Journal of Biopharmaceutical Statistics, 17: 1–13, 2007

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400601044691

Taylor & Francis
Taylor & Francis Group

STATISTICAL AND REGULATORY ISSUES WITH THE APPLICATION OF PROPENSITY SCORE ANALYSIS TO NONRANDOMIZED MEDICAL DEVICE CLINICAL STUDIES

Lilly Q. Yue
CDRH, U.S. Food and Drug Administration, Rockville, MD, USA

Propensity score analysis is a versatile statistical method used mainly in observational studies for improving treatment comparison by adjusting for up to a relatively large number of potentially confounding covariates. Recently, there has been an increased interest in applying this method to nonrandomized medical device clinical studies. In the application of the methodology, some statistical and regulatory issues arise in both study design and analysis of study results, such as the need for pre-specifying clinically relevant covariates to be measured, appropriate patient populations, and the essential elements of statistical analysis, planning sample size in the context of propensity score methodology, handling missing covariates in generating propensity scores, and assessing the success of the propensity score method by evaluating treatment group overlap in terms of the distributions of propensity scores. In this paper, the advantages and limitations of this methodology will be revisited, and the above issues will be discussed and illustrated with examples from a regulatory perspective.

Key Words: Medical devices; Nonrandomized studies; Propensity score analysis.

1. INTRODUCTION

Nonrandomized medical device clinical studies are sometimes conducted when indeed appropriate. In fact, nonrandomized studies are more common in medical devices than in drugs and biological products, for practical or ethical reasons. In such studies, depending on clinical knowledge as well as previous study results, information on many baseline covariates, for example, baseline demographics and risk factors, is usually collected. In a nonrandomized study, the advantages of a well-designed and conducted randomized clinical trial are no longer available. In a randomized trial, a patient is randomly assigned to either treatment A or B so that each patient has a known probability of receiving each treatment, avoiding both obvious and non-obvious clinical selection of patients for one treatment or the

Received July 21
No official support or endorsement by the Food and Drug Administration of this presentation is intended or should be inferred.
Address correspondence to Lilly Q. Yue, CDRH, U.S. Food and Drug Administration, 1350 Piccard Dr., Rockville, MD 20850, USA; E-mail: lilly.yue@fda.hhs.gov

other. All patient covariates, observed or unobserved, are expected to be balanced, i.e., nearly equally distributed, between the two treatment groups, which is one of the most important benefits of randomization. Assumptions underlying many statistical tests for comparison are met or hold approximately and the observed treatment difference is an unbiased estimate of true treatment difference (Rubin, 1997). However, the above advantages are not guaranteed in a small or a poorly designed or conducted randomized trial. In contrast, in a nonrandomized study, patients are not randomly assigned to treatment A or B; instead, the chance of being assigned to treatment A rather than treatment B may vary from patient to patient depending on patient baseline covariates. As a result, the treatment groups might not be comparable before the start of treatment, due to imbalance of the baseline covariates, and direct treatment comparison, such as linear or logistic regression, may not be valid. Therefore, it is necessary to adjust for covariate imbalance in treatment comparison. To do so, traditional matching and subclassification in terms of baseline covariates as well as regression (covariate) adjustment could be used. However, when there are many imbalanced covariates, matching is often not possible, and likewise subclassification could be impracticable because the number of subclasses grows exponentially as the number of covariates increases. Regression analysis may not work due to the problem of over-fitting. It is common in the area of cardiovascular devices for there to be a relatively large number of clinically relevant covariates. To address the complicated and common issue, the propensity score methodology, developed by Rosenbaum and Rubin (1983), can be used. By replacing the entire collection of covariates with a scalar function that appropriately summarizes these covariates, the propensity score technique allows the straightforward assessment whether the treatment groups overlap enough regarding baseline covariates to allow for a sensible treatment comparison. When sufficient overlap is present, the methodology allows a straightforward treatment comparison that reflects adjustment for imbalances in all observed covariates (Braitman and Rosenbaum, 2002; D'Agostino, 1998; D'Agostino and Rubin, 2000; Rosenbaum and Rubin, 1983, 1984; Rubin, 1997).

Approximately six years ago, propensity score methods were introduced to the Center for Devices and Radiological Health (CDRH) at the FDA, and since then have been recommended in some nonrandomized device studies. Also, some applications of the methodology have appeared in the medical literature (Blackstone, 2002; Grunkemeier et al., 2002; Wolfgang et al., 2002). In this paper, we will revisit the methodology, and discuss statistical issues encountered in its application, for instance, handling missing baseline covariate values, evaluating the treatment group comparability, assessing the resulting covariate distribution balance, and accounting for propensity score adjustment in sample size estimation. Also, we will point out the limitation of its use and try to provide some insight into the appropriate application, with two hypothetical examples motivated by our review experience. We then discuss regulatory requirements when using the propensity score analysis as primary analysis in medical device submissions. Finally, we conclude that propensity score methodology can in some cases be a good alternative to traditional covariate adjustment methods to adjust for many covariates and reduce bias in treatment comparison. However, it should be noted that the propensity score methods can only adjust for imbalance in observed covariates but not in unobserved ones, while randomization tends to balance the

distribution of all covariates, observed and unobserved. Hence, nonrandomized studies with propensity score analysis are still inferior to randomized trials in terms of the level of scientific evidence. Therefore, well-designed and conducted randomized trials are still preferred and strongly recommended whenever possible, and the availability of propensity score analysis for nonrandomized studies is not an excuse for one not to conduct a randomized trial when it is deemed necessary and feasible.

2. PROPENSITY SCORE ANALYSIS

The propensity score $e(\boldsymbol{x})$ for a subject with a vector $\boldsymbol{x}$ of *observed* covariates is the conditional probability of receiving treatment A $(Z = 1)$ rather than treatment B $(Z = 0)$ given $\boldsymbol{x}$:

$$e(\boldsymbol{x}) = \Pr(Z = 1 \mid \boldsymbol{x})$$

Rosenbaum and Rubin (1983) state that the propensity score $e(\boldsymbol{x})$ is a balancing score in the sense that it is a function of the observed covariates $\boldsymbol{x}$ such that the conditional distribution of $\boldsymbol{x}$ given $e(\boldsymbol{x})$ is the same for subjects who had received treatment A and subjects who had received treatment B. They also show that if treatment assignment is strongly ignorable, that is, the treatment assignment Z and the outcome Y are conditionally independent given the covariates, $\boldsymbol{x}$, then the average treatment effect at each value of the propensity score is an unbiased estimate of the true treatment effect at that propensity score, and therefore, matching, subclassification and covariate adjustment on the propensity score can produce unbiased estimates of the treatment effect.

The propensity score is generated by modeling the probability of treatment group membership, based on the baseline covariates, for example, by a multiple logistic regression or discriminant analysis. In this statistical modeling, the outcome is an event: treatment actually received, A or B, and predictor variables include all observed covariates and some interactions of various orders, e.g., age, gender, duration of disease,..., age $*$ duration,.... However, the clinical outcome variable of interest, such as major complication event, is not involved. The propensity score model is not used to make any statistical inference concerning treatment comparison, and is instead employed to find propensity scores which are used to match patients and therefore create balance between the two treatment groups. Thus, estimating a propensity score with many terms does not create a problem in model fitting and does not introduce bias in treatment comparison (Rubin, 1997).

Once the propensity score model is constructed, the propensity score is estimated for each subject. A group of patients with the same propensity score are equally likely to have been assigned to treatment A. Within a group of patients with the same propensity score, some patients actually received treatment A and some received treatment B, just as if they had been randomly allocated to whichever treatment they actually received. It is just like "randomized after the fact." Therefore, the comparison between the two patients with the same propensity score, where one received treatment A and the other received treatment B, is expected to be

balanced with respect to the observed baseline covariates. The above assumes that unobserved covariates do not affect treatment assignment given observed covariates.

After the propensity score model is constructed and a propensity score is estimated for each patient, three common covariate adjustment methods based on propensity scores can be used for treatment comparison: matching, subclassification and regression. There exist multiple ways to do matching (D'Agostino, 1998) with "nearest available matching on the estimated propensity score" being the easiest one. Then, treatment comparison of the clinical outcome Y is based on matched sets of patients. In subclassification, all patients are ordered based on their propensity scores, and then divided into 5 or 6 subgroups or strata of approximately equal size. Within each subclass or stratum, propensity scores are relatively homogeneous; i.e., all patients have similar probability of receiving treatment A. Rosenbaum and Rubin (1983, 1984) and Rubin (1997) show that within each subclass, if the propensity scores are relatively constant, then the distribution of all covariates should be approximately the same in both treatment groups, and hence the two treatment groups are comparable. They also state that subclassification on the propensity score tends to balance all k covariates that are used to estimate the propensity score, and often five strata based on the propensity score will remove over 90% of the bias in each of these covariates, if the covariate distributions overlap sufficiently well. Then, an overall treatment comparison in a clinical outcome can be obtained by a weighted average of the subgroup-specific treatment comparisons. However, if study size is small or imbalance is severe, some subclasses may contain patients from only one treatment group so that the treatment comparison in that subclass is impossible. In regression (covariate) adjustment, the relationship of a clinical outcome and treatment received is modeled with estimated propensity score as a covariate. Also, a subset of original important covariates can be included in the model. Regression analysis within propensity score subclasses can sometimes provide a more efficient estimator of treatment effect (D'Agostino, 1998; Rosenbaum and Rubin, 1983, 1984; Rubin, 1997).

Moreover, Rubin (1997) points out that the propensity score is a one-dimensional summary of the observed covariates such that when the propensity scores are balanced across two treatment groups, the distribution of all the covariates are balanced in expectation across the two groups. Therefore, this dimension reduction of covariates allows the straightforward assessment of whether the two treatment groups overlap enough with respect to baseline covariates to allow a sensible treatment comparison. When such overlap is present, the propensity score analysis allows a straightforward treatment comparison that reflects adjustment for differences in all observed baseline covariates.

3. APPLICATION EXAMPLES

Example 1. Suppose that a multicenter, nonconcurrent, two-arm study was conducted to demonstrate the safety and effectiveness of a new device through the comparison of the new device to a medical therapy without device. Suppose that there were 200 patients in the study, among whom about 2/3 received the experimental device and the others the control treatment. The primary effectiveness endpoint was treatment success, defined based on pre-specified clinical criteria. The hypothesis testing for the primary endpoint was superiority in terms of

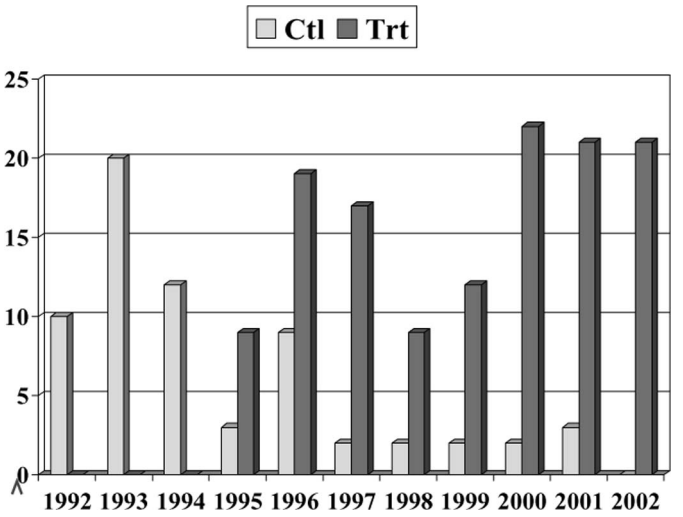

Figure 1 Enrollment time.

treatment success rate. There were about 20% patients with at least one missing covariate value and 20 imbalanced important baseline covariates, identified at a significance level 0.05, respectively. Additionally, since medical treatment for the disease had changed dramatically over the previous decade, the enrollment time is an important covariate to consider. As it turned out, a majority of control patients were treated in the early 1990s and, as is shown in Fig. 1, there is considerable imbalance in the distribution of enrollment time between the two treatment groups. Due to the above-mentioned imbalances, it was concluded that the two treatment groups were not comparable, any direct treatment comparisons on the effectiveness endpoint would be inappropriate, and all p-values from direct treatment comparisons would be uninterpretable. To attempt to resolve the problem, propensity score analysis was performed, where the propensity score was defined as the conditional probability of receiving the device, given a patient's baseline covariates.

In order to include all those statistically imbalanced as well as clinically important baseline covariates in the propensity score model, missing covariate values were handled by multiple imputations through a Markov Chain Monte Carlo (MCMC) method. Without imputation, the 20% patients with missing covariate values would have to be excluded from the propensity score modeling. Sometimes a single data set for propensity score analysis is generated, which is called generalized propensity score analysis (D'Agostino and Rubin, 2000).

At the completion of the propensity score modeling, it was noticed that while the fitted propensity score model predicted the treatment group membership very well, the distribution of the propensity scores generated by the model in the two treatment groups may not overlap enough to allow a sensible treatment comparison (See Figs. 2–3). Of course, the traditional logistic regression analysis might be able to provide "significant" treatment comparison with different modeling, but the inability

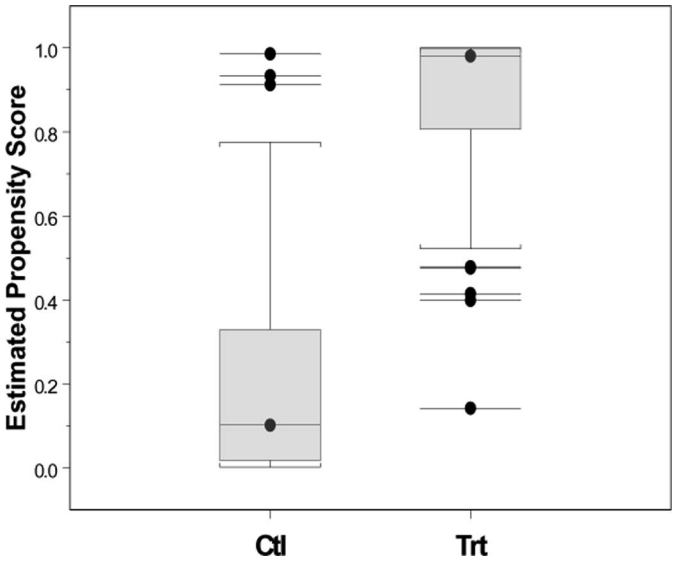

Figure 2 Estimated propensity scores (with time).

to provide warnings about treatment group comparability in the traditional logistic regression analysis of a clinical outcome is actually a pitfall, as pointed out by Rubin (1997). It can be seen that (Fig. 2) with the enrollment time included in the modeling, there is very little overlap in the distribution of propensity scores between the treatment and control groups. Excluding the enrollment time from the propensity score model, the overlap was better but still not satisfactory (Fig. 3). The treatment group comparability was further evaluated using propensity score

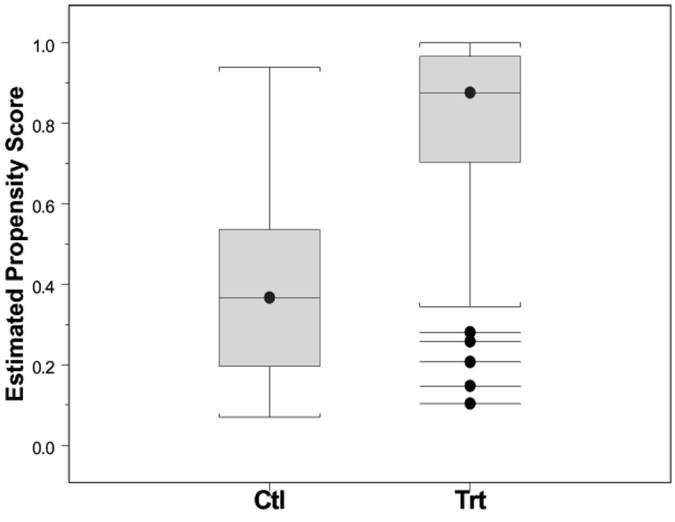

Figure 3 Estimated propensity scores (without time).

Table 1 Distribution of patients in propensity score quintiles

		1	2	3	4	5	Total
with time	Ctl	39	19	8	1	0	67
		58%	28%	12%	2%	0%	
	Trt	1	21	33	38	40	133
		0%	16%	25%	29%	30%	
without time	Ctl	30	25	8	4	0	67
		45%	37%	12%	6%	0%	
	Trt	11	14	32	36	40	133
		8%	11%	24%	27%	30%	

quintiles (Table 1). The fifth quintile contains 30% of the treated patients but does not contain any control patients to compare with, and the first quintile contains 58% of the control patients but only 1 treated patient. Although both Cochran-Mantel-Haenszel test controlling for the propensity score quintile (see Table 2) and logistic regression using propensity score as a continuous covariate provided statistically "significant" treatment comparisons, the results were necessarily unreliable, due to the exclusion of 30% of the treated patients in the Cochran-Mantel-Haenszel test and the failure of logistic regression analysis to give any warning of the lack of treatment group comparability. Hence, it was concluded that the two treatment groups did not overlap sufficiently to allow a sensible treatment comparison, and therefore any treatment comparisons adjusted for imbalanced covariates were problematic and all significant p-values were uninterpretable.

Example 2. Suppose that the effectiveness of a new cardiovascular device was investigated in a nonrandomized noninferiority study through the comparison of the new device with an active control, using the nine-month major adverse cardiac event (MACE) rate as the primary endpoint. A noninferiority margin of 10% was pre-determined with respect to the MACE incidence rate, based on clinical judgment. To reduce potential bias in the treatment comparison, a propensity score analysis was pre-specified in the protocol to simultaneously adjust for 15 pre-identified common baseline demographics and risk factors and baseline lesion characteristics. There were 800 patients in the study with about 1/3 in the new device group, where propensity score adjustment was not taken into account in the planning of sample size. Although it seemed to have worked well in this particular case, sample size calculation without considering propensity score adjustment could be inappropriate in other situations.

Table 2 Distribution of treatment success in propensity score quintiles (with time; S: success, N: sample size)

		1	2	3	4	5	Total
	N	39	19	8	1	0	67
Ctl	S	16	9	1	0		26
	N	1	21	33	38	40	133
Trt	S	0	14	27	24	23	88

Table 3 Covariate balance checking before and after propensity score stratification adjustment

	Mean		*p*-value	
Covariate	New	Control	Before	After
C1	0.25	0.40	<.0001	0.4645
C2	0.28	0.21	0.0421	0.8608
C3	2.41	2.75	0.0003	0.3096
C4	11.02	12.16	<.0001	0.5008
C5	3.00	3.08	0.0202	0.2556
C6	62.75	66.81	<.0001	0.4053

The propensity score was defined as the conditional probability that a patient would have been assigned to the new device, given the patient's baseline covariates. Prior to fitting a propensity score model of the 15 baseline covariates, 6 apparently imbalanced ones were identified by comparing the two treatment groups (Table 3). Starting with the main effects of the 15 covariates and all interactions and quadratic terms of the six apparently imbalanced covariates, a logistic regression model with a stepwise selection process (0.15 for stay and leave out) was used to build the propensity score model. The final propensity score model included the main effects of the six apparently imbalanced covariates as well as a quadratic term. The entire study population was then divided into propensity score quintiles, with 160 subjects each. Consistent with the definition of the propensity score, most patients with the new device were in the quintiles 3, 4, and 5 (Table 4).

After propensity scores have been generated for all the patients, the overlap of the two treatment groups was evaluated with respect to those propensity scores by box plots in Fig. 4. By definition, patients with the new device tend to have higher propensity scores than patients with the control device. Some patients in the new device group had higher estimated probability of receiving the new device than any patients in the control group, suggesting that some combinations of covariate values not appearing in the control group.

Recall that the propensity score method is supposed to simultaneously balance the covariates between the two treatment groups within each propensity score quintile stratum/subclass. Checking for balance in those covariates after stratification is therefore critical, although often not provided in medical device submissions. In this example, the balance was examined for each of the 15 covariates

Table 4 Distribution of patients at the five propensity score quintiles

Quintile	Control	New	Total
1	135	25	160
2	120	40	160
3	117	43	160
4	113	47	160
5	48	112	160
Total	533	267	800

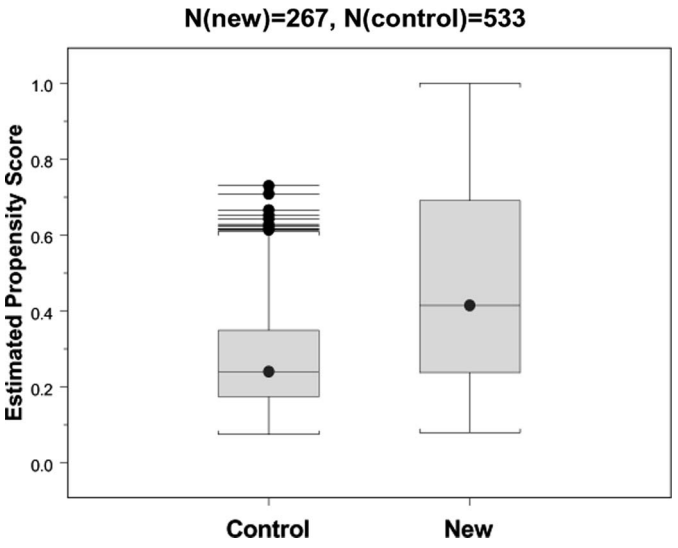

Figure 4 Boxplots of the estimated propensity score.

by a two-way analysis of variance model, 2 (treatment groups) $\times$ 5 (propensity score quintiles). The two-way interaction of the treatment and propensity score quintile was first examined for each covariate. There was only one covariate with a significant 2-way interaction at the significance level of 0.15, detected by the two-way ANOVA. The p-value for treatment comparison after adjustment for propensity score was compared with that before the adjustment. It was found that after propensity score adjustment, the six previously identified imbalanced covariates were not significantly different between the two treatment groups at significance level 0.05 (Table 3). The balance of the six covariates within each propensity score quintile was also checked using bar chart. For example, Fig. 5

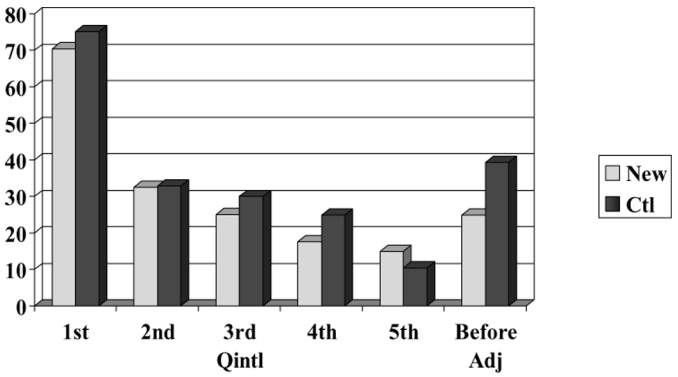

Figure 5 Percentage of patients with covariate C_1 in the propensity score quintiles.

Table 5 Balance check for the percentage of patients with covariate C_1

Stratification	Treatment	Estimate (%)				
Before						
	New			25		
	Control			40		
	Control-New			15		
After			Quintiles			
		1	2	3	4	5
	New	70.4	32.6	25.0	17.6	15.0
	Control	75.2	32.8	30.0	24.8	10.4

displays the balance within the propensity score subclasses for the proportion of patients with covariate C_1. While the overall difference between the two treatment groups on the covariate was 15% before the stratification, the difference was much smaller within each propensity score subclass (Table 5).

The above results on balance were thought to be satisfactory. But what if the results were not satisfactory? Rubin (1997) indicates that if important within-subclass differences between treatment groups had been found on some covariates, then either the propensity score prediction model would need to be reformulated or it would have to be concluded that the covariate distributions did not overlap sufficiently to allow stratification to adjust for these covariates, as is the case in Example 1.

The directly adjusted nine-month MACE rate estimate obtained from the subclass-specific rates, with subclass total weights, for the two treatment groups and a 95% confidence interval estimate of treatment difference, new — control, were calculated. Compared to the prespecified noninferiority margin, 10%, the noninferiority of the new device to the control was concluded.

4. LIMITATIONS

Propensity score methods can simultaneously adjust for many imbalanced covariates and reduce bias in treatment comparison. However, caution is needed in the application of the methodology, especially in nonrandomized medical device studies, where the sample size is often relatively small.

Propensity score methods can only adjust for observed covariates and not for unobserved ones. This is always a limitation of nonrandomized studies compared with randomized trials where randomization tends to balance both observed and unobserved covariates. Therefore, some sensitivity analysis for hidden bias is often desirable in a medical device submission. Propensity score methods may not eliminate all selection bias, due to the limitation of propensity score modeling, which typically uses a linear combination of covariates (Rubin, 1997). Braitman and Rosenbaum (2002) state that the propensity score methods work better under the three conditions: First, when event is rare, for example, the number of the major complication events in ablation catheter study is usually small, about 3% or 4%. Second, there are a large number of patients in each treatment group. Although it is expected that all observed covariates could be balanced between the two treatment

groups by propensity scores, in a small nonrandomized study, substantial imbalance of some covariates might be unavoidable despite the use of a sensible propensity score. Third, there are many covariates measured. It is seriously degraded when important variables influencing selection have not been collected.

The propensity score technique is not the only way of adjusting for covariates. Also, it may or may not be helpful in a particular comparative study (Blackstone, 2002; Grunkemeier et al., 2002; Wolfgang et al., 2002 also in Example 1 above). Compared to randomized trials which are considered the highest level of evidence for treatment comparison, nonrandomized studies using propensity score methods are less rigorous, and therefore not as definitive as randomized trials. Therefore, although covariates between two treatment groups may be balanced well using propensity scores, the statistical inference, e.g., *p*-values and confidence intervals, obtained from such treatment comparisons may still carry lower level of scientific assurance, compared to those from randomized trials.

5. SOME STATISTICAL AND REGULATORY ISSUES

In nonrandomized confirmatory medical device studies, propensity score methodology is an addition to, not a substitute for, more traditional covariate adjustment methods. If propensity score analysis is selected to be the primary data analysis, it needs to be planned in advance through the pre-specification of the methodology, as post-hoc propensity score analysis only provides supporting evidence. In addition, a number of statistical and regulatory issues need to be considered in study design stage and final study report.

In the protocol, it is important to prespecify as many as possible clinically relevant baseline covariates that will be collected in the study and used in the data analysis, as well as specific propensity score method that will be used, for example, stratification. Also, sensitivity analysis should be planned for unobserved covariates.

It is crucial to select comparable patient populations, since propensity score analysis does not work (in fact, no statistical method works) when there is serious imbalance in baseline covariates between the two treatment groups. It is sometimes impossible to predict in advance whether the patient population with a new device is comparable to that with a control, which is a danger with nonrandomized study and has been a serious issue in nonrandomized device study submissions. In fact, an unsuccessful nonrandomized study is more burdensome than a randomized trail at the outset.

It is inappropriate to conduct propensity score analysis when sample size is small, which could result in some propensity score subclasses that may contain patients from only one treatment group and then the exclusion of these patients from the data analysis. Such exclusion of patients should be discouraged as it could lead to biased treatment comparison.

It is noticed that the propensity score adjustment should be taken into account in sample size estimation. A fundamental rule of sample size determination is that the method of sample size estimation should be based on the planned method of data analysis. It is not hard to imagine that the required sample size in a study depends on treatment group comparability. However, it is sometimes difficult to predict in advance and hence hard to make assumptions on the treatment group comparability.

In the final study analysis, ignoring missing covariate values has been a problem in device submissions with propensity score analysis. It is common, for example, in cardiovascular device studies that a large number of baseline covariates are measured and collected. This could result in a high proportion of patients with at least one covariate value missing. Without any action on the missing covariate values, these patients could be automatically excluded from propensity score modeling with logistic regression and then from treatment comparison with respect to clinical outcome. Again, excluding such a large number of patients from treatment comparison leads to questionable study results.

Furthermore, assessment of the success of the propensity score estimation by checking the resulting balance of the distributions of covariates and evaluation of treatment group comparability by the distributions of propensity scores should be included in the final study report.

6. SUMMARY

Propensity score methodology generalizes the technique that uses just one confounding covariate to allow simultaneous adjustment for many observed covariates and thus reduce bias in treatment comparisons. It gives an alternative to traditional covariate adjustment methods, and especially provides a convenient way to assess treatment group comparability. However, it should be used wisely due to its limitations discussed above. Also, in regulatory environment, some special issues need to be well considered in the study protocol, for example, prespecification of the methodology along with measurement and inclusion of every important covariate, appropriate selection of patient populations, and sample size estimation in the context of propensity scores. In the final study report, missing covariate imputation, assessment of the success of the propensity score estimation and evaluation of treatment group comparability should be demonstrated. Moreover, due to the limitations of observational studies, randomized trials are still preferred and strongly encouraged whenever possible, especially in the development of new technology in the medical device world.

ACKNOWLEDGMENTS

The author thanks the Editor-in-Chief of the journal, Dr. Shein-Chung Chow, for his invitation of the article and Drs. Greg Campbell and Heng Li for their valuable comments during the preparation of this manuscript.

REFERENCES

Blackstone, E. H. (2002). Comparing apples and oranges. *J. Thorac. and Cardiovasc. Surg.* 1:8–15.

Braitman, L., Rosenbaum, P. R. (2002). Rare outcomes, common treatments: Analytic strategies using propensity scores. *Ann. Intern. Med.* 137:693–696.

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a nonrandomized control group. *Stat. Med.* 17:2265–2281.

D'Agostino, R. B., Jr. Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *JASA* 95:749–759.

Grunkemeier, G. L., Payne, N., Jin, R., Handy, J. R., Jr. (2002). Propensity score analysis of stroke after off-pump coronary artery bypass grafting. *Ann. Thorac. Surg.* 74:301–305.

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

Rosenbaum, P. R., Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *JASA* 79:516–524.

Rubin, D. B. (1997). Estimating casual effects from large data sets using propensity scores. *Ann. Intern. Med.* 127:757–763.

Wolfgang, C., Winkelmayer, W. C., Glynn, R., Mittleman, M., Levin, R., Pliskin, J., Avorn, J. (2002). Comparing mortality of elder patients on hemodialysis versus peritoneal dialysis: A propensity score approach. *J. Am Soc Nephrol* 13:2353–2362.