Biomedical Statistics and Informatics
2021; 6(2): 32-41
http://www.sciencepublishinggroup.com/j/bsi
doi: 10.11648/j.bsi.20210602.12
ISSN: 2578-871X (Print); ISSN: 2578-8728 (Online)

Bias Adjustment Methods for Analysis of a Non-randomized Controlled Trials of Right Heart Catheterization for Patients in ICU

Yi Xu, Yeqian Liu

Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA

Email address:
Yeqian.liu@mtsu.edu (Yeqian Liu)

To cite this article:
Yi Xu, Yeqian Liu. Bias Adjustment Methods for Analysis of a Non-randomized Controlled Trials of Right Heart Catheterization for Patients in ICU. *Biomedical Statistics and Informatics*. Vol. 6, No. 2, 2021, pp. 32-41. doi: 10.11648/j.bsi.20210602.12

Received: June 21, 2021; **Accepted**: July 7, 2021; **Published**: July 19, 2021

Abstract: Kaplan-Meier estimate or proportional hazards regression is commonly used directly to estimate the effect of treatment on survival time in randomized clinical studies. However, such methods usually lead to biased estimate of treatment effect in non-randomized or observational studies because the treated and untreated groups cannot be compared directly due to potential systematical difference in baseline characteristics. Researchers have developed various methods for adjusting biased estimates by balancing out confounding covariates such as matching or stratification on propensity score, inverse probability treatment weighting. However, very few studies have compared the performance of these methods. In this paper, we conducted an intensive case study to compare the performance of various bias correction methods for non-randomized studies and applied these methods to the right-heart catheterization (RHC) study to investigate the impact of RHC on the survival time of critically ill patients in the intensive care unit. Our findings suggest that, after bias adjustment procedures, RHC was associated with increased mortality. The inverse probability treatment weighting outperforms other bias adjustment methods in terms of bias, mean-squared error of the hazard ratio estimators, type I error and power. In general, a combination of these bias adjustment methods could be applied to make the estimation of the treatment effect more efficient.

Keywords: Confounder, Right Heart Catheterization, Propensity Score, Proportional Hazards Model, Kaplan-Meier Estimate, Non-randomized Study

1. Introduction

In randomized clinical studies, the effect of treatment on patients' survival time can be estimated by comparing treated and untreated subjects directly. In this case, Kaplan-Meier estimate or proportional hazards regression is used directly to estimate the effect of treatment on survival time. However, it is not easy to materialize a randomized study in daily life. There is an increasing number of nonrandomized studies in recent years. In an observational (or nonrandomized) study, the treated and untreated groups cannot be compared directly because they may systematically differ at baseline characteristics. For example, the patients' health condition and medical history are essential factors when doctors make a diagnosis. The treatment assignment to a patient is dependent on covariates like age, gender, health condition, and medical history, etc. As a result, the effect of medical treatment on patients' survival time may be confounded by their baseline covariates. Therefore, systematic differences in baseline characteristics between the treated and untreated groups must be considered in assessing the impact of treatment on survival time in observational studies.

The propensity score plays an important role in balancing the treated and untreated subjects to make them comparable. Rosenbaum and Rubin proposed that propensity score is the conditional probability assignment to a particular treatment given a vector of observed covariates [1-2]. They indicated that adjustment for the scalar propensity score contributes to control all confounders and eliminate bias due to observed covariates. Propensity Score is a scalar function of the covariates that includes the information required to achieve the balance of distribution of baseline covariates. The most common methods based on propensity score are matching, stratification, regression adjustment, and probability weighting [3-4]. With the

application of the propensity score, the treated and untreated patients who have similar propensity scores will have a similar distribution of observed background covariates. Therefore, the effect of treatment will be unrelated to confounders, as a result of which, treated and the untreated subject is comparable like what we could attain in randomized studies.

The dataset that motivated this paper pertains to day 1 of hospitalization and the treatment variable "swang1" is whether or not a patient received a Right Heart Catheterization (RHC), also called the Swan-Ganz catheter, on the first day in which the patient qualified for the SUPPORT study [5]. RHC is a test used to see how well your heart is pumping (how much blood it pumps per minute) and to measure the pressures in the heart and lungs. In an RHC, the doctor guides a special catheter (a small, hollow tube) to the right side of the heart then passes the tube into the pulmonary artery. The doctor observes blood flow through the heart and measures the pressures inside the heart and lungs. A sensitivity analysis provided some evidence that patients receiving RHC had decreased survival time. But the sensitivity analysis indicated that any unmeasured confounder would have to be somewhat strong to explain away the results. Our goal is to estimate the effect of RHC treatment on the patients' survival time after reducing the confounding bias. However, systematic differences in patients in the two groups may exist, and these differences could lead to a biased estimate of treatment effect; which is known as the causal effect in a nonrandomized study.

The RHC dataset includes the treatment variable "swang1". Denote the "treatment 0" as not receiving RHC and "treatment 1" receiving RHC. The observed time and censored indicator of each patient could be indicated from the variable "dthdte", which means "date of death", and variable "lstctdte" which represents "date of last contact". The patients with "NA" in "date of death" are recognized as censored and otherwise, they are uncensored. The observed time of each censored patient in the study is determined by the "date of last contact" and the observed time of each uncensored patient in the study is determined by the "date of death". Besides, there are 50 covariates included in the dataset with information of 5735 patients: "age", "sex", "race", "edu"(years of education), "income", "ninsclas"(medical insurance), "cat1"(primary disease category), "das2d3pc"(Duke activity status index), "dnr1"(DNR status on day1), "ca"(cancer), "surv2md1"(Support model estimate of the prob. of surviving 2 months), "aps1"(APACHE score), "scoma1"(Glasgow Coma Score), " wtkilo1"(weight), "temp1"(temperature), "meanbp1"(mean blood pressure), "resp1"(respiratory rate), "hrt1"(heart rate), "pafi1" (PaO2/FIO2 ratio), "paco21" (PaCo2), "ph1"(PH), "wblc1"(WBC), "hema1"(hematocrit), "sod1"(sodium), "pot1"(potassium), "crea1"(creatinine), "bili1"(bilirubin), "alb1"(albumin), and categories of admission diagnosis: "resp", "card", "neuro", "gastr", "renal", "meta", "hema", "seps", "trauma", "ortho"; categories of comorbidities illness: "cardiohx", "chfhx", "dementhx", "psychhx", "chrpulhx", "renalhx", "liverhx", "gibledhx", "malighx", "immunhx", "transhx", "amihx".

As mentioned before, the distributions of baseline covariates between treatment 0 and treatment 1 subjects are quite different. What's more, as we will see in the matching methods, the distributions of the propensity score in the two treatment groups are different, which reveals the systematic difference in the two studies and the problem of confounding. The remainder of the article focus on the application and comparison of the following three methods. Section 2 introduce matching on propensity score method. Section 3 introduce stratification on propensity score method. Section 4 introduce inverse probability treatment weighting method. We apply each method to the Right Heart Catheterization study to compare the survival time of RHC treated group and control group. The article concludes with a discussion on the choice of methods under different scenarios in Section 5.

2. Matching on Propensity Score

The propensity score is presented in both randomized trials and observational studies. In randomized trials, the true propensity score is known and defined by the study design. In observational studies, true propensity scores are generally not known but can be estimated through the data [6]. The propensity score is the conditional probability assignment to a particular treatment given a vector of observed covariates [1]:

$$e(X_i) = Pr(Z_i = 1|X_i)$$

Where the dependent variable is binary, Z_i=1 is associated with the RHC treatment and Z_i=0 is corresponding to control. X_i, i=0, 1 are observed values of the vector of covariates [6].

Propensity scores are generally calculated using one of two methods: Logistic regression or Classification and Regression Tree Analysis [6]. In practice, the propensity score is most often estimated using a logistic regression model, in which treatment status is regressed on observed baseline characteristics [7]. The estimated propensity score is the predicted probability of treatment derived from the fitted regression model [8].

$$ln\frac{e(X_i)}{1-e(X_i)} = ln\frac{Pr(Z_i = 1|X_i)}{1-Pr(Z_i = 1|X_i)} = \alpha + \beta^T X_i$$

Where the parameters α, β are estimated by maximum likelihood logistic regression.

Matching is a commonly used method to select "matched" pairs on background covariates that we believe need to be controlled. Even though it seems difficult to find patients who are similar on all important covariates, especially when there are plenty of covariates of interest, propensity score matching solves this problem by allowing us to control for as many covariates as we want simultaneously by matching a single scalar variable [9]. Rosenbaum and Rubin introduced three techniques for constructing a matched sample: (i) nearest available matching on the estimated propensity score; (ii) Mahalanobis metric matching including the propensity score; and (iii) nearest available Mahalanobis metric matching within calipers defined by the propensity score.

Therefore, once the propensity scores are estimated by the logistic regression method, we apply the nearest available matching approach to reduce the confounding bias in the RHC study. In this method, the absolute difference between the estimated propensity scores for the control and treated groups is minimized [6]. Given randomly ordered control and treated subjects, the first treated subject is selected along with a control subject with a propensity score closest in value to it [10]. Generally, if a treated subject and a control subject have the same propensity score, the observed covariates are automatically controlled for [6]. Therefore, any differences between the treatment and control groups will be accounted for and will not be a result of the observed covariates.

To confirm the effect of the propensity score matching method on reducing systematic difference, it is necessary to compare the covariates between treatment 0 and treatment 1 before and after matching. Our goal is to reduce the difference in the mean of individual covariate between treatment 0 and treatment 1 after matching method. To decide whether there is a significant difference in the mean of individual covariate between treatment 0 and treatment 1, visualizations like box plots, bar plots are carried out first and then a two-sample t-test is applied to compare the results statistically.

Since there are 50 covariates in the dataset makes it too complicated to conclude the changes that matching influenced, and according to the variable description, not all of the covariates are useful in the model. There may be some errors in analyzing the results of matching without any variable selection. The Lasso method has been tried first for variable selection in the Cox model. LASSO can be computed via R Package glmnet [11]. But the final results showed that there are still 42 covariates remaining in the model whose coefficient is larger than zero. It is not convenient to implement a comparison among all of the 42 covariates. Then we can try to use the stepwise package which provides the final model with 28 covariates. Apparently, it is not the perfect result even though it provides a much simpler model. We can do further selection from the final 28 covariates.

According to the Cox matching adjusted model with the selected covariates, table 1 comparing the P-value results of the Cox match adjusted model with full covariates and the 28 covariates from the variable selection. The majority of the 28 covariates in the stepwise final model have smaller P-value, which means the corresponding covariates are more significant in this model. Meanwhile, the P-value of some covariates increases relatively. Therefore, those covariates whose P-value becomes smaller while are less than 0.05 before and after variable selection are reasonable to represent the most significant ones. It is more convenient to concentrate on these 9 covariates and compare the mean of them after the matching method.

In order to confirm the effects of matching, first of all, we draw the boxplots and bar plots of these covariates chosen from stepwise before and after matching. Here we use the boxplot of "surv2md1", "das2d3pc" and the results of propensity score and bar plots of "hema", "chfhx", "meta", "chrpulhx", "psychhx", "dnr1Yes", "renal". Although plots are showing the approximate equivalence between treatment 0 and 1, in favor of unbiased estimate of treatment effect before matching, it is not statistically significant at the 0.05 level of significance. As a result, it is not enough to conclude the effect of matching only by the plots of covariates. Further statistical steps are necessary. To be specific, a two-sample t-test is applied here to test whether the difference of a covariate's mean in treatment 0 and treatment 1 is zero.

Table 1. *Comparison of the P-value results of the Cox match adjusted model with full covariates and the 28 covariates from the variable selection.*

Stepwise final model:		
covariates	**before variable selection**	**after variable selection**
swang1	0.004955	0.00599
surv2md1	1.92E-08	< 2e-16
hemaYes	1.81E-13	3.74E-15
das2d3pc	0.002177	0.000423
ninsclasMedicare	0.124848	0.011959
ninsclasMedicare & Medicaid	0.053879	0.208157
ninsclasNo insurance	0.026792	0.037224
ninsclasPrivate	0.515965	0.973867
ninsclasPrivate & Medicare	0.285651	0.805867
gastrYes	2.87E-06	8.14E-06
chfhx	1.58E-03	0.00012
metaYes	2.75E-04	0.000165
chrpulhx	0.000105	0.0000127
transhx	0.28426	0.000258
psychhx	3.52E-03	0.002359
dnr1Yes	1.08E-03	0.000182
renalYes	0.000866	0.000335
hrt1	4.77E-05	0.000196
liverhx	0.804028	0.045018
cardYes	0.000171	0.000259
respYes	0.00042	0.000468
neuroYes	0.001887	0.001994
pafi1	0.109493	0.009555
bili1	0.002447	0.02724
sod1	0.021681	0.025196
meanbp1	0.035547	0.104889
dementhx	0.041714	0.052488
cardiohx	0.062822	0.140563

Table 2 indicates the mean and standard deviation of covariates in subsets under treat 0 and treat 1 before matching. Among the 9 significant covariates, there are 8 covariates with P-value less than 0.05, which is sufficient evidence to reject the null hypothesis and conclude that the confounding bias exists. Similarly, the visualization and two-sample t-test are conducted for relevant data after matching. It can be seen from the box plot of the PS's before and after matching that the unbalance has been reduced a lot after matching. Also, the test statistics and P-value in table 3 revealed that the differences between covariates under treatment 0 and 1 decreases, since most covariates' P-value are larger than 0.05. Even though the P-value of "survmd1" and "dnr1" is still less than 0.05, the significance becomes less with the P-value increasing much more.

Since the systematic differences between the patients in treatment 0 group and treatment 1 group have been greatly reduced, the effect of treatment on survival time could be compared directly. Figure 3 is the comparison plot of the Kaplan-Meier estimates before and after matching. The log-

rank test statistic is 19.35 with P-value 1.00e-05 before matching, and 23.65 with P-value 1.20E-06 after matching. In other words, the result of the treatment effect (P-value 1.00e-05) is not accurate statistically without matching adjustment. The results provided evidence that the difference of survival functions between the two groups is more significant at significance level 0.05 after propensity score matching and the patients who received RHC had lower survival time than those who did not receive RHC.

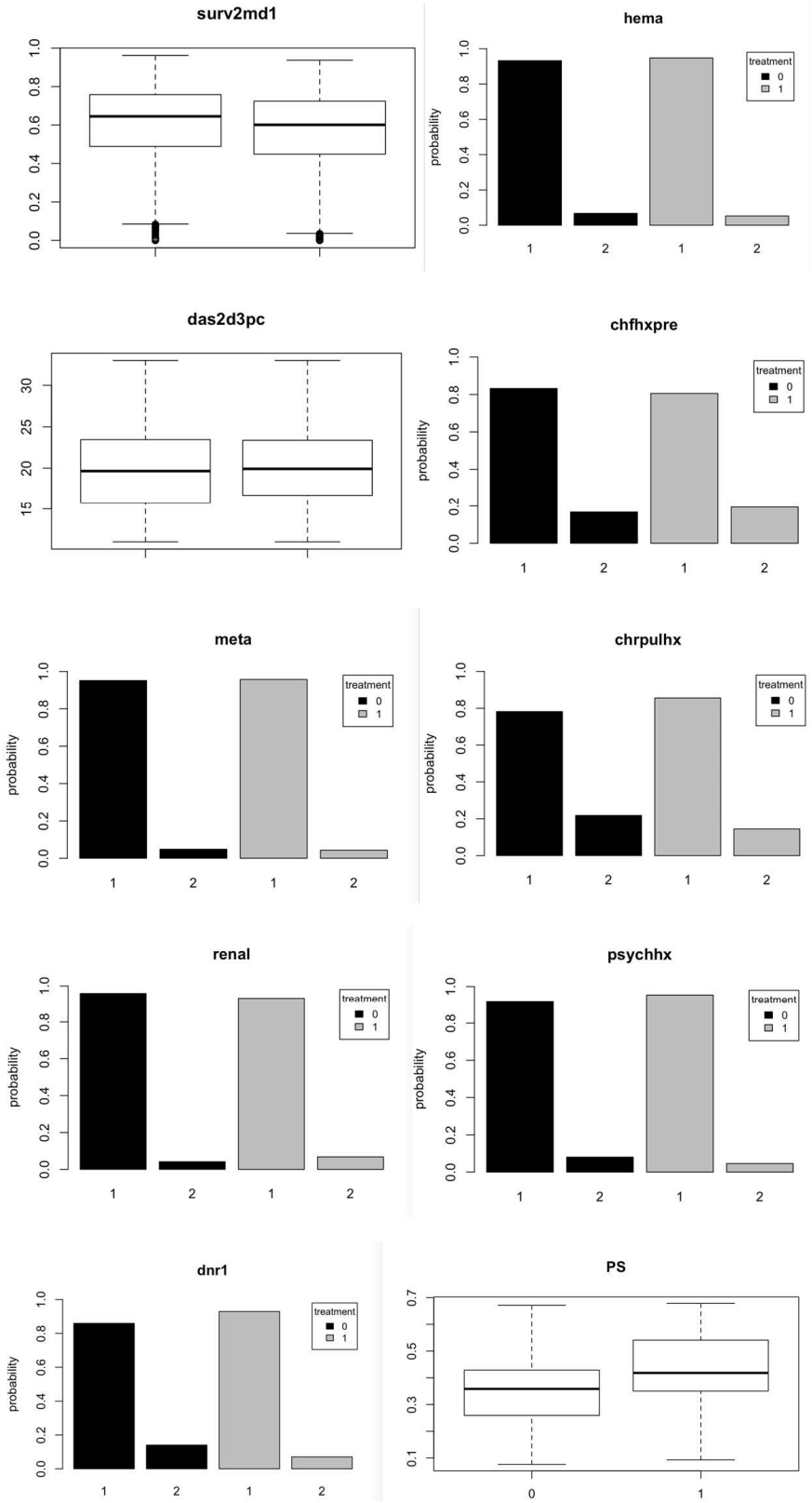


Figure 1. *Comparison for covariates before matching.*

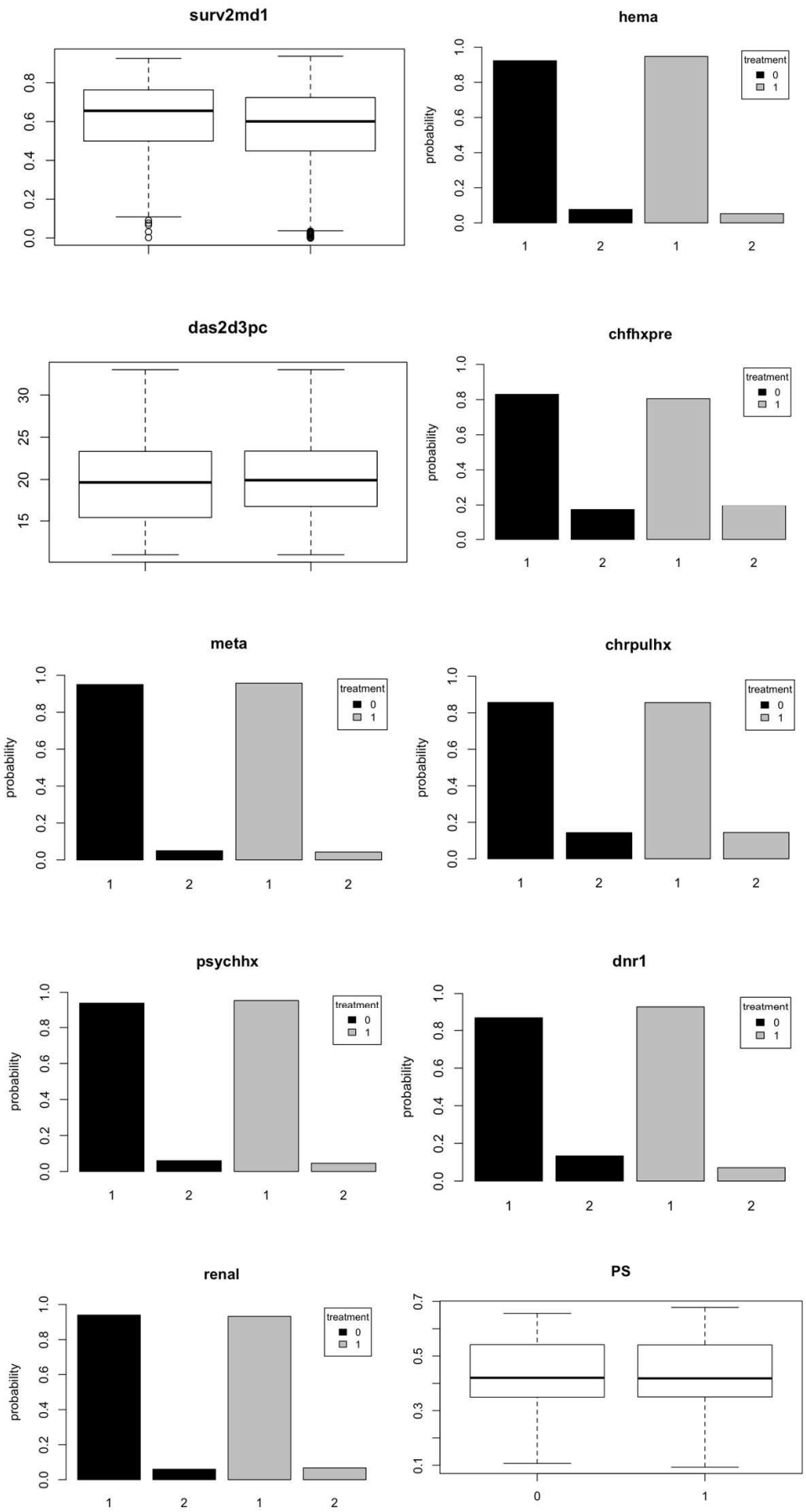


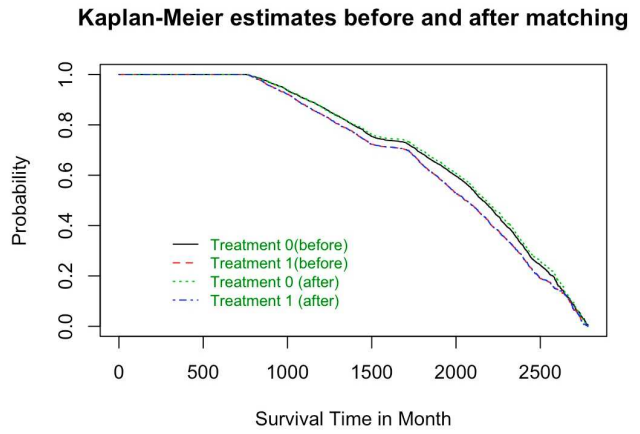
Figure 2. *Comparison for covariates after matching.*

Table 2. *The mean and standard deviation of covariates before matching.*

before matching	treatment 1 (N=2184)		treatment 0 (N=3551)		Comparison	
covariates	**Mean**	**SD**	**Mean**	**SD**	**Test statistics**	**P-value**
surv2md1	0.57	0.2	0.61	0.19	7.3275	2.67E-13
hemaYes	0.05	0.22	0.07	0.25	2.2391	0.02519
das2d3pc	20.7	5.03	20.37	5.48	-2.2784	0.02274
chfhx	0.19	0.4	0.17	0.37	-2.5732	0.0101
metaYes	0.04	0.2	0.05	0.21	1.0255	0.3052
chrpulhx	0.14	0.35	0.22	0.41	6.9412	4.32E-12
psychhx	0.05	0.21	0.08	0.27	5.1115	3.30E-07
dnr1Yes	0.07	0.26	0.14	0.35	8.0912	7.15E-16
renalYes	0.07	0.25	0.04	0.2	-4.3965	1.12E-05
PS	0.43	0.13	0.35	0.14	-22.637	< 2.2e-16

Table 3. *The mean and standard deviation of covariates after matching.*

after matching	treatment 1 (N=2184)		treatment 0 (N=2184)		Comparison	
covariates	**Mean**	**SD**	**Mean**	**SD**	**Test statistics**	**P-value**
surv2md1	0.57	0.2	0.62	0.19	5.325	1.09E-07
hemaYes	0.05	0.22	0.08	0.27	2.2329	0.02564
das2d3pc	20.7	5.03	20.15	5.39	-2.3187	0.02048
chfhx	0.19	0.4	0.17	0.38	-1.3622	0.1732
metaYes	0.04	0.2	0.05	0.22	0.78292	0.4337
chrpulhx	0.14	0.35	0.14	0.35	-0.055431	0.9558
psychhx	0.05	0.21	0.06	0.24	1.4305	0.1527
dnr1Yes	0.07	0.26	0.13	0.34	4.8859	1.09E-06
renalYes	0.07	0.25	0.06	0.24	-0.67794	0.4979
PS	0.43	0.13	0.43	0.13	0.27235	0.7854



	before matching	after matching
hazard ratio	1.159	1.20062
95% CI	(1.085,1.237)	(1.115,1.293)
Log-rank	19.35	23.65
P-val	1.00E-05	1.20E-06

Figure 3. *Comparison plot of the Kaplan-Meier estimates before and after matching.*

3. Stratification on Propensity Score

Stratification on propensity score can also ameliorate the confounding effects of covariates. Each observation for the subject is classified into a propensity quantile based on the propensity score [12]. According to Rosenbaum and Rubin's results, creating five strata based on a continuous variable like the propensity quantile with the stratum boundaries determined by its distribution in the exposed and the comparison group combined eliminates approximately 90% of measured confounding [13]. Therefore, the patients can be assigned to five strata using the propensity score quantile as the cut-off. Within each stratum, the treated patients and untreated patients will have roughly similar propensity score values, also a similar distribution of the measured baseline covariates. The effect of the treatment can be estimated by comparing the outcomes directly between subjects with treatment 0 and subjects with treatment 1 in one stratum if the propensity score has been estimated correctly [7].

To confirm that the systematic difference has been reduced after stratification, it is necessary to compare the covariates' mean under treatment 0 and treatment 1 before and after stratification. The same problem occurs here as with matching when there are 50 covariates in the dataset, which is too complex to conclude whether the stratification makes a difference. Variable selection will be operated again as before. Similarly, the Lasso method has been tried for variable selection but there are 32 covariates left in the final result with coefficients larger than zero. Therefore, I still apply stepwise here aiming to obtain a simpler model and then 28 covariates are selected from the stepwise function with stratification. A further selection is similar as before.

According to the Cox stratification adjusted model with the selected covariates, table 4 comparing the P-value results of the Cox stratification adjusted model with full covariates and the 28 covariates from the variable selection. The majority of the 28 covariates in the stepwise final model have smaller P-value, which means the corresponding covariates are more significant in this model. Meanwhile, the P-value of some covariates increases relatively. So, those covariates

whose P-value becomes smaller while are less than 0.05 before and after variable selection are chosen to represent the most significant ones. It is reasonable to concentrate on these 8 covariates and compare the mean of them after the stratification.

Table 4. *Comparison of the P-value results of the Cox stratification adjusted model with full covariates and the 28 covariates from the variable selection.*

Stepwise final model (stratification)		
covariates	**before variable selection**	**after variable selection**
swang1	1.87E-05	2.23E-06
surv2md1	2.25E-15	< 2e-16
hemaYes	< 2e-16	< 2e-16
das2d3pc	1.37E-04	2.71E-06
ninsclasMedicare	1.11E-02	8.61E-03
ninsclasMedicare & Medicaid	0.498918	0.681657
ninsclasNo insurance	3.71E-03	4.88E-04
ninsclasPrivate	0.385573	0.928321
ninsclasPrivate & Medicare	0.414613	0.375537
dnr1Yes	1.89E-06	7.78E-08
bili1	3.92E-06	1.03E-05
metaYes	2.52E-07	8.61E-07
chfhx	3.21E-04	2.06E-06
psychhx	1.34E-04	7.06E-05
hrt1	0.000138	0.000312
traumaYes	0.051076	0.004354
gastrYes	0.0000523	0.013768
transhx	0.051076	0.001106
neuroYes	2.82E-08	0.000635
sod1	5.28E-03	2.36E-03
sexMale	0.111296	0.027512
amihx	0.011171	0.013947
age	0.001246	0.003172
meanbp1	0.078033	0.035246
renalYes	0.589645	0.028471
alb1	0.136271	0.086081
dementhx	0.017683	0.041219
gibledhx	0.186017	0.061819
chrpulhx	0.022417	0.073299

To confirm the effects of stratification, a two-sample t-test is applied here to test whether the difference of a covariate's mean in treatment 0 and treatment 1 is zero, which is related to test whether the systematic difference in covariates has been reduced. Table 5 shows the mean and standard deviation of corresponding covariates in subsets under treatment 0 and treatment 1 before stratification. All of the 8 covariates' P-value is less than 0.05. That is sufficient evidence to reject the null hypothesis and conclude that the confounding bias exists and the stratification adjustment is necessary when evaluating the effect of treatment on survival time. Similarly, the two-sample t-test is conducted for relevant data after stratification. It can be seen from the test statistics and P-value in table 6 that the systematic differences between covariates under treatment 0 and treatment 1 decreases, since most covariates' P-value increased and the significance of the difference in mean between treatment 0 and treatment 1 decreased. Even though the P-value of the covariates is still less than 0.05, the significance becomes less with the P-value increasing. The reason for the zero P-value is that the stratified two-sample t-test function defines the extreme P-value as zero.

To compare the mean of selected covariates in the subject of treatment 0 and subject of treatment 1 more accurately and sufficiently, table 7 and 8 indicate the mean of each covariate in each stratification group and use two-sample t-test respectively to test whether there is a significant difference in the mean of covariates between treatment 0 and treatment 1 after stratification. Apparently, most of the P-values are larger than 0.05 which concludes to fail to reject the null hypothesis and illustrates that the systematic difference and confounding bias are reduced.

Since systematic differences between the patients in treatment 0 group and treatment 1 group have been greatly reduced, the effect of treatment on survival time is comparable. Figures 4 and 5 is the Cox proportional hazard regression model for treatment 0 and treatment 1 after stratification. It is obvious that the balance of the covariates is better achieved after stratification. Figure 6 are the comparison plots of the Kaplan-Meier estimates between treatment 0 and 1 in each stratification group. As we can see from the five plots, the survival time of patients after RHC treatment is relatively decreased, which leads to the same conclusion as propensity score matching.

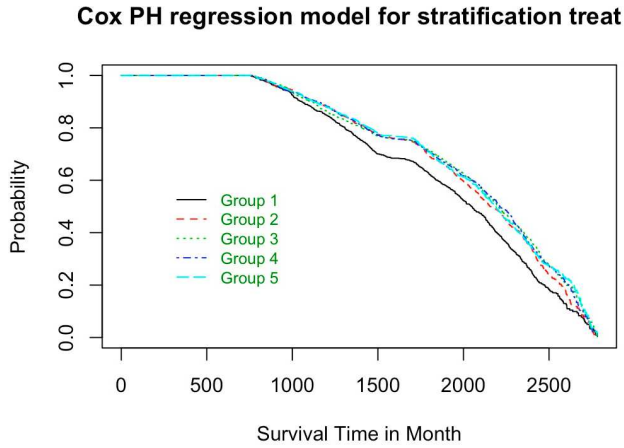


Figure 4. *Cox proportional hazard regression model for treatment 0 after stratification.*

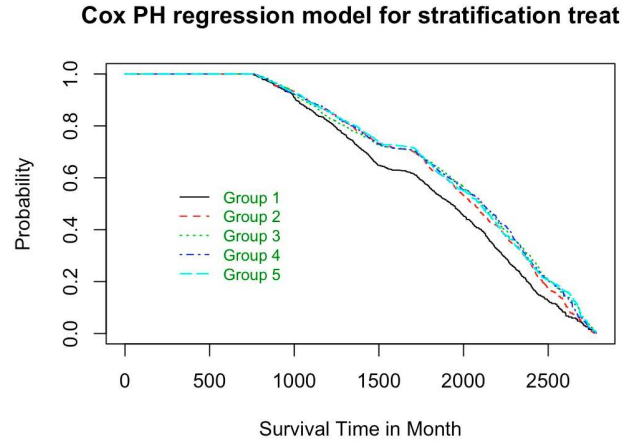


Figure 5. *Cox proportional hazard regression model for treatment 1 after stratification.*

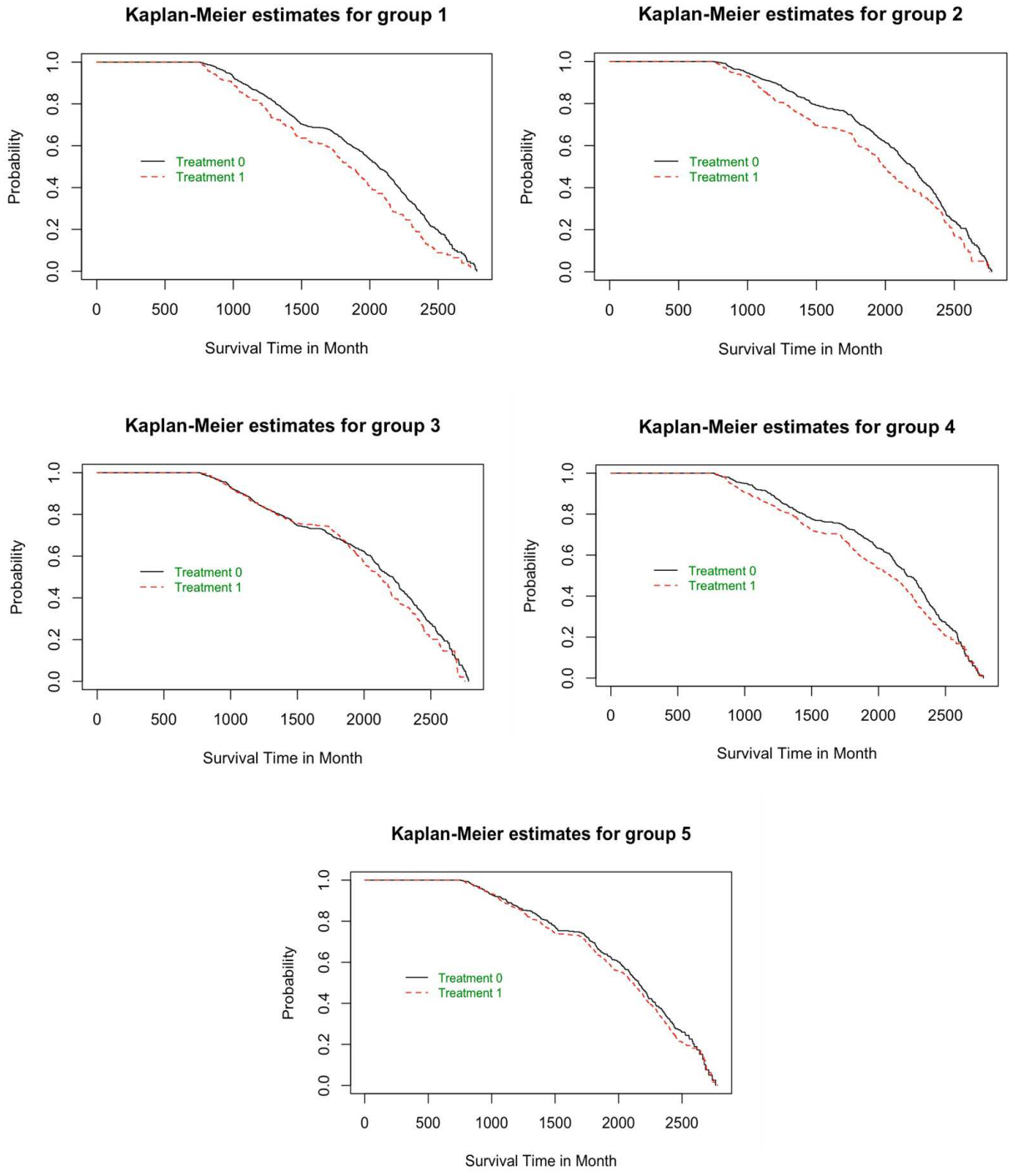


Figure 6. *Comparison of the Kaplan-Meier estimates in each stratification group.*

Table 5. *The mean and standard deviation of covariates before stratification.*

before stratification	treatment 1 (N=2184)		treatment 0 (N=3551)		Comparison	
covariates	Mean	SD	Mean	SD	Test statistics	P-value
surv2md1	0.57	0.2	0.61	0.19	7.3275	2.67E-13
hemaYes	0.05	0.22	0.07	0.25	2.2391	0.02519
das2d3pc	20.7	5.03	20.37	5.48	-2.2784	0.02274
dnr1Yes	0.07	0.26	0.14	0.35	8.0912	7.15E-16
chfhx	0.19	0.4	0.17	0.37	-2.5732	0.0101
psychhx	0.05	0.21	0.08	0.27	5.1115	3.30E-07
sod1	137.04	7.68	136.33	7.6	3.3864	0.000713
PS	0.43	0.13	0.35	0.14	-22.637	< 2.2e-16

Table 6. *The mean and standard deviation of covariates after stratification.*

after stratification	treatment 1 (N=2184)		treatment 0 (N=3551)		Comparison	
covariates	Mean	SD	Mean	SD	Test statistics	P-value
surv2md1	0.57	0.2	0.61	0.19	0.03874876	0
hemaYes	0.05	0.22	0.07	0.25	0.01464931	0.02702
das2d3pc	20.7	5.03	20.37	5.48	-0.3293159	0.02354
dnr1Yes	0.07	0.26	0.14	0.35	0.0695531	0
chfhx	0.19	0.4	0.17	0.37	-0.02675702	0.01164
psychhx	0.05	0.21	0.08	0.27	0.03475315	0
sod1	137.04	7.68	136.33	7.6	0.7042972	0.00052
PS	0.43	0.13	0.35	0.14	-0.08168341	0

Table 7. *The mean of covariates in each stratification group.*

after stratification	Mean	N (%)	PS	surv2md1	hemaYes	das2d3pc	dnr1Yes	chfhx	psychhx	sod1
Group 1	treat 0	938 (26.4%)	0.173	0.586	0.018	19.809	0.172	0.143	0.071	137.280
	treat 1	209 (9.6%)	0.191	0.448	0.048	20.056	0.105	0.158	0.053	136.785
Group 2	treat 0	783 (22.1%)	0.317	0.611	0.061	19.531	0.193	0.147	0.110	136.718
	treat 1	364 (16.7%)	0.320	0.563	0.044	19.983	0.088	0.173	0.060	136.217
Group 3	treat 0	707 (19.9%)	0.382	0.603	0.103	20.769	0.105	0.154	0.086	136.992
	treat 1	440 (20.1%)	0.383	0.567	0.061	21.075	0.064	0.136	0.048	136.925
Group 4	treat 0	658 (18.5%)	0.445	0.633	0.131	21.669	0.096	0.246	0.065	136.766
	treat 1	489 (22.4%)	0.450	0.600	0.090	21.284	0.057	0.264	0.041	135.370
Group 5	treat 0	465 (13.1%)	0.576	0.613	0.032	20.481	0.108	0.163	0.062	137.538
	treat 1	682 (31.2%)	0.583	0.586	0.026	20.622	0.066	0.205	0.038	136.565

Table 8. *Comparison of the mean of covariates in each stratification group.*

after stratification		N (%)	PS	surv2md1	hemaYes	das2d3pc	dnr1Yes	chfhx	psychhx	sod1
Group 1	Test statistics	938 (26.4%)	-4.5104	8.0084	-2.5682	-0.6369	2.3736	-0.55698	0.97569	0.96355
	P-value	209 (9.6%)	7.14E-06	2.84E-15	0.01035	0.5243	0.01778	0.5776	0.3294	0.3355
Group 2	Test statistics	783 (22.1%)	-1.6964	4.3185	1.191	-1.4198	4.554	-1.1405	2.672	0.97937
	P-value	364 (16.7%)	0.09008	1.71E-05	0.2339	0.1559	5.83E-06	0.2543	0.007646	0.3276
Group 3	Test statistics	707 (19.9%)	-0.67512	3.1893	2.4497	-0.94883	2.3777	0.82698	2.4687	0.14676
	P-value	440 (20.1%)	0.500	0.001	0.014	0.343	0.018	0.408	0.014	0.883
Group 4	Test statistics	658 (18.5%)	-3.1341	2.6375	2.1539	1.1147	2.3888	-0.67711	1.7983	3.2663
	P-value	489 (22.4%)	0.002	0.008	0.031	0.265	0.017	0.499	0.072	0.001
Group 5	Test statistics	465 (13.1%)	-2.412	2.7371	0.58299	-0.45618	2.5111	-1.7803	1.8879	1.8697
	P-value	682 (31.2%)	0.016	0.006	0.560	0.648	0.012	0.075	0.059	0.062

4. Inverse Probability of Treatment Weighting

Kaplan Meier estimator is widely used in clinical studies to compare survival time between different treatment groups. However, if certain covariates corresponding to low survival rates are more strongly represented in one group than another, which is considered as over-represented, the survival estimated by the Kaplan-Meier method form one group would appear to be worse than survival estimated from the other group. Another approach reducing confounding effects was proposed by Xie and Liu in 2005 [14]. They developed the Adjusted Kaplan Meier estimator (AKME) using the inverse probability of treatment weighting (IPTW). The estimated propensity score, the probability of being treated in a certain group conditioning on a set of covariates, is used to construct the weights for subjects. A weight is assigned to each individual as the inverse of the propensity score. For example, a subject with a higher propensity score, which is considered as over-represented, is assigned with a lower weight. On the other hand, subjects with a lower propensity score, considered as under-represented, will be given a higher weight [15]. They also proposed a weighted log-rank test for statistical comparison of the survival functions of the two groups.

As with the matching and stratification, we apply the IPTW method to the Right Heart Catheterization study. The propensity score of each patient is estimated using logistic regression in the same way. Then the Kaplan-Meier estimators of both the treatment group and control group are adjusted with the weight as the inverse of the propensity score. If the propensity score is estimated correctly, the sampling bias will be removed after weighting adjustment. Figure 7 shows the Kaplan-Meier estimator on the survival function of the two groups before and after weighting adjustment. We can see from the plot that the survival curve of the subject with treatment 1 is lower than the subject with treatment 0. It becomes more obvious after adjustment. We also perform the log-rank test for statistical comparison of the survival functions. Table 9 shows the comparison of the hazard ratio estimate with or without IPTW procedure. The log-rank test statistic without weighting is 19.35 with P-value 1.00E-05, while the weighted log-rank test statistic is 75.45 with a P-value less than 2.00e-16. We conclude that the difference in survival functions between the two groups is

more significant at significance level 0.05 after weighting with the inverse of the propensity score. Moreover, the plot shows that the survival time of subjects with treatment 1, who received the RHC, tends to be lower than the survival time of those not receiving RHC.

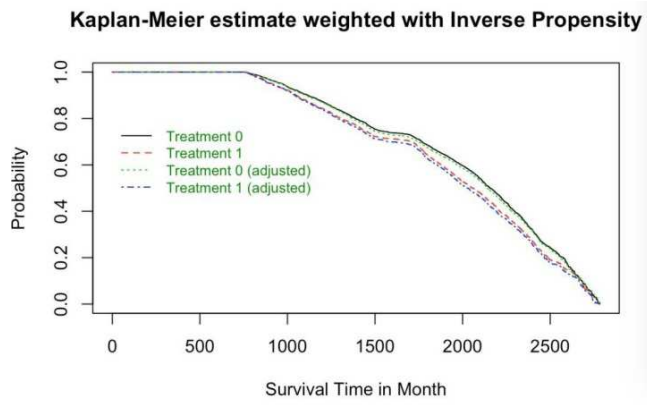


Figure 7. *The Kaplan-Meier estimator on survival function before and after weighting adjustment.*

Table 9. *Comparison of the hazard ratio estimate before and after IPTW procedure.*

	before IPTW	**after IPTW**
hazard ratio	1.159	1.18284
95% CI	(1.085, 1.237)	(1.139, 1.229)
Log-rank	19.35	75.45
P-val	1.00E-05	<2e-16

5. Discussions and Conclusions

In this paper, we discussed three bias adjustment methods for causal inference in non-randomized clinical trials. According to the application results from three bias adjustment methods on the Right Heart Catheterization study, we conclude from the Cox proportional-hazards regression that patients receiving RHC had decreased survival time. Moreover, the difference in survival time between the two groups becomes more significant at significance level 0.05 after reducing the confounding bias.

Matching on propensity score is a good method for removing the bias between the treated group and the control group on the background covariates. It is preferred when the sample size of the control group is much large than the sample size of the treatment group. Stratification is preferred when the sample size is large enough since the estimation would be unreliable if there are not enough patients in each stratum. The IPTW method showed better performance in general. One may consider matching or stratification when the control group variance is much larger than the variance of the treatment group. Overall, a combination of the methods could be applied to make the estimation of the treatment effect more efficient.

Acknowledgements

The authors are grateful to the editor and anonymous referee for their beneficial and accurate comments that improved this paper.

References

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

[2] Schober, P., Vetter, T. R. (2020). Propensity Score Matching in Observational Research. Anesthesia & Analgesia. 130 (6): 1616-1617.

[3] Rosenbaum, P. R. (1987). Model-based direct adjustment. Journal of the American Statistical Association. 82, 387-394.

[4] Granger, E., Watkins, T., Sergeant, J. C. (2020). A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. BMC Medical Research Methodolgy. 20, 132.

[5] Connors, A. F., Speroff, T. & Dawson, N. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. Journal of American Medical Association. 18, 294-1295.

[6] Thavaneswaran, A., Lix, L. (2008). Propensity score matching in observational studies. Manitoba Centre for Health Policy.

[7] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research. 46 (3): 399-424.

[8] Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine. 28 (25): 3083-107.

[9] Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science. 25, 1-21.

[10] D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to non-randomized control group. Statistics in Medicine. 17, 2265-2281.

[11] Nabi, R., Su, X. (2017). An R package for sparse estimation of cox proportional hazards models via approximated information criteria. The R Journal. 9 (1): 2073-4859.

[12] Leon, A. C., Hedeker, D. (2011). Propensity score stratification for observational comparison of repeated binary outcomes. Statistics and Its Interface. 4, 489–498.

[13] Rosenbaum, P. R., Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association. 79, 516–524.

[14] Xie, J., Liu, C. (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Statistics in Medicine. 24, 3089-3110.

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