Propensity score methods to adjust for confounding in assessing treatment effects: bias and precision

Z Wang

Citation

Z Wang. *Propensity score methods to adjust for confounding in assessing treatment effects: bias and precision*. The Internet Journal of Epidemiology. 2008 Volume 7 Number 2.

Abstract

There is an increasing interest in the use of propensity score (PS) methods for confounding control, with generally three ways of estimating adjusted treatment effects in pharmacoepidemiological studies: 1) stratification on PS, 2) matching on PS and 3) using PS as a covariate. To assess bias and precision of different methods, we conducted simulations in three scenarios: 1) treatment had no effect but the crude estimate showed a protective effect; 2) treatment was protective and the crude estimate was more extreme; and 3) treatment increased the risk but the crude estimate showed protective. Adjusting for confounders in all methods shifted the effect estimates toward the true values. Adjusted odds ratios using the PS stratification and the method using PS as a covariate were biased due to either residual confounding or over-adjustment. Matching on PS produced less biased average estimates than other methods but the precision of effect estimates was lower.

Sponsor: The National Health and Medical Research Council (NH&MRC) of Australia (511013).

INTRODUCTION

Propensity score, introduced by Rosenbaum and Rubin, [1] is the conditional probability of a subject's receiving the treatment of interest given a set of covariates. The use of propensity score is increasing for confounding control, especially for evaluating treatment effect using observational data. [2] However, as suggested by Sturmer et al, there is little evidence that propensity score methods yield substantially different estimates compared with conventional regression methods. [2] Several simulation studies have conducted to evaluate the performance of propensity score methods. [3-6] In a Monte Carlo simulation study, Austin et al shows that conditioning on the propensity score produces a biased estimation of the true conditional odds ratio and the true conditional hazard ratio. [5] In another Monte Carlo simulation study, Brookhart et al suggest that standard model building tools designed to create good predictive models of the exposure will not always lead to optimal propensity score models. [6] On the other hand, Cepeda et al found that propensity score estimates were less biased than the logistic regression estimates when there were six or fewer events per confounder. [3]

Generally, there are three ways to apply propensity scores: 1) stratification on the propensity score, 2) matching on the propensity score, and 3) using the propensity score as a covariate. [2] Little is known about the effect of different ways of using propensity scores on the bias and precision of treatment effect estimates. Simulation studies use computer intensive procedures to assess the performance of statistical methods in relation to a known truth. [7] In this study, we used simulations to examine different propensity score methods and logistic regression methods in assessing the treatment effects. We mainly focused on comparing biases and precisions of those methods under different scenarios with various sample sizes.

METHODS

DATA SIMULATION PROCEDURES

As in typical epidemiological studies of assessing treatment effect, we started with two groups: treatment and non-treatment groups. The variable X was coded 1 for treatment and 0 for non-treatment. The random number generator in Stata [8] was used to generate five confounding variables and the outcome variable. Among the five confounding variables, two were continuous and three dichotomous. Three random dichotomous variables were coded 1 and 0. First, we generated three uniform variables U1, U2 and U3 with values between 0 and 1. For the non-treatment group, we set a dichotomous variable W1 to be 1 if U1< 0.20 and 0 otherwise, W2 = 1 if U2< 0.10 and 0 otherwise, and W3 = 1 if U4 0.4 < 0.40 and 0 otherwise. For the treatment group,

we set W1 = 1 if U1< 0.60, W2=1 if U2< 0.50 and W3 = 1 if U3 < 0.20. Two random continuous variables were generated with expected means of 0.25 and 0.20, respectively, in the treatment group and -0.25 and -0.20 in the non-treatment group for W4 and W5 respectively. The standard deviations for both variables in both populations were 1. The above procedures generated five variables (W1-W5) associated with the treatment (X).

Outcome variable (Y) was modeled using logistic regression as a function of (confounding variables (W1-W5) and treatment (X) variable in three scenarios:

Scenario 1. The odds ratio as a measure of treatment effect was set to be 0.70. The odds ratios for confounders W1, W2, W3, W4 and W5 were 0.3, 0.5, 3.0, 0.4 and 0.5. Baseline probability of having the outcome (Y=1) was 0.30 when all Ws and X were 0. The probability of a subject with a specific combination of Ws and X was estimated:

$$\text{logit}(Y) = \ln(0.3/0.7) + \ln(0.3)*W1+\ln(0.5)*W2+\ln(3.0)*W3 + \ln(0.4)*W4 + \ln(0.5)*W5 + \ln(\text{true OR})*X$$

where the true OR = 0.70 in the scenario 1.

$$\Pr(Y|X, Wi) = \exp(\text{logit}(Y))/(1 + \exp(\text{logit}(Y))$$

The outcome variable (Y) was set to be 1 if the randomly generated uniform number was less than Pr(Y|X, Wi), and to be 0 if otherwise.

Scenario 2. The associations between confounders (Wi) and the outcome (Y) were the same as those in Scenario 1 but there was no treatment effect (the true OR = 1).

Scenario 3. The associations between confounders (Wi) and the outcome (Y) were the same as those in Scenario 1 and 2 but the true OR = 1.6.

SAMPLE SIZES AND NUMBERS OF SIMULATIONS

We performed 4 different sets of simulations with 50, 100, 500 and 1000 subjects, respectively, in the treatment group, and the same numbers in the control group. We generated 36000 dataset with 3000 datasets for each combination of scenarios and sample sizes.

ADJUSTMENT FOR CONFOUNDING

In each of the 36000 simulated studies, we estimated the crude and adjusted odds ratios using conventional logistic regression and three propensity score methods.

Logistic regression method: To estimate the effect of the treatment on the outcome, we applied logistic regression with the outcome (Y) as dependent variable and all confounding factors (Wi) and treatment variable (X) as independent variables.

Propensity score stratification: We obtained the propensity score of the treatment (X), the probability of being treated, using logistic regression with the treatment (X) as a dependent variable and all confounders (Wi) as independent variables. The propensity scores were divided into five strata with 20^{th}, 40^{th}, 60^{th} and 80^{th} percentiles as the cutoffs. Then, we used the outcome variable (Y) as the dependent variable and treatment (X) and the categories of the propensity score were independent variables in logistic regressions.

Propensity score matching: The propensity score matching refers to the pairing of treated and untreated subjects with similar values of the propensity scores and the discarding unmatched subjects. As proposed by Rubin, all propensity scores were transformed to the logit scale, which is referred to as the linear propensity score. [9, 10] We matched each treated subject to a untreated subject with the closest propensity score (1:1 matching) within the range of linear propensity score ±0.25. If there were no untreated subjects within the range for a treated subject, this subject would not be included in the conditional logistic regression. A unique identification number was assigned to each matched pair, and this variable was used as the identifier variable for the matched groups in the conditional logistic regression, in which the dependent variable was the outcome (Y) and the independent variable was the treatment (X).

Propensity modeling: We took the linear propensity score, a continuous variable, as a covariate in the logistic regression. The dependent variable was the outcome (Y) and independent variables were the treatment (X) and the linear propensity score. In this study, we only assessed linear relationship between the propensity score and the outcome.

MEASURES OF INTEREST

Bias: Odds ratios by four different methods were calculated and compared with the true values, which were 0.7 in scenario 1, 1.0 in scenario 2 and 1.6 in scenario 3. The differences between the true and estimated odds ratios indicated the bias of the effect estimates. Average differences of log odds ratios were presented according to the methods, sample sizes and scenarios.

Precision: We calculated standard errors of log odds ratios as a measure for precision. Since the same data sources were used for all four methods, the average standard errors among different methods were compared.

We conducted all simulations and analyses using Stata 10. [8] Matching was a tedious and time consuming procedure, so we developed a Stata program (CMATCH) to perform this task. The change-in-estimate approach has been recommended for selecting confounders for control. [11] All confounding variables in this study were true confounders and all were included in the analyses. Confounder selection was not the focus of this study. However, the distortion of these confounders to the odds ratio will be demonstrated using a Stata program. [12]

RESULTS

CHARACTERISTICS OF DATASET SIMULATIONS

Table 1 shows the characteristics of treated and untreated groups. Those confounding variables were substantially different between two groups. Table 2 shows the numbers of cases in 3 scenarios according to sample sizes. The numbers of cases were very small, ranging from 3 to 10 per confounder, when sample size was 50 in the treatment group. When the sample size was 1000, there were over 100 cases per confounder.

Figure 1

Table 1. Mean (minimum and maximum) values of confounding variables in 3000 simulations for each sample size

Confounders	Non-treatment	Treatment
Sample size = 50		
W1	10 (2, 22)	30 (17, 41)
W2	5 (0, 14)	25 (13, 36)
W3	20 (9, 32)	10 (1, 19)
W4	-0.25 (-0.78, 0.22)	0.25 (-0.22, 0.80)
SD for W4	1.00 (0.61, 1.35)	1.00 (0.67, 1.34)
W5	-0.20 (-0.79, 0.30)	0.20 (-0.30, 0.70)
SD for W5	1.00 (0.69, 1.33)	0.99 (0.70, 1.35)
Sample size = 100		
W1	20 (9, 34)	60 (41, 78)
W2	10 (2, 23)	50 (34, 73)
W3	40 (20, 57)	20 (7, 34)
W4	-0.25 (-0.59, 0.07)	0.25 (-0.12, 0.66)
SD for W4	1.00 (0.79, 1.30)	0.99 (0.76, 1.22)
W5	-0.20 (-0.54, 0.16)	0.20 (-0.15, 0.57)
SD for W5	1.00 (0.78, 1.25)	1.00 (0.75, 1.23)
Sample size = 500		
W1	100 (71, 128)	300 (267, 333)
W2	50 (29, 74)	251 (215, 284)
W3	200 (165, 236)	100 (73, 132)
W4	-0.25 (-0.39, -0.11)	0.25 (0.09, 0.43)
SD for W4	1.00 (0.90, 1.11)	1.00 (0.91, 1.11)
W5	-0.20 (-0.36, -0.04)	0.20 (0.08, 0.36)
SD for W5	1.00 (0.89, 1.12)	1.00 (0.88, 1.10)
Sample size = 1000		
W1	200 (159, 250)	600 (536, 653)
W2	100 (66, 136)	499 (447, 559)
W3	400 (348, 467)	200 (149, 242)
W4	-0.25 (-0.38, -0.15)	0.25 (0.13, 0.36)
SD for W4	1.00 (0.93, 1.08)	1.00 (0.92, 1.11)
W5	-0.20 (-0.30, -0.09)	0.20 (0.10, 0.31)
SD for W5	1.00 (0.92, 1.09)	1.00 (0.91, 1.09)

Figure 2

Table 2. Mean (minimum and maximum) number of cases by sample size and scenario: 3000 simulations in each sample size and scenario combination

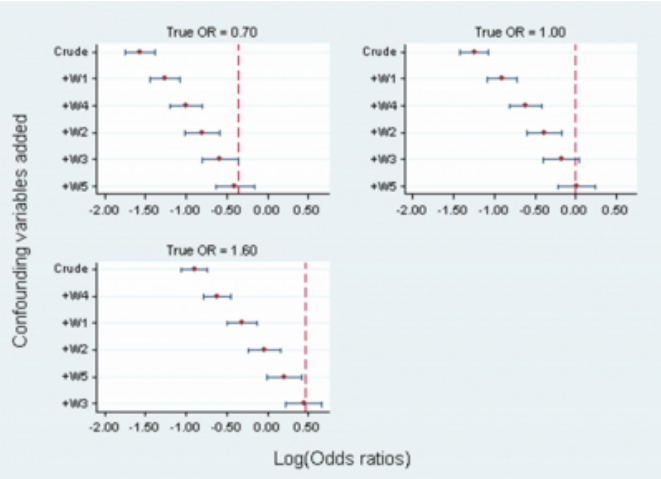
Sample size	Scenario 1: OR = 1.0	Scenario 2: OR = 0.7	True OR = 1. 6
50	31 (15, 46)	29 (15, 44)	33 (20, 51)
100	61 (40, 84)	57 (37, 80)	67 (43, 91)
500	306 (263, 350)	288 (249, 331)	336 (289, 382)
1000	613 (535, 683)	576 (504, 644)	672 (594, 742)

One example data set (one of the 3000 sets with the sample size of 1000) was randomly selected to demonstrate the presence of confounding effects from five variables W1-W5, using the change-in-effect estimate method. [12] Figure 1 shows the effect estimates after adjusting for each of confounders according to the magnitude of the change-in-effect estimate in a stepwise fashion. All five confounding variables contributed to the distortion of odds ratio estimates

(the change-in-estimate) in three scenarios. Adjusting for confounding variables altered the effect estimates from huge protective effects (crude odds ratios) to the true effect values.

Figure 3

Figure 1. Crude and adjusted odds ratios in one data set: variables added in a stepwise fashion according to the change-in-estimate. Vertical dash lines represent true odds ratios.



EFFECT ESTIMATES

Table 3 shows the average odds ratios according to sample size and scenario combinations. The crude odds ratios in three scenarios showed a strong protective effect of the treatment. The odds ratios adjusting for confounding factors using logistic regression and three propensity score methods were closer to the true values, indicating that the treatment had 1) no effect in scenario 1, 2) a protective effect in scenario 2 but to a much less extent than the crude estimate, and 3) a risk effect in scenario 3, opposite to the crude estimate. However, those methods performed differently in terms of biases and precisions of their effect estimates.

Figure 4

Table 3. Mean odds ratio by sample size and scenario: 3000 simulations in each sample size and scenario combination

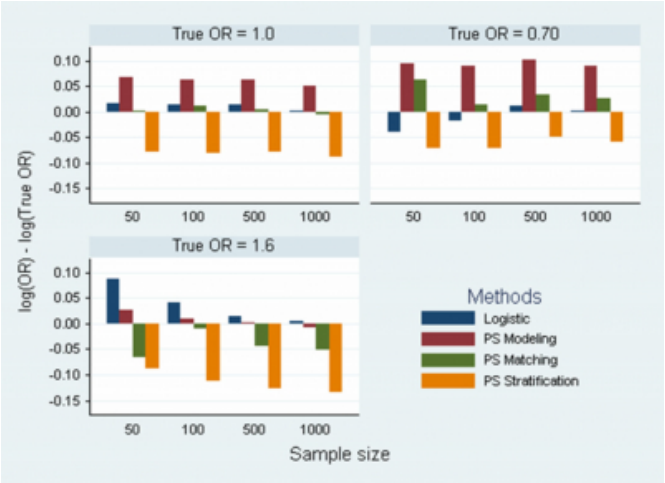
	Scenario 1: True OR = 1.0	Scenario 2: True OR = 0.7	Scenario 3: True OR = 1.6
Sample size = 50			
Crude	0.266	0.197	0.388
Logistic	1.017	0.673	1.745
Modeling	1.069	0.769	1.642
Matching	1.000	0.745	1.496
Stratification	0.925	0.651	1.464
Sample size = 100			
Crude	0.271	0.202	0.393
Logistic	1.015	0.688	1.668
Modeling	1.064	0.766	1.617
Matching	1.011	0.710	1.584
Stratification	0.922	0.652	1.432
Sample size = 500			
Crude	0.276	0.209	0.396
Logistic	1.015	0.708	1.622
Modeling	1.063	0.774	1.601
Matching	1.004	0.724	1.529
Stratification	0.925	0.667	1.409
Sample size = 1000			
Crude	0.276	0.208	0.396
Logistic	1.002	0.699	1.607
Modeling	1.050	0.765	1.589
Matching	0.995	0.718	1.519
Stratification	0.915	0.660	1.399

BIAS

Figure 2 shows the differences between the estimated and true log odds ratios. The propensity score stratification method consistently produced an odds ratio away from the true value in a direction towards the crude odds ratio regardless of the sample size and the magnitude of the true value. In two of the three scenarios (1 and 2), the propensity score linear modeling yielded average odds ratios that were higher than the true values and in a direction that was further away from the crude effect estimate. The bias from the propensity score matching tended to be less extreme than those from propensity score stratification and propensity score modeling.

Figure 5

Figure 2. Mean bias by methods, scenario and sample size: 3000 simulations per sample size and scenario combination.

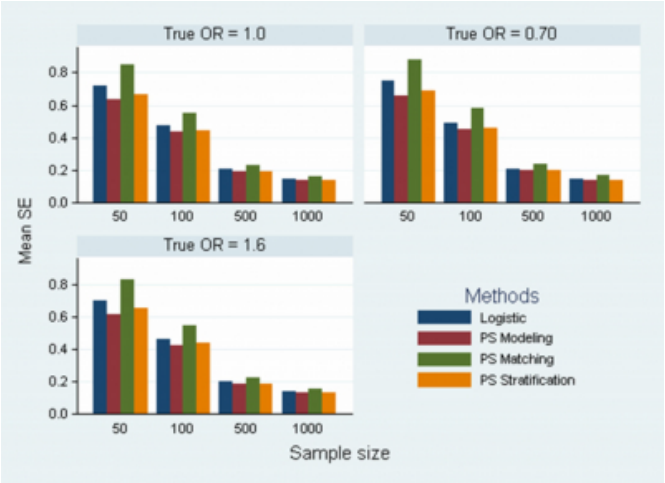


PRECISION

Mean standard errors are shown in Figure 3. The propensity score stratification and propensity score modeling had lower mean standard errors than the conventional logistic regression. The propensity matching had highest mean standard errors.

Figure 6

Figure 3. Mean bias by methods, scenario and sample size: 3000 simulations per sample size and scenario combination.

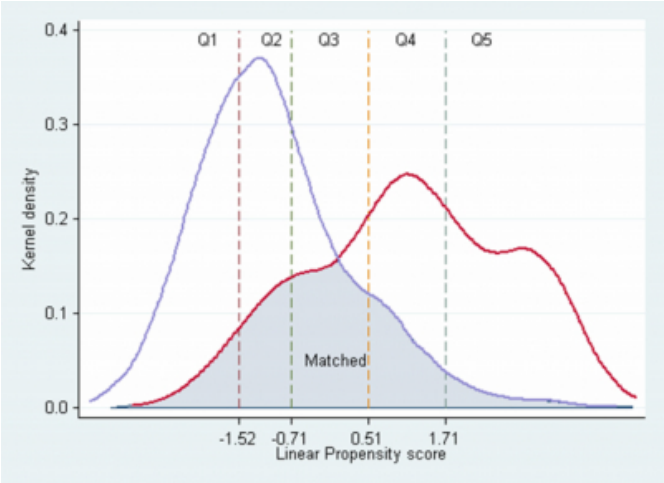


We calculated the numbers of pairs used in the propensity score matching methods. On average, 42 (min: 23, max: 61) and 220 (175, 264) and 443 (378, 513) treated subjects were matched to the untreated subjects, which were 42%, 44% and 44% of the total subjects, for simulations with 100, 500 and 1000 treated subjects, respectively.

To further explore possible explanations of the higher bias using the stratification method and higher mean standard errors using the propensity score matching method. Using the example dataset for Figure 1, we generated Figure 4, which demonstrates the striking difference in the distribution of propensity scores between treatment and non-treatment groups. Even within each propensity score stratum, two groups still had different propensity scores. The figure also shows that a small proportion (shaded area) of treated and untreated subjects could be matched.

Figure 7

Figure 4. Distributions of linear propensity scores in treatment and non-treatment groups: vertical lines are the stratification cutoffs and shaded area represents the matching between treatment and non-treatment subjects



DISCUSSION

In this study, we found that the propensity score methods provided biased effect estimates. Residual confounding persists in the propensity score stratification method regardless of sample size and the strength and direction of the true treatment effects. Using the propensity score as a linear predictor also produced biased effect estimates but the direction of this bias can be different from that of residual confounding. Matching by propensity scores excluded a large proportion of subjects and resulted in the effect estimates with less precision.

Several systematic reviews have been conducted on this topic. [2, 13, 14] Sturmer et al found 13% studies using a propensity score method had an effect estimate that differed by more than 20% from that obtained with a conventional regression model. [2] Shah et al found the statistical significance of the association differed between two methods in 10% of the effect estimates, in which the association was

statistically significant using conventional regress but not significant using propensity score methods. [13] Most observational studies had similar results whether using conventional regression or using propensity scores to adjust for confounding. [2, 13] Drake shows that omitting a confounder in the propensity score method produces biases comparable to those in a conventional regression model. [15] Using an example dataset from Hosmer and Lemeshow, [16] Drake and Fisher reported that the propensity score method leads to a different conclusion with regard to the effects of smoking on birthweight. [17] However, it is difficult to assess the performance of different methods using real data sets because the true values of the treatment effects are unknown.

Simulations studies provide an opportunity to assess the performance of different statistical methods in relation to a known truth using computer intensive procedures. [7] Several simulation studies have been conducted on the propensity score methods. [3, 5, 6, 18, 19] Brookhart et al revealed that the model best predicted exposure did not yield the optimal propensity score model in terms of efficiency when including a non-confounder in the propensity score model. [6] They suggested that variables that are unrelated to the exposure but related to the outcome should be always included in a propensity score model. [6, 9] Austin et al found that failure to include an important confounding variable in the propensity score model can result in variable imbalance between exposed and unexposed subjects and result in biased estimation of the effect. [5] In this study, all variables were true confounding factors and performances of different propensity score methods were assessed using the same data set with the same confounding variables.

Cepeda et al found that the propensity score estimates were less biased that the logistic regression estimates when there were six or fewer events per confounder. Overall the propensity score was more robust, more precise and had more power than logistic regression. [3] The purpose of this study was not to compare the logistic regression estimates with those of different propensity score methods. Since we carried out the simulations according to the known logistic regression models to generate data, logistical regression models were theoretically correct. However, we demonstrated some potential problems of different propensity score methods. In this study, even when the number of cases was about six per confounder (when treatment group n = 50), the propensity score methods produced biased estimates. The statistical power with such a small sample size is too low to provide a reasonable effect estimates regardless of the methods. Even if there were no confounding in this study, the sample sizes required to detect an odds ratio of 0.7 and 1.6 should be 638 and 314 with 80% power.

Among the three ways of applying propensity scores, the propensity stratification method produced biased effect estimates toward the crude estimate, indicating the presence of residual confounding. We did not check if the distribution of the confounders in the treated and un-treated groups in each stratum were similar. However, the presence of residual confounding is likely to be a common phenomenon because within each stratum the treatment subjects can still have higher propensity scores than their untreated counterparts, as illustrated in Figure 4.

Matching by propensity scores can efficiently balance the propensity scores between two groups at the expense of losing a large proportion of the subjects. In our simulated data only 42% to 44% subjects were matched, the precision of the estimates of the matching method were lower than those of other methods.

The linear modeling of propensity scores as a continuous variable can also produce biased estimates. In two of the three scenarios, the linear modeling produced biased estimates to the opposite side of the crude estimate, indicating an over-adjustment. We did not explore whether fitting a non-linear relationship in the model would change the magnitude of bias or alter the direction of bias. Only one set of confounding variables were used in all three scenarios and four methods. In the real world, the relationships among confounders, treatment and outcome can be more complicated. Therefore, the magnitude and the direction of bias of propensity score methods are likely to vary accordingly.

CONCLUSION

Propensity score methods potentially produce biased effect estimates. Residual confounding is common when using the propensity stratification method. Propensity matching results in lower precision of effect estimates. Linear modeling of propensity scores may not appropriate for all data. Better understanding of the benefits, limitations and appropriate use of the propensity score methods are needed before they are widely used.

ACKNOWLEDGEMENT

Zhiqiang Wang was supported by the National Health and Medical Research Council (NHMRC) of Australia (511013).

References

1. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.
2. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2006;59:437-47.
3. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158:280-7.
4. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf 2008.
5. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med 2007;26:734-53.
6. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149-56.
7. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med 2006;25:4279-92.
8. Stata Statistical Software: Release 10 [program]. College Station, TX: StataCorp LP, 2007.
9. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996;52:249-64.
10. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services & Outcome Research Methodology 2001;2:169-188.
11. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol 1993;138:923-36.
12. Wang Z. Two postestimation commands for assessing confounding effects in epidemiological studies Stata Journal 2007;7:183-196.
13. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 2005;58:550-9.
14. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf 2004;13:841-53.
15. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 1993;49:1231-1236.
16. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: John Wiley, 1989.
17. Drake C, Fisher L. Prognostic models and the propensity score. Int J Epidemiol 1995;24:183-7.
18. Austin PC, Grootendorst P, Normand SL, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. Stat Med 2007;26:754-68.
19. Judkins DR, Morganstein D, Zador P, Piesse A, Barrett B, Mukhopadhyay P. Variable selection and raking in propensity scoring. Stat Med 2007;26:1022-33.

Author Information

Zhiqiang Wang, PhD
Centre for Chronic Disease, School of Medicine, the University of Queensland