble tool that is nonparametric in nature, the risk of model misspecification is minimized. While many other nonparametric machine learning tools, such as boosting, support vector machines, and artificial neural networks (see, e.g., Refs. [5,25], and references therein), can be used for the same purpose of estimating propensity score, RF is deemed as the top performer or among the top in predictive modeling among many other competitors. See, e.g., Refs. [1,3,4,12,16] for simulation studies, real data examples, and discussions.

3.2. Missing data method

One of the main advantages of decision tree based methods is the ease to handle missing data. Unlike other modeling methods, it is not necessary to impute or delete any missing data before constructing trees in random forest. The method used to deal with missing data is the so-called surrogate-split method, as detailed in CART [2]. Since surrogate-split is not implemented in the R package *randomForest*, we detail our algorithm below.

At each split, denote the splitting variable and the cut-point of the primary split by X_1 and c_1. For each of the other randomly selected covariates, $X_2,\ldots,X_m$, all cut-points are examined to find the best cut-off, denoted by $c_2,\ldots,c_m$ for $X_2,\ldots,X_m$ respectively. The best cut-off for surrogate splits is defined as achieving the highest agreement with the primary split: a 2×2 table is formed based on the primary split and a possible cut-point for one of the selected variables $X_2,\ldots,X_m$, and the agreement is defined as the proportion of data falling along the diagonal. All of the randomly selected variables, $X_2,\ldots,X_m$, are then ranked based on their highest agreement with the primary split. During the node-splitting process, any observation is sent to the left or right child node based on the primary split. If an observation has missing value on the primary splitting variable, it will be sent based on the first surrogate split. If an observation has missing values on both the primary and first surrogate splitting variables, it will then be sent based on the second surrogate split. If an observation has missing values on all of the selected covariates ($X_1,\ldots,X_m$) at a split, it will then be sent with the majority.

The Kappa statistic [7] was also applied to determine the order of surrogate covariates, and it yielded very similar results to the percentage of agreement method. Since a percentage of agreement is easy to calculate and interpret and is what CART uses, all the results presented in this paper are based on surrogate splits ranked by the percentage of agreement.

3.3. Calculation of propensity score and proximity matrix

Once a tree is constructed, the propensity score for each subject and the distance measure between any pair of subjects are easily calculated, using only information contained in the terminal nodes. Note that the propensity score is a conditional probability, hence the percent of treated at the terminal node where any subject falls into gives the estimated propensity score for the subject, given the tree structure. For each terminal node, the percentage of treated subjects is calculated and this is passed to all the subjects in the terminal node as their estimated propensity score. The distance between any two subjects in the same terminal node is defined as 0, while the distance between any two subjects from different terminal nodes is defined as 1. Because we use all of the data to construct each tree in the random forest, there is a propensity score for each subject and a distance measure between any pair of subjects in the data based on each tree. The propensity score and the proximity matrix from all trees in the forest are then averaged to arrive at the final estimates of propensity score and proximity matrix. The algorithm is summarized in Appendix A.

Tree based methods handle variables of all types in a unified way by essentially treating all variables as discrete (each introducing a finite number of ways of partitioning data). Referring to one single tree in the RF, any pair of subjects falling into different terminal nodes have been assigned a distance of 1. This is the standard way of computing the proximity/distance matrix in RF. It is worth noting that there are several variants for computing the distance in our setting. First of all, observations that fall into different terminal nodes can be different to varying degrees. One alternative way of assigning distance is based on the *p*-value computed from a two-sample statistic that compares each pair of terminal nodes. See Ref. [26] for examples. Secondly, the treatment variable is used as the target variable in constructing RF. Alternatively, one could have used a pseudo-outcome that is randomly generated. The resultant distance matrix with the latter approach will be solely based on the covariates, resembling more the Mahalanobis distance. Nevertheless, using RF that predicts the treatment assignment has the advantage of taking the propensity score information into account. This also allows the calculation of propensity score and proximity matrix to be based on one single forest.

3.4. Matching based on propensity score and/or proximity

Once the propensity scores and distance measures are calculated, the matching methods are similar to those in Ref. [8] using the propensity score only, the distance measure only, or both propensity score and distance measures. In order to evaluate the different matching methods, all treated subjects are randomly ordered and the same random sequence is used for all three matching approaches below.

3.4.1. The nearest available matching based on propensity score

Based on the random sequence of treated subjects, the first treated subject is matched with the control subject who has the most similar propensity score. Then both subjects are matched and marked as unavailable for other subjects to match. This process is repeated until all treated subjects have their own match.

3.4.2. The nearest available matching based on distance

Based on the random sequence of treated subjects, the first treated subject is matched with the control subject who has the smallest distance from the treated subject. Then both subjects are considered to be matched and marked as unavailable for other subjects to match. This process is repeated until all treated subjects have their own match.

3.4.3. The nearest available matching based on distance within calipers defined by the propensity score

Based on the random sequence of treated subjects, all control subjects whose propensity scores are within a small range, say half standard deviation of the propensity scores of all subjects, of the first treated subject are considered. Within this small set of control subjects, the subject with the smallest distance from the treated subject is selected. Both are then considered as being matched and unavailable for other subjects to match. All the other control subjects in the small set become available to be matched with other treated subjects. This process is repeated until all treated subjects have their own match.

4. Results

4.1. Matching results

Table 2 presents the distributions and comparisons of all background covariates in the full dataset prior to matching. As can be seen from the table, most of covariates are unbalanced, with all but five covariates reaching statistical significance at the 0.01 level. The last column of Table 2 presents the standardized difference in percent given by $\frac{\bar{x}_{NS}-\bar{x}_S}{\sqrt{(s^2_{NS}+s^2_S)/2}} \times 100$, where $\bar{x}$ and s^2 denote sample mean and sample variance, and the subscripts NS and S refer to nonsmokers and smokers respectively. Half of the standardized differences exceed 10%. The covariates that are most unbalanced are alcohol and drug use, with their standardized differences exceeding 50%. One can see that only 14.21% of smokers use alcohol while 33.71% of nonsmokers use alcohol. Similarly, the percentages of drug users among smokers and nonsmokers are 40.40% and 56.72%, respectively. Since alcohol and drug use may affect an individual's lifestyle, which may in turn affect their body mass index (BMI), it is important to remove large differences in such background information when studying the effect of smoking on BMI. Note from Table 1, there is about 40% missing data in drug use, which would create difficulties for traditional matching methods using logistic regression and Mahalanobis distance.

To facilitate the evaluation and comparison of the proposed matching methods based on random forest, the same randomly selected 500 smokers were used to construct matched non-smoker groups. Table 3 presents the *p*-values and standardized differences for the comparison of the 19 background covariates between the random 500 smokers (treated subjects) versus 500 randomly selected nonsmokers, as well as 500 matched nonsmokers (controls) based on the three matching methods. As can be seen from the table, the methods of distance only matching and distance matching within calipers of propensity score produced well balanced smoker and non-smoker groups for all 19 covariates. To be conservative, we may define the treated and control groups as balanced for a covariate if the corresponding comparison has a *p*-value greater than 0.10, or a standardized difference less than 10%. However, the matching method based only on propensity score did not perform as well as the other two matching methods: many of the background covariates are still statistically significantly different (unbalanced) between the smoker and non-smoker groups after matching. This might be explained by the structure of the tree models. In order to have a small distance, the two subjects have to be in the same terminal node in most trees, which has two implications: (1) they largely share the same covariate values for those covariates that are important predictors of treatment, (2) they also have similar propensity scores. On the other hand, two subjects may have similar propensity scores even when they are in different terminal nodes. Ref. [23] found that the nearest available Mahalanobis metric matching within calipers defined by the propensity score performed the best since it used both propensity score and Mahalanobis distance when seeking matched subjects, as also reported in Ref. [8]. In a sense, our matching results are consistent with their findings considering that the distance (or proximity) we calculate also includes information on propensity.

Since matching based on proximity matrix alone and matching based on proximity within calipers defined by the propensity score can both produce well balanced samples and the former is a bit easier and more straightforward to implement, we will use the nearest available matching based on distance (proximity) in the rest of the paper. Fig. 1 presents histograms and density curves of the logit of propensity scores for smokers and nonsmokers. The same 500 smokers were used in both (a) and (b), while the 500 nonsmokers were selected randomly in (a) and by distance in (b). The plots show much better overlap in the distributions of propensity scores between the treatment groups after matching, indicating that the covariate distributions are much more balanced after matching.

4.2. Engineered samples

The purpose of matching is to produce balanced treatment groups for data collected from an observational study. As a result, the effect of treatment on the response may be uncovered with a direct comparison of the outcome variable from the two groups. For the NHANES data, however, the mean of BMI is similar between the two smoking groups with 28.5 kg/m^2 for non-smokers and 28.1 kg/m^2 for smokers. (The

Table 2
Group comparisons prior to matching (all data). Mean and standard deviation (SD) are presented for continuous variables, while number and percentage (%) of observations are presented for categorical variables. The standardized difference in % is given by $\frac{\bar{x}_{NS}-\bar{x}_S}{\sqrt{(s^2_{NS}+s^2_S)/2}} \times 100$, where $\bar{x}$ and s^2 denote sample mean and sample variance, and the subscripts NS and S refer to nonsmokers and smokers respectively.

Covariates		Nonsmoker ($N_0 = 4221$)		Smoker ($N_1 = 4067$)		*p*-Value	Std Diff %
Age	Mean (SD)	48.2	(18.19)	51.6	(17.05)	<0.0001	−18.9
Race	Hispanic	1373	(32.53%)	1067	(26.24%)	<0.0001	–
	White	145	(3.44%)	112	(2.75%)		
	Black	1820	(43.12%)	2135	(52.50%)		
	Other	883	(20.92%)	753			
BirthPlace	1 = US	3003	(71.14%)	3298	(81.09%)	<0.0001	23.5
	2 = Not US	1216	(28.81%)	768	(18.88%)		
Education	1 = Less than 9th grade	705	(16.70%)	631	(15.52%)	<0.0001	13.5
	2 = 9–11th grade	658	(15.59%)	833	(20.48%)		
	3 = High school graduate	892	(21.13%)	1029	(25.30%)		
	4 = Some college or AA degree	1039	(24.62%)	987	(24.27%)		
	5 = College graduate or above	922	(21.84%)	581	(14.29%)		
Marital	1 = Yes	1658	(39.28%)	1675	(41.19%)	0.1031	3.7
	0 = No	2344	(55.53%)	2196	(54.00%)		
HouseholdSize	1	505	(11.96%)	597	(14.68%)	<0.0001	17.8
	2	1242	(29.42%)	1404	(34.52%)		
	3	764	(18.10%)	720	(17.70%)		
	4	706	(16.73%)	615	(15.12%)		
	5	461	(10.92%)	354	(8.70%)		
	6	209	(4.95%)	156			
	7 or more	334	(7.91%)	221			
FMPIR	Mean (SD)	2.7	(1.63)	2.6	(1.60)		10.7
Alcohol	1 = Yes	1423	(33.71%)	578	(14.21%)	<0.0001	−50.7
	0 = No	2494	(59.09%)	3306	(81.29%)		
Drug	1 = Yes	2394	(56.72%)	1643	(40.40%)	<0.0001	−59.4
	0 = No	241	(5.71%)	781	(19.20%)		
HIC	1 = Yes	802	(19.00%)	887	(21.81%)	0.0022	6.8
	0 = No	3350	(79.37%)	3130	(76.96%)		
PHIC	1 = Yes	696	(16.49%)	813	(19.99%)	<0.0001	12.3
	0 = No	2616	(61.98%)	2289	(56.28%)		
Owned	1 = Yes	1389	(32.91%)	1383	(34.01%)	0.3821	2.0
	0 = No	2765	(65.51%)	2640	(64.91%)		
PHSAVG	1 = Sit during the day, does not walk about much	1006	(23.83%)	1009	(24.81%)	<0.0001	−6.9
	2 = Stand or walk a lot, but does not lift things	2346	(55.58%)	2071	(50.92%)		
	3 = Lift light load and climb stairs or hills often	631	(14.95%)	646	(15.88%)		
	4 = Do heavy work or carry heavy loads	231	(5.47%)	339	(8.34%)		
PHSVIG	1 = Yes	2861	(67.78%)	2967	(72.95%)	<0.0001	11.4
	0 = No	1358	(32.17%)	1098	(27.00%)		
PHSMOD	1 = Yes	2427	(57.50%)	2442	(60.04%)	0.0196	5.2
	0 = No	1791	(42.43%)	1622	(39.88%)		
PHSCOM	−1 = Less active	719	(17.03%)	791	(19.45%)	0.0002	1.0
	0 = About the same	1886	(44.68%)	1648	(40.52%)		
	1 = More active	1554	(36.82%)	1566	(38.51%)		
PHS10YR	−1 = Less active	1951	(46.22%)	2164	(53.21%)		7.1
	0 = About the same	954	(22.60%)	921	(22.65%)		
	1 = More active	471	(11.16%)	440	(10.82%)		
SBP	Mean (SD)	126.0	(21.08%)	126.5	(19.84%)		−2.8
DBP	Mean (SD)	72.2	(13.37%)	71.9	(13.62%)	0.3874	1.9

CDC defines the BMI range of 25.0 to 29.9 as "overweight".) In addition, the BMI values between smokers and non-smokers are compared in four scenarios: the randomly selected 500 smokers compared to a random sample of 500 non-smokers, and the same 500 randomly selected smokers compared to the 500 non-smokers selected by each of the three matching methods. All four comparisons yielded statistically non-significant results with *p*-values ranging from 0.14 to 0.45.

This subsection is designed to demonstrate how matching may facilitate the comparison of the response between two treatment groups by removing the bias resulting from unbalanced distributions in important covariates. To this end, we first generate unbalanced smoker/nonsmoker samples (or "engineered" samples) with respect to BMI. Next, matched treatment groups will be selected using the proposed RF based method, and BMI distributions will be compared again in matched samples. We start from all data as summarized in Table 1. Using R package *randomForest*, it was found that SBP and age are two most important inputs affecting BMI. After excluding all subjects with SBP values missing, a sample of 8006 subjects (3930 smokers and 4076 non-smokers) remains. The median age of this sample is 48.75 years and the median SBP is 123 mm/Hg. Among all the smokers in this sample, we include all subjects with both age and SBP below the median, 1/2 of those with age above the median and SBP below the median, 1/2 of those with age below the median and SBP above the median, and 1/4 of those with both age and SBP above the median. So members of the "engineered" smoker group, on average, are younger and have lower systolic blood pressure. As a result, they are also thinner (see Table 5 for BMI comparisons). Among all the non-smokers in this sample, we do the opposite: we include everyone with both age and SBP above the median, 1/2 of those with one variable (age or SBP) below the median and one variable (SBP or age) above the median, and 1/4 of those with both age and SBP below the median. So, members of the "engineered" non-smoker group, on average, are older and have higher systolic blood pressure. As a result,

Table 3
Group comparisons after matching (Based on the same 500 randomly selected smokers). The standardized difference in % is given by $\frac{\bar{x}_{NS}-\bar{x}_S}{\sqrt{(s^2_{NS}+s^2_S)/2}} \times 100$, where $\bar{x}$ and s^2 denote sample mean and sample variance, and the subscripts NS and S refer to nonsmokers and smokers respectively.

Covariates	Randomly picked		Propensity score		Distance		Distance within calipers of propensity score	
	p-Value	Std Diff %	*p*-Value	Std Diff %	*p*-Value	Std Diff %	*p*-Value	Std Diff %
Age	<0.0001	−25.5	0.0012	18.1	0.8273	−4.5	0.6791	−4.9
Race	0.0019	–	0.0003	–	0.6771	–	0.5315	–
BirthPlace	0.0033	19.1	<0.0001	−35.3	1	0.5	0.8074	1.5
Education	0.0940	10.4	0.0316	−0.5	0.6643	1.4	0.8525	1.7
Marital	0.2571	−7.8	0.3865	−17.0	0.9749	−4.4	0.9720	−4.0
HouseholdSize	0.0015	27.5	0.0036	−20.5	0.8577	0.0	0.7747	0.9
FMPIR	0.4186	5.4	0.0079	−1.8	0.4022	2.5	0.4316	2.7
Alcohol	<0.0001	−52.7	<0.0001	35.3	0.7310	−6.1	0.9340	−4.6
Drug	<0.0001	−59.6	<0.0001	55.9	0.6974	3.1	0.9738	2.3
HIC	0.4456	5.4	0.9955	−7.5	0.7817	−3.9	0.7936	−3.2
PHIC	0.1419	11.4	0.1240	−15.9	0.2360	−2.7	0.2534	−2.3
Owned	0.4626	5.1	0.4238	4.3	0.9092	8.3	0.9630	8.1
PHSAVG	0.2699	−7.1	0.6687	1.9	0.8359	1.6	0.8454	2.2
PHSVIG	0.0062	17.8	0.3059	−7.6	0.7178	4.1	0.6641	4.6
PHSMOD	0.2980	7.0	0.9490	0.8	0.6527	5.9	0.4022	10.6
PHSCOM	0.4835	1.4	0.3898	−1.4	0.8186	7.6	0.7456	9.2
PHS10YR	0.7653	3.7	0.9732	−22.4	0.3533	−4.0	0.4824	−2.1
SBP	0.1763	−8.7	0.2178	−4.2	0.8968	−9.7	0.9240	−8.4
DBP	0.1562	9.1	0.3929	−0.4	0.2263	0.8	0.2317	0.8

they are heavier (see Table 5). The engineered samples include 2237 smokers and 2280 non-smokers.

A random sample of 250 smokers and 250 non-smokers were selected from the engineered samples. These samples are not balanced, as can be seen from the *p*-values presented in the middle column of Table 4. Using the same 250 smokers, the distance based matching method was applied: for each smoker, the non-smoker with the smallest proximity to the smoker in the remaining available (unmatched) non-smoker reservoir is selected as its match. The two matched smoker and non-smoker groups are now balanced in terms of all 19 covariates, see the *p*-values in the last column of Table 4.

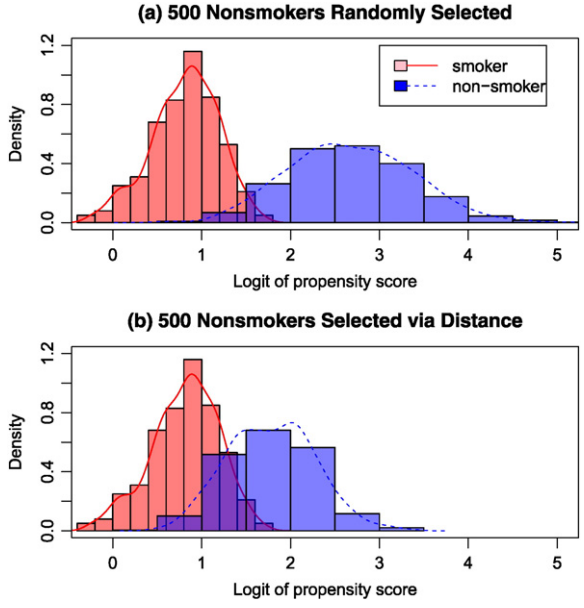

Fig. 1. Histograms and density curves of the logit of propensity scores for 500 smokers and 500 nonsmokers. The same smokers were used in both (a) and (b), while the nonsmokers were selected randomly in (a) and by distance in (b).

Presented in Table 5 are BMI comparisons based on three different sets of smokers and non-smokers. The first comparison was for the full engineered samples of 2237 smokers and 2280 non-smokers. By managing the distribution of age and SBP in the two groups, the mean BMI value is 27.6 kg/m^2 for smokers and 29.0 kg/m^2 for non-smokers. Since the CDC interprets BMI values of 25.0 to 29.9 as overweight, and BMI values of above 30.0 as obese, the average values of BMI among smokers and non-smoker both fall within the "overweight" range with the average BMI for non-smokers rather close to the obese range. The BMI values are also statistically significantly different between smokers and non-smokers. The next comparison is between two random samples of 250 smokers and 250 non-smokers. As can be seen from Table 5, the sample averages are similar to the full data averages, with the comparison again being statistically significant with a *p*-value of 0.006. The standardized differences for the two comparisons between

Table 4
Group comparisons before and after matching based on engineered samples. The prior-to-matching results are based on randomly selected samples of 250 smokers and 250 non-smokers. The after-matching results are based on the same 250 smokers, with 250 non-smokers selected by the nearest available distance.

Covariates	Prior to matching	After matching by distance
Age	<0.0001	0.7258
Race	0.2204	0.8606
BirthPlace	1	0.6744
Education	0.0324	0.2802
Marital	0.9132	0.8769
HouseholdSize	0.0094	0.4780
FMPIR	0.8647	0.1742
Alcohol	<0.0001	0.9156
Drug	<0.0001	0.1929
HIC	0.0065	0.9795
PHIC	0.0095	0.4355
Owned	0.0357	0.8815
PHSAVG	<0.0001	0.9189
PHSVIG	0.6229	0.3898
PHSMOD	0.1068	0.8580
PHSCOM	0.0029	0.9224
PHS10YR	0.3197	0.3245
SBP	<0.0001	0.7138
DBP	0.3328	0.5615

Table 5
The BMI comparison in engineered dataset (prior and after matching). The standardized difference in % is given by $\frac{\bar{x}_{NS}-\bar{x}_S}{\sqrt{(s^2_{NS}+s^2_S)/2}}\times 100$, where $\bar{x}$ and s^2 denote sample mean and sample variance, and the subscripts NS and S refer to nonsmokers and smokers respectively.

Dataset	BMI		*p*-Value	Std Diff in %
	Smoker	Nonsmoker		
Full engineered dataset, smoker (N = 2237) vs nonsmoker (N = 2280)	27.6	29.0	<0.0001	21.8
Randomly selected, smoker (N = 250) vs nonsmoker (N = 250)	27.5	29.1	0.0060	24.7
Matched by distance, smoker (N = 250) vs nonsmoker (N = 250)	27.5	28.0	0.3473	8.4

unmatched samples are 21.8% and 24.7%, respectively. The last comparison is based on the same 250 smokers, with the 250 non-smokers selected by the matching method based on proximity. The mean values of BMI in the matched samples are 27.5 and 28.0 for smokers and non-smokers respectively, with both of these means resting in the middle of the "overweight" range. The BMI comparison based on matched samples is no longer statistically significant, with a *p*-value of 0.35. The standardized difference for this last comparison between matched samples is 8.4%, indicating similar BMI values among smokers and non-smokers. This exercise demonstrates that, when the difference in outcome (BMI) is due to unbalanced samples, by producing balanced samples using the proposed method, the unreal effect of treatment (smoking) rightfully disappears and a direct comparison of the outcome is now valid based on matched samples.

4.3. Missing data investigation

The purpose of this subsection is to investigate how well the proposed method may be able to handle missing data. In order to have a reasonably large dataset with no missing covariates, four covariates with most missing data, *FMPIR*, *Drug*, *PHIC*, and *PHS10YR*, were excluded from the full dataset as summarized in Table 1. Then, all observations with any missing value in one of the 15 remaining covariates were excluded, resulting in a complete case dataset with 3505 smokers and 3531 nonsmokers. Next, three different datasets were generated, each with varying amount of missing data: up to 10%, 25%, and 40% missing data, respectively. Here is how a dataset with up to 10% missing data was generated. Two variables were randomly picked, one for smokers and one for non-smokers. For these variables, 10% of the data were replaced by NA's. In our simulations, 10% of smokers had PHSVIG missing and 10% of non-smokers had HIC missing. All the other variables had the missing rate uniformly distributed between 0 and 10%. Datasets with 25% and 40% missing data were generated similarly, that is, one variable had 25% (or 40%) missing data and all other variables had the rate of missing data randomly generated from a uniform distribution between 0 and 25% (or 40%). For each of these four datasets, with no missing data and rate of missing up to 10%, 25%, and 40% respectively, the matching method based on nearest available distance (or proximity) was applied: a sample of 500 cases was randomly selected from the 3505 cases, and 500 controls were selected based on proximity from the available reservoir of 3531 controls. The results are presented in Table 6. It can be seen that the random forest based matching method is able to find well balanced treatment groups for all these four datasets. In addition, the matching performance barely deteriorates with increasing missing rate: the smallest *p*-values comparing the two groups among the 15 covariates are 0.25, 0.47, 0.23, and 0.16 for the maximum missing rate of 0%, 10%, 25%, and 40%, respectively, all comfortably above the 0.10 threshold we use for balanced samples.

Table 6
Matching results with varying missing rates.

Covariate	Complete	Up to 10% missing	Up to 25% missing	Up to 40% missing
Age	0.6530	0.5836	0.9999	0.8776
Race	0.8350	0.5846	0.6473	0.9380
BirthPlace	0.9341	0.9792	0.6652	0.9227
Education	0.9807	0.8469	0.6292	0.7773
Marital	0.6057	0.8959	0.8662	0.6646
HouseholdSize	0.7798	0.4720	0.8825	0.9564
Alcohol	0.9294	0.7330	0.7440	0.2217
HIC	0.8750	0.8777	0.9779	0.1626
Owned	0.7346	0.7442	0.7103	0.5789
PHSAVG	0.7569	0.7557	0.2284	0.5594
PHSVIG	0.2492	0.4717	0.4931	0.6770
PHSMOD	0.3979	0.9475	0.9195	0.2980
PHSCOMP	0.4617	0.6859	0.4462	0.3963
SBP	0.3229	0.5338	0.9345	0.8959
DBP	0.7164	0.7805	0.4816	0.7423

5. Discussion

In this paper, random forest based matching methods are proposed to construct balanced treatment groups based on data from observational studies. Though the matching methods are similar to those discussed in Refs. [8,23], we propose to estimate both the propensity score and distance (proximity) by constructing a random forest with treatment as the output and all other covariates as inputs. The proposed methods were applied to the data from the National Health and Nutrition Examination Survey (NHANES), with the aim of producing balanced smoker and non-smoker groups in the study of body mass index (BMI). Our results show that the matching methods based on distance alone or on distance within calipers defined by the propensity score can produce well balanced treatment groups. On the other hand, the matching method based on propensity score alone yielded matching results that are less satisfactory. Among the many advantages (see Section 1) for using random forest based matching methods, the ease to handle missing data is particularly desirable for large observational datasets that often contain a considerable amount of missing data, such as the NHANES data. While the missing data have to be eliminated or imputed beforehand in the traditional logistic regression or discriminant analysis based matching methods, the proposed random forest based approach uses surrogate splits for handling missing covariates and hence can use all observations with some covariates missing. Our simulation studies show that the proposed methods can handle data with up to 40% missing data in covariates without much deterioration in matching results, compared to data without any missing values. In summary, the proposed random forest based matching methods may be used to produce balanced treatment groups for data from large observational studies and thus provide better estimation of the treatment effect.

Acknowledgments

This research was supported in part by the National Institutes of Health grant R01-DE019656.

Appendix A. Algorithm for calculating propensity score and proximity matrix based on random forest

Suppose that the total sample size including treated and control subjects is *n*, and that the total number of trees in the random forest is *B*. Let

$(S_1,\ldots,S_n)$ and $(D_{ij})_{n\times n}$ denote the propensity score vector and proximity matrix.

Initialize: set $S_i = 0$ and $D_{ij} = 0$ for $i, j = 1,\ldots,n$.

For $b = 1,\ldots,B$, do

- Using all data, grow a tree with treatment as output and all other covariates as inputs. At each split, search over m randomly selected inputs. No pruning.

- Propensity score calculation:
  - Compute the percentage of treated observations s_i for the terminal node to which the ith subject belongs.
  - Update: $S_i = S_i + s_i$

- Proximity matrix calculation:
  - Compute the proximity matrix $(d_{ij})_{n\times n}$ by assigning a value of 0 to the (i, j) cell if the ith and jth subjects are in the same terminal node and 1 otherwise.
  - Update: $D_{ij} = D_{ij} + d_{ij}$

Enddo.

Average: $S_i = S_i/B$ and $D_{ij} = D_{ij}/B$

References

[1] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[2] L. Breiman, J. Friedman, R. Olsen, C. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, California, 1984.
[3] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, Proceedings of the 23rd International Conference on Machine Learning 2006, pp. 161–168.
[4] R. Caruana, N. Karampatziakis, A. Yessenalina, An empirical evaluation of supervised learning in high dimensions, Proceedings of the 25th International Conference on Machine Learning 2008, pp. 96–103.
[5] H.N. Cham, Propensity Score Estimation with Random Forests(Ph.D. Dissertation) Arizona State University, 2013.
[6] A. Chiolero, D. Faeh, F. Paccaud, J. Cornuz, Consequences of smoking for body weight, body fat distribution, and insulin resistance, Am. J. Clin. Nutr. 87 (4) (2008) 801–809.
[7] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[8] R.B. D'Agostino, Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group, Stat. Med. 17 (1998) 2265–2281.
[9] Y. Ding, J.S. Simonoff, An investigation of missing data methods for classification trees applied to binary response data, J. Mach. Learn. Res. 11 (2010) 131–170.
[10] A. Hapfelmeier, T. Hothorn, K. Ulm, Recursive partitioning on incomplete data using surrogate decisions and multiple imputation, Comput. Stat. Data Anal. 56 (2012) 1552–1565.
[11] J.R. Hayes, J.I. Groner, Using multiple imputation and propensity scores to test the effect of car seats and seat belt usage on injury severity from trauma registry data, J. Pediatr. Surg. 43 (5) (2008) 924–927.
[12] Hastie, Tibshirani, Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2009.
[13] Y. He, Missing Data Imputation for Tree-Based Models(Ph.D. Dissertation) University of California at Los Angeles, 2006.
[14] J. Hill, Reducing bias in treatment effect estimation in observational studies suffering from missing data, Working Paper 04–01, Institute for Social and Economic Research and Policy, University, Columbia, 2004.
[15] B.K. Lee, J. Lessler, E.A. Stuart, Improving propensity score weighting using machine learning, Stat. Med. 29 (2010) 337–346.
[16] A. Liaw, Classification and regression by randomForest, R News 2 (2002) 18–22.
[17] P. McCulloch, S. Lee, R. Higgins, K. McCall, D.S. Schade, Effect of smoking on hemoglobin A1c and body mass index in patients with type 2 diabetes mellitus, J. Investig. Med. 50 (4) (2002) 284–287.
[18] P.C. Mahalanobis, On the generalised distance in statistics, Proc. Natl. Inst. Sci. India 2 (1) (1936) 49–55.
[19] R. Mitra, J.P. Reiter, A comparison of two methods of estimating propensity scores after multiple imputation, Stat. Methods Med. Res. (2006)http://dx.doi.org/10.1177/0962280212445945.
[20] A. Rieger, T. Hothorn, C. Strobl, Random forests with missing values in the covariates, Technical Report # 79, Department of Statistics, University of Munich, 2010 (Available at http://epub.ub.uni-muenchen.de/11481/1/techreport.pdf).
[21] P.R. Rosenbaum, D.B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 4155.
[22] P.R. Rosenbaum, D.B. Rubin, Reducing bias in observational studies using subclassification on the propensity score, J. Am. Stat. Soc. 79 (1984) 516–524.
[23] P.R. Rosenbaum, D.B. Rubin, Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, Am. Stat. 39 (1) (1985) 33–38.
[24] S. Setoguchi, S. Schneeweiss, M.A. Brookhart, R.J. Glynn, E.F. Cook, Evaluating uses of data mining techniques in propensity score estimation: a simulation study, Pharmacoepidemiol. Drug Saf. 17 (2008) 546–555.
[25] E.A. Stuart, Matching methods for causal inference: a review and a look forward, Stat. Sci. 25 (1) (2010) 1–21.
[26] X.G. Su, J. Kang, J. Fan, R. Levine, X. Yan, Facilitating score and causal inference trees for large observational data, J. Mach. Learn. Res. 13 (2012) 2955–2994.
[27] D. Westreich, J. Lessler, M.J. Funk, Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression, J. Clin. Epidemiol. 63 (2010) 826–833.