An interactive web application for propensity score matching with R shiny; example of thrombophilia

Emre Demir[1], Serdal Kenan Kose[2], Omer Faruk Akmese[3], Engin Yildirim[4]

[1]Hitit University, Faculty of Medicine, Department of Biostatistics, Corum, Turkey
[2]Ankara University, Faculty of Medicine, Department of Biostatistics, Ankara, Turkey
[3]Hitit University, Osmancik Omer Derindere Vocational School, Department of Computer Technologies, Corum, Turkey
[4]Hitit University, Faculty of Medicine, Department of Gynecology and Obstetrics, Corum, Turkey

Abstract

Aim: The aim of this study was to develop a new web-based R Shiny package that calculates propensity score using many algorithms such as logistic regression, machine learning, and performs matching analysis with balance evaluation. In addition, it was aimed to explain the process of matching analysis on a real data set by comparing the number of live births between those with methylenetetrahydrofolate reductase (MTHFR) homozygous mutations and those without mutations in women hospitalized due to abortion in the gynecology and obstetrics clinic.
Material and Methods: The web-based application was developed using R shiny. The "matchIt" library was used for matching analysis and PS prediction. The "cobalt" library was used to evaluate balance and generate plots.
Results: The abortion variable, which was statistically significantly different in the groups before matching (p=0.010), was similar in the groups after matching (p=0.743). In addition, when the descriptive statistics and p values of the other variables were examined, it was seen that almost full balance was achieved after matching and the confounder variables were similar distributed in groups. After matching analysis, it was determined that the result variable "livebirths" did not show statistically significant difference in the groups (p=0.864).
Conclusion: In this study, we developed an interactive web application for matching analysis based on propensity score. It is thought that this application will facilitate the studies of the researchers.

Keywords: Propensity score; matching; logistic regression; machine learning; R shiny

INTRODUCTION

Randomization is performed in order to ensure the random assignment of patients to experimental groups, to prevent bias in clinical trials, and to ensure similar distribution of the groups in terms of many covariates such as age, gender, body mass index (BMI) (1-3). In experimental studies where there is an effect of another confounding variable other than the main effect investigated in case and control groups, confounding variables lead to biased results. In scientific studies, while examining causality relationships between variables, accurate presentation of findings can only be achieved by eliminating the effects of confounder variables (4,5).

In clinical trials conducted to determine cause-effect relationships, it is assumed that bias is prevented by ensuring a balanced distribution of the characteristics of the patient (age, gender, BMI, etc.) and disease-related (stage, family history, etc.) in the experimental and control groups before randomization (3). In non-randomized observational or quasi-experimental clinical trials investigating the effect of a specific factor, case-control groups do not have similar characteristics in terms of many covariates/factors. Therefore, bias may be found in statistical test results showing the effect of treatment/exposure where the unbalanced confounding variable is associated with the dependent variable being investigated (6-8).

In case of imbalance in the groups, models were developed in which the effects of the confounder variables on the dependent variable were purified and the pure effect of the treatment could be determined. Matching methods based

on PS estimation reduce the bias of observational studies (7). The aim of the matching was to select individuals from the control group who had similar characteristics to the covariates values such as age, gender and BMI of the individuals in the treatment group. When the number of covariates is high, PS gives more successful results than traditional methods (stratification, restriction, covariance analysis) because it reflects the summary measure of a large number of covariate information (4,7,8). Therefore, in clinical trials where randomization cannot be performed, matching of case and control groups according to PS is a method that should be taken into consideration in research to reduce bias and to determine the actual effect of treatment by reducing the effect of confounder factors (8). D'agostino (1998) stated that a quasi-randomized trial could be created when matching analysis based on the probability of being treated (propensity score) was used to correct the estimate of the exposure effect in observational studies (8).

In order to examine the importance of the Propensity score (PS) in the observational studies, a literature review was conducted using the Web of Science database (Accessed 01.10.2019). As a result of the literature search in the title, abstract and keywords of the articles by using "propensity score" and "propensity-score" keywords, 21184 publications were found between 1975 and 2018. 19015 of these publications were indexed in the article category.

When the 19015 articles are evaluated by country, the top ten countries that use PS the most frequently were: USA (9359), China (1324), Japan (1303), Germany (1289), Canada (1264), South Korea (1213), England (1156), Italy (1080), Taiwan (938) and France (820). Almost half (49.2%) of the articles published on PS were from USA authors. According to the rankings between countries using the PS it was most Turkey ranks 31 with 76 articles. When the distribution of publications by years was examined, it was seen that 13135 (69.1%) of 19015 articles were published in the last five years (2014-2018).

The research areas where PS is widely used are shown in Table 1. In medical research, PS has been widely used in Cardiovascular System Cardiology with 2950 (15.5%) publications and Surgery with 2600 (13.6%) publications. According to the findings of literature review, matching analysis based on PS widely used in many medical research fields and it is an important method that is increasing in use day by day. However, the fact that PS is not used adequately in the studies conducted in our country compared to the international usage shows the importance of this article subject and the developed application (9).

In this study, we aimed to develop a new interactive web-based R Shiny package that can perform matching analysis by calculating propensity score using many algorithms such as logistic regression and machine learning and can perform statistical and graphical balance evaluation after matching. The sample for matching analysis was selected among those with methylene tetrahydrofolate reductase homozygote mutation and non-mutation in women hospitalized due to abortion (involuntary abortion) in the gynecology and obstetrics clinic. The outcome variable in the study was the number of live births and was compared between women with mutations and healthy volunteers. The similar structure of the groups in terms of age, gravida, parita, abortion number, IUGR and Precclampsia will allow determining the importance of the mutation in low etiopathogenesis and providing appropriate medication.

MATERIAL and METHODS

Rosenbaum and Rubin (1983) showed that PS can be estimated based on the covariates observed using a logit model. The dependent variable; treatment variable and independent variable; the covariates; Rosenbaum and Rubin (1983) defined PS as the conditional probability of being assigned to a particular treatment group based on observed covariates. To illustrate the treatment (Ti = 1) and control group (Ti = 0), the probability of assigned treatment group based on the covariates (x) is expressed as in the following equation (2,9,10).

$$P(T_i = 1 / X_1, \ldots, X_P) = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P)}$$

After this study, logit model (Logistic regression analysis) was widely used in PS calculations in health studies. In recent years, the use of Machine Learning (ML) methods has been proposed as an alternative to Logistic Regression (LR), the classic method for PS prediction, in models with interaction effects and nonlinear relationships between the probability of assigning subjects to treatment groups and the covariates (9-14).

PS is a summary score based on covariates given to individuals. Matching is a sampling method used to identify the group that is most similar to the covariate distribution of the treatment group among potential controls (8). In the treatment and control group, two subjects with the same PS score may considered to be assigned randomly to each group with equal probability. Thus, the critical assumption regarding the independence of the treatment assignment mechanism is theoretically satisfied (8).

In the matching analysis based on PS, the j th observation with a propensity score Pj in the control group matches an observation with a propensity score Pi in the treatment group. Here, Pi and Pj, which have the smallest difference in propensity scores, are selected. In other words, it is attempted to match the patient with the control that is closest to the confounder variable properties (9,10).

The steps of the matching analysis can be listed as follows;

1. Determination of the covariates/factors that need to achieve balance in groups

2. Prediction of PS to be used in matching analysis and performing matching using an appropriate matching method

3. Evaluation of matching balance (Standardized Mean Differences (SOF), Variance Ratios (VO), Kolmogorov-Smirnov (KS) Test Statistics, Graphical approach) according to covariates/factors

4. Estimating treatment effect from a matched data set (15).

The developed web application is available by internet at:"http://propensityscorematching.shinyapps.io/propensityscorematching". In the application developed using R shiny (R Core Team, 2019), "MatchIt" library developed by Ho et al. (2007) was used for matching analysis and PS prediction (16). In the present study, the nearest neighbor matching method based on PS was used for matching analysis. Graphic drawings used for balance evaluation were created by using "cobalt" library (17). Using the application, we developed, all processes of matching analysis on a sample set of thrombocytes from Gynecology and Obstetrics were explained in the results section and the findings were reported. Data set were collected from patients who were followed up during pregnancy at the Pregnant Policlinic of Hitit University Faculty of Medicine, Department of Obstetrics and Gynecology and ethical approval was obtained from Hitit University Non-Interventional Research Ethics Committee.

RESULTS

The sample data set for thrombophilia which will be used in the application is presented in Figure 1. In addition, variable properties are defined in Table 2. In the variable "Groups" painted with orange in Figure 1, "1" shows women with a homozygous MTHFR mutation in women lying due to an abortion, while "0" shows women without a mutation. The variable "Live Births" painted with orange is the outcome variable (the main hypothesis of the study), and whether these two groups differ significantly will be investigated. Age, Gravida, Parita, Abortion, IUGR and Precclampsia variables painted with yellow color are variables that are wanted to be distributed similarly in groups and will be used in matching analysis.

The commands for opening the data set in the R shiny interactive web application were shown in Figure 2. When the commands in Figure 2 are applied, the data set opened in the R shiny interactive web application will appear as below the Figure 2.

Descriptive statistics (mean, standard deviation, median, minimum, maximum, 1st and 3rd quarter values) of the variables in the data set are calculated before matching analysis is performed in the web application. The application also provides a summary of the structure of the variables in the menu section Figure 2.

PS for matching analysis in the web application is calculated by Logistic regression analysis by default.

Figure 1. The sample data set for research on thrombophilia (Gravida: Total number of pregnancies; Parity: Number of births; Abortion: Number of abortions; IUGR: Intra uterine growth restriction; Preeclampsia: High blood pressure and proteinuria in pregnancy; Live births: Number of live births)

Figure 2. Screenshot of opening data set in R shiny interactive web application

492

Table 1. Research area categories where propensity score is widely used					
Research Area Categories	**Number of articles**	**% (percentage of 19015)**	**Research Area Categories**	**Number of articles**	**% (percentage of 19015)**
Cardiac Cardiovascular Systems	2950	15.5	Radiology Nuclear Medicine Medical Imaging	280	1.4
Surgery	2600	13.6	Anesthesiology	272	1.4
Economics	1673	8.7	Immunology	253	1.3
Oncology	1508	7.9	Orthopedics	250	1.3
Public Environmental Occupational Health	1326	6.9	Transplantation	249	1.3
Medicine General Internal	1298	6.8	Pediatrics	240	1.2
Respiratory System	1173	6.1	Obstetrics Gynecology	236	1.2
Health Care Sciences Services	948	4.9	Medical Informatics	235	1.2
Statistics Probability	727	3.8	Environmental Studies	217	1.1
Pharmacology Pharmacy	675	3.5	Criminology Penology	213	1.1
Gastroenterology Hepatology	649	3.4	Management	213	1.1
Urology Nephrology	648	3.4	Geriatrics Gerontology	208	1.0
Peripheral Vascular Disease	615	3.2	Development Studies	206	1.0
Health Policy Services	605	3.1	Hematology	205	1.0
Clinical Neurology	591	3.1	Sociology	199	1.0
Multidisciplinary Sciences	521	2.7	Microbiology	184	0.9
Critical Care Medicine	448	2.3	Business	174	0.9
Psychiatry	404	2.1	Rheumatology	171	0.8
Medicine Research Experimental	386	2.0	Social Work	165	0.8
Education Educational Research	348	1.8	Neurosciences	164	0.8
Mathematical Computational Biology	319	1.6	Psychology Developmental	158	0.8
Infectious Diseases	318	1.6	Environmental Sciences	157	0.8
Business Finance	315	1.6	Family Studies	150	0.7
Social Sciences Mathematical Methods	305	1.6	Agricultural Economics Policy	149	0.7
Endocrinology Metabolism	292	1.5	Psychology Clinical	142	0.7

Table 2. The sample data set for research on thrombophilia			
Variables	Variable types/ Levels of Measurements	Values	Roles
Groups	Categorical / Nominal	1= Homozygous, 2=Control	Experimental groups/Treatment variable
Age	Numeric / Continuous / Ratio	Years	Covariate /Confounder
Gravida	Numeric / Discrete / Ratio	Total number of pregnancies	Covariate /Confounder
Parity	Numeric / Discrete / Ratio	Number of births	Covariate /Confounder
Abortion	Numeric / Discrete /Ratio	Number of abortions	Covariate /Confounder
IUGR	Categorical / Nominal	Intra uterine growth restriction 1= Present 0=Absent	Covariate /Confounder
Preeclampsia	Categorical / Nominal	High blood pressure and proteinuria in pregnancy 1= Present 0=Absent	Covariate /Confounder
Live Births	Numeric / Discrete /Ratio	Number of live births	Outcome/ Dependent variable

However, using the "Select Algorithm" option, matching analysis can be performed by using some machine learning algorithms such as artificial neural networks, some other methods such as quadratic discriminant analysis, multivariable adaptive regression curves (MARS) and combined estimator algorithms which have been shown to be effective in reducing bias in observational studies (9,10). After selecting the algorithm in Figure 3, the variable that specifies the groups to be matched (case-control) should be selected using "select dependent variable" option. After selecting this variable, it is necessary to select the confounding variables which are to be distributed similar between case-control groups. In this case study, we will investigate whether "Live Births", the result variable to be compared between case and control groups, is significantly different between the groups. In this design, Age, Gravida, Parita, Abortion, IUGR and Preeclampsia variables shown in Figure 3 were chosen as the confounding variables.

Once the selections shown in Figure 3 are made, the application will display the matching results under the "Model menu" (Figure 4). After the matching analysis, both the statistics of the whole unpaired data set and
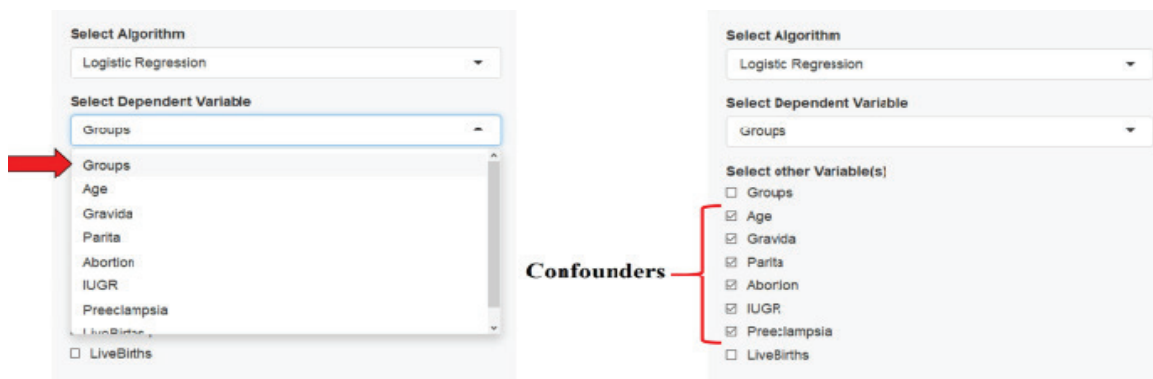

Figure 3. Screenshot in R shiny interactive web application for selection of algorithm for propensity score calculation, selection of variables showing groups, and selection of independent variables that are desirable to be distributed similarly in groups
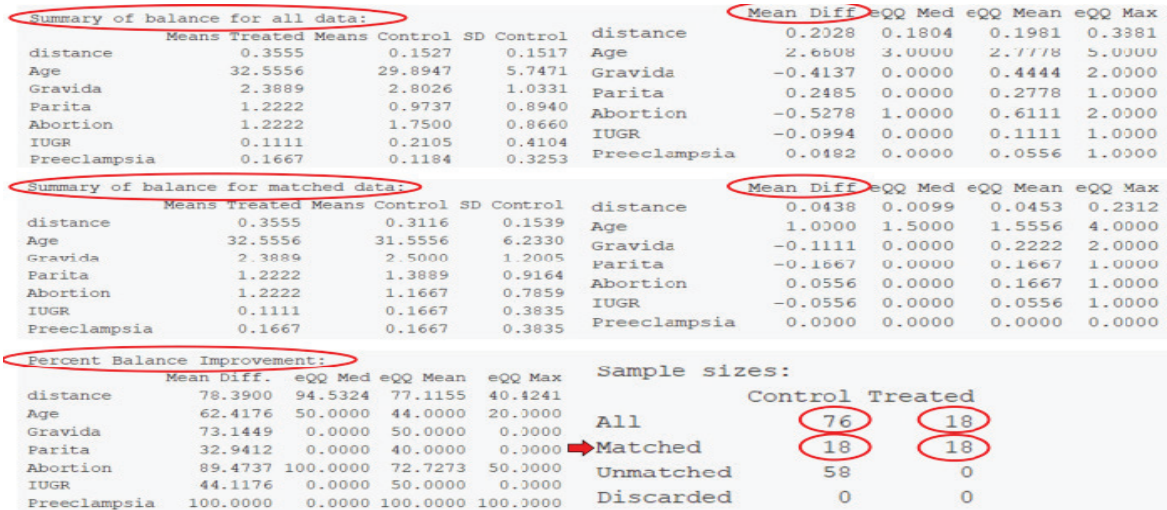
Figure 4. Screenshot of balance results after matching analysis based on propensity score in the R shiny interactive web application

the statistics of the matched data set are calculated. In addition, the results show how the mean differences for each confounding variable are improved as a percentage. Thus, the interpretation of the matching balance can be made. For example, before the matching analysis, the mean age of the case group was 32.55 and the mean age of the control group was 29.89, while the mean of the control group after matching was 31.55 and approached the case group. It is seen that the mean differences decreased from 2,6608 before matching to 1.00 after matching in the Figure 4. In addition, 62.41% improvement in balance was observed as a result of matching for the age variable. In the preeclampsia variable, the balance was 100% and the mean difference in the case control group was 0 (Figure 4).

The screenshot showing the download of the new data set as a result of the matching analysis in R shiny is shown in Figure 5. After the matching analysis, the new data set can be downloaded with the "download" option on the bottom left side under the "model" menu. After downloading the match data set, users will be able to perform statistical analysis by opening the data set in a statistical package program such as SPSS. In addition, in Figure 5, the PS score of each individual is calculated under the heading "distance". When the figure is examined, the confounding variable values and PS values of 18 patients and 18 controls matching the patients are shown.
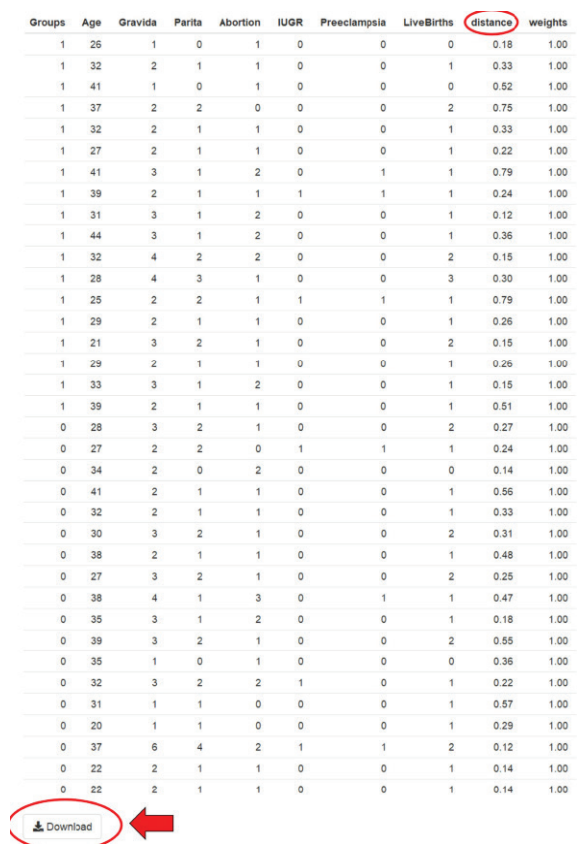
The screen shot of the balance evaluation of categorical variables and continuous variables obtained as a result of the matching analysis in R shiny interactive web application is shown in Figure 6. Figure 6.a (left) shows the distribution of Preeclampsia, a categorical variable, in groups before matching analysis. Figure 6.a (right) shows the distribution of Preeclampsia variable in groups after matching analysis. When the figure is examined, it is seen that the preeclampsia proportions before the matching analysis are not equal between the groups, but after the matching analysis they become exactly equal.

Figure 5. Screenshot showing the download of the new data set after matching analysis in the R shiny interactive web application.

In Figure 6.b, the overlap of the distribution line for a continuous variable will show that the balance between the groups is fully achieved.

An important balance assessment test between the groups; Kolmogorov-Smirnov test statistics, variance ratios and mean differences are compared in the "BalPlots for Balance" menu is shown in Figure 7.

Figure 6. Screenshot of the balance evaluation of categorical (a) and numerical (b) variables after matching analysis in R shiny interactive web application.

Table 3. Statistical results before and after matching analysis						
	Before matching			**After matching**		
	Control (n=76)	Homozygous (n=18)	P values	Control (n=18)	Homozygous (n=18)	P values
Age	29.89±5.75	32.56±6.37	0.087[a]	31.56±6.23	32.56±6.37	0.637[a]
Gravida	2.8±1.03	2.39±0.85	0.113[b]	2.5±1.2	2.39±0.85	0.938[b]
	3 (1-6)	2 (1-4)		2 (1-6)	2 (1-4)	
Parity	0.97±0.89	1.22±0.73	0.200[b]	1.39±0.92	1.22±0.73	0.628[b]
	1 (0-4)	1 (0-3)		1 (0-4)	1 (0-3)	
Abortion	1.75±0.87	1.22±0.55	**0.010[b]**	1.17±0.79	1.22±0.55	0.743[b]
	2 (0-4)	1 (0-2)		1 (0-3)	1 (0-2)	
IUGR	16/76	2/18	0.509[c]	3/18	2/18	1.000[c]
Preeclampsia	9/76	3/18	0.694[c]	3/18	3/18	1.000[c]
LiveBirths	0.89±0.83	1.17±0.71	0.149[b]	1.17±0.62	1.17±0.71	0.864[b]
	1 (0-3)	1 (0-3)		1 (0-2)	1 (0-3)	
[a]: Student's t test (mean±SD) b: Mann Whitney U test (mean±SD and median (min-max)) [c]: Fisher exact test (n)						

The fact that the KS statistic is close to 0 indicates that the distribution of the treatment and control groups according to covariates is consistent and the balance is achieved. The fact that the mean differences are 0 indicates that the groups are similar and the balance is achieved. The ratio of variance close to 1 indicates that the balance is achieved between the groups. (9,10). Figure 7 shows that the before matching (blue color) Kolmogorov-Smirnov test statistics are far from 0, while after matching they are closer to 0.

When the figure of mean differences is examined, it is seen that after the matching, the differences (red color) are distributed around 0 and the balance is achieved.

The findings before and after the matching analysis are given in Table 3. When the table was examined, the abortion variable, which was statistically significantly different in the groups before matching (p=0.010), was similar in the groups after matching (p=0.743). In addition, when the descriptive statistics and p values of the other variables

were examined, it was seen that almost full balance was achieved after the matching and the confounding variables were similar in the groups. After the matching analysis, the result variable "live births" was not significantly different between the groups (p=0.864).

When the clinical results are evaluated, it can be said that the MTHFR homozygous mutations does not have a significant effect on the number of live births. However, prospective studies with larger sample sizes will provide more reliable results for this effect.
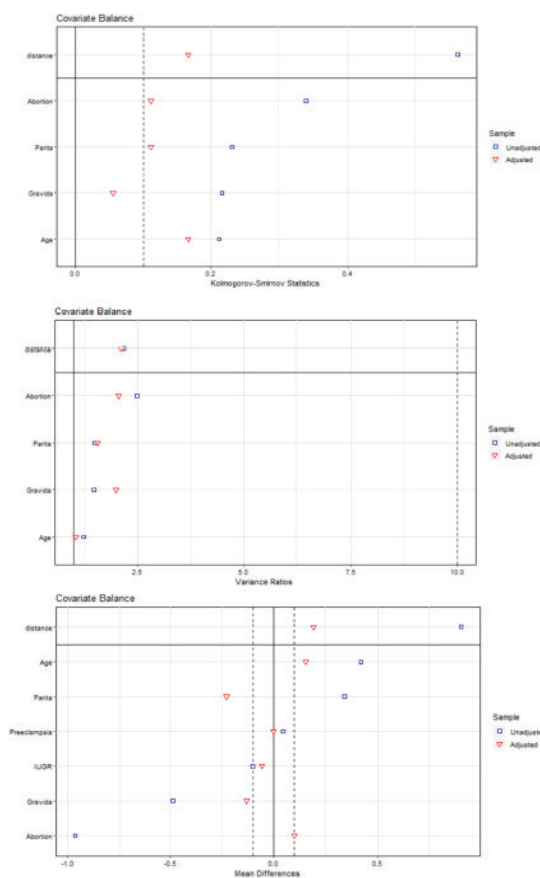
Figure 7. Screenshot of the balance evaluation of variables after matching analysis in the R shiny interactive web application.

DISCUSSION

The cost of conducting a randomized clinical trial and the difficulty of obtaining an ethical approval directs researchers to low-cost observational studies. In observational studies, after the matching analysis, the difference between treatment and control group means or treatment effect varies according to the choice of matching control group. To reduce bias in observational studies, matching analysis based PS is widely used, especially in statistical analysis in the medical field. In this study; An interactive web application was developed in which researchers can easily perform matching analysis based on the PS and balance assessment, and the application steps on a sample data set for thrombophilia were shown

with examples. In addition, sample research findings were reported and presented.

In the literature, matching analysis could be performed with a limited number of programs such as "matching" and "matchlt" in the 'R package' until the last years (16,18). However, medical doctors or health researchers had difficulty in matching analysis since these packages required coding knowledge. We believe that matching analysis is not widely used in Turkey due to the lack of an easy-to-use package program. In recent years, matching analysis can be done in new versions of the SPSS package. However, because the SPSS package is paid and many universities are still using the old versions, there is a need for a new package program that provides free and easy-to-use matching analysis. The superiority of this package is that it is offered to researchers free of charge and the balance assessments can be performed easily. In addition, our application is the first application in the literature that can match with R shiny.

CONCLUSION

In experimental research, the factors that may affect the outcome of the experiment, other than the factor that is investigated, are attempted to be controlled by randomization by the researcher. Therefore, cause-effect relationship can be determined accurately. However, confounder variables are often found in non-randomized observational studies. If these variables are ignored, the results of the experiments will be biased. Therefore, in clinical trials where randomization cannot be performed, matching of case and control groups according to PS is an important method to be taken into consideration in studies to reduce bias and determine the actual effect of treatment. In this study, an interactive web application was developed for matching analysis based on PS and the usage of the application was explained with a sample data set. It is thought that this application will facilitate the studies of the researchers.

Emre Demir ORCID: 0000-0002-3834-3864
Serdal Kenan Kose ORCID: 0000-0002-6734-3847
Omer Faruk Akmese ORCID: 0000-0002-5877-0177
Engin Yildirim ORCID:0000-0001-7937-4141

REFERENCES

1. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educational Psychology 1974;66:688-701.

2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.

3. Rosenberger WF, Lachin JM. Randomization in clinical trials: theory and practice. New York: John Wiley 2002; 1-133

4. Kaspar EÇ, Bekiroğlu N, Genceli M. Propensity score in observational studies and an application in medical sciences. Turkiye Klinikleri J Biostat 2010;2:1-10.

5. Gökmen D, Alkan A, Bakırarar B, et al. Bilimsel araştırma yöntemleri. Ankara: Ankara Üniversitesi Basımevi 2018;1-61.

6. Hill HA, Kleinbaum DG. Bias in observational studies. In: Armitage P, ed. Encyclopedia of Biostatistics. 2nd edition. Chichester: John Wiley and Sons 2005;323-30.

7. Hoffmeister H, Szklo M, Thamm M. Bias in observational studies. Epidemiological Practices in Research on Small Effects. Berlin: Springer 1998;59-60.

8. D'agostino RB JR. Tutorial in biostatistics: propensity score methods for bias reduction in the comparıson of a treatment to a non-randomized control group. Statistics in Medicine 1988;17:2265-81.

9. Demir E. Development of new propensity score estimation models with machine learning algorithms for optimal matching analysis in non-randomized clinical trials. Ph.D. dissertation, Ankara University, Ankara 2019.

10. Demir E, Köse SK. Development of new propensity score estimation models with machine learning algorithms for optimal matching analysis in non-randomized clinical trials and an interactive web application with R shiny. XXI. National and IV. International Biostatistics Congress 2019;149-62.

11. Mccaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods 2004;9:403-25.

12. Setoguchi S, Schneeweiss S, Brookhart MA, et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and drug safety 2008;17:546-55.

13. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statistics in Medicine 2010;29:337-46.

14. Pirracchio R, Petersen M, Van Der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. American Journal of Epidemiology 2014;181:108-19.

15. Stuart EA. Matching methods for causal ınference: a review and a look forward. Statistical Science 2010;25:1-21.

16. Ho D, Imai K, King G, et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 2007;15:199-36.

17. Greifer, N. Covariate balance tables and plots: A guide to the cobalt package. https://mran.microsoft.com/snapshot/2017-11-12/web/packages/cobalt/vignettes/cobalt_basic_use.html, access date 01.07.2019.

18. Sekhon, JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. Journal of Statistical Software 2011;42:1-52.