Minimizing Bias in Observational Comparative Clinical Studies

Lilly Q. Yue

U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Keyword: Minimize bias; Observational study; Propensity scores

Abstract

Although well-controlled and conducted randomized clinical trials are viewed as gold standard in the safety and effectiveness evaluation of medical products, including drugs, biological products and medical devices, observational (non-randomized) comparative studies play an important role in the medical product evaluation, due to ethical or practical reasons, in both pre-market and post-market regulatory settings. However, various biases could be introduced at every stage and into every aspect of the observational study, and adversely impact the interpretation of the study results. Among existing statistical techniques for addressing some of the bias issues, propensity score methodology is one increasingly used in regulatory settings, due to its unique feature of separating "study design" and "outcome analysis". In this presentation, we will focus on the propensity score approaches in minimizing bias in medical device studies through adequate study design and data analysis.

1. Introduction

Carefully designed and well-conducted randomized controlled trials (RCTs) provide the highest level of evidence in the safety and effectiveness evaluation of medical products. One of the key advantages of RCT is that with randomization, all baseline covariate (observed and unobserved) distributions tend to be balanced across two treatment groups, leading to unbiased estimation of treatment effect. Another critical feature of RCT is that the study is "prospectively" designed; that is, it is designed without access to any outcome data from either treatment group, resulting in convincing and interpretable treatment effect estimation on outcomes. However, an RCT may not be feasible in some circumstances due to practical or ethical reasons, could be costly, or may not reflect real-world medical practice. Alternatively, observational (nonrandomized) comparative studies are playing a substantial role in device evaluations in both pre-market and post-market settings. The observational comparative studies could be conducted using 1) a concurrent (but non-randomized) control, 2) a historical control formed from patients with existing data collected from earlier studies of a previously approved device, or 3) a control extracted from a well-designed and executed registry database. In the current "Big Data" era, observational comparative studies could be performed using administrative data, such as the Centers for Medicare and Medicaid (CMS) claims database, to estimate treatment effects of intervention. But, various biases could be introduced at every stage and into every aspect of the observational study, and adversely impact the interpretation of the study results. For example, while treatment assignment is

determined in the study design of an RCT, it is frequently based on patient characteristics and determined by physician judgment or patient preference in observational studies with concurrent controls. This often leads to a systematic difference in the distribution of baseline covariates between interventional treatment group and control group, and this difference then result in bias in treatment effect estimation. There may be more differences in baseline covariate distributions between the treatment groups in the studies using non-concurrent controls, due to, for example, temporal bias caused by retrospective use of existing data or evolution of medical practice or technology, or a difference in definition and adjudication of clinical outcomes. All of these challenging issues lead to doubts about treatment group comparability, and hence the interpretability of study results and the ability of regulatory decision-making (Li & Yue, 2008). Fortunately, there exist some statistical methods that could be used to address some of such challenges, such as regression (covariate) analysis or propensity score methodology, introduced by Rosenbaum and Rubin (1983, 1984). However, these statistical methods can only adjust for observed confounding covariates but not for unobserved. Also, when there are substantial differences in baseline covariates between two treatment groups, the statistical methods may not be able to reduce bias. Therefore, minimizing bias should start from study design. This presentation highlights two areas: 1) minimizing bias in control group selection; 2) minimizing "opportunistic bias", described by Rubin, 2001.

2. Minimizing Bias in Control Group Selection

As in all observational comparative studies, there always exists risk that the treatment groups are not comparable even with concurrent control group. For example, one treatment group could have much sicker patients, or two treatment group patients may have different physiological factors. To reduce the treatment selection bias, treatment group comparability should be carefully considered in control group selection. One of the major concerns with using a historical control is the presence of potential temporal bias, due to rapidly evolving medical technology and/or learning effect of device use. For a historical control to be suitable for the regulatory purpose, the unexplained temporal bias needs to be minimal. If a control group is extracted from a national/international registry database, clinical comparability regrading patient population, treatment management, patient follow-up, definition of endpoints and clinical event adjudication should be thoroughly thought through. In a case that a control group is selected from OUS studies, the investigation of similarity in patient population and medical practice across multi-regions is critical in minimizing bias.

3. Minimizing Opportunistic Bias

Opportunistic bias occurs when an observational study is designed as if it was an RCT, and then outcome analyses are repeatedly performed after the study is completed, with both covariates and outcome data in sight. For example, Yue (2012) describes a premarket study with a historical control, in which, without a prospective study design, two fitted propensity score estimation models were submitted to the Food and Drug Administration (FDA): one with 10 of 35 covariates leading to a so-called "significant" outcome analysis result and the other with 15 covariates but "insignificant" outcome analysis result. As the propensity scores were estimated with both covariates and outcomes data in sight, it could be argued that more propensity score models had been

tried but results were not submitted to the FDA. Rubin (2001) points out "it is essentially impossible to be objective when a variety of analyses are being done, each producing an answer, favorable, neutral, or unfavorable to the investigator's interests". He further (2001, 2007, and 2008) advocates that the two appealing features of RCT mentioned in the Introduction can and should be duplicated for designing observational comparative studies. In doing so, propensity score methods, such as matching or stratification on propensity scores or inverse probability weighting using propensity scores, could be used to design observational studies in a way analogous to the way an RCT is designed: without seeing any outcome data. In regulatory settings, the outcome-free study design could be implemented in two stages (Yue et al, 2014).

Stage I - *Initial study planning by a sponsor.* This design stage mimics RCT planning that is performed by a sponsor and begins before the investigational study starts. The tasks performed in this stage include, but are not limited to, control group selection, baseline covariate identification, sample size estimation and power consideration, and propensity score method(s) to be used. In addition, some commitments need to be made: 1) identify an independent statistician who is masked to the outcome data of treatment and control groups and will perform the study design in Stage II; 2) establish firewalls to protect outcomes of treatment and control groups from leaking; and 3) be aware that there may be a need to change control group if treatment group incomparability is identified at the design Stage II, or to increase sample size if lower study power is noticed.

Stage II - *Approximating RCT by the independent statistician identified in Stage I.* This stage involves 1) developing the propensity scores model as a function of the baseline covariates and then estimating the propensity score for each patient; 2) checking that distributional balance of propensity scores between the two groups has been achieved; 3) assuring that the distributions of the covariates adjusted for propensity scores have been balanced between the two treatment groups; and 4) specifying a final statistical analysis plan for the treatment effect estimation with respect to clinical outcomes. This design stage should start as soon as all patients are enrolled, and the design should be accomplished by the independent statistician identified in the design Stage I, without access to any outcome data of either treatment group. Based on the resulting study design, the treatment comparability should be assured, and sample size and power should be re-evaluated. It is also critical to communicate and reach agreement with the FDA on the final study design at this stage.

The implementation of the two-stage study design process has started taking place in medical device submissions. The expected benefits of such a prospective design include: 1) avoidance of debates regarding "study design" at the final outcome analysis stage; 2) an increase in the integrity of study design and the creditability of study results; 3) an improvement in the consistency, transparency, predictability, and efficiency of regulatory review process; and 4) an increased flexibility regarding control group selection, sample size estimation and propensity score method(s) to be used. In addition, there are savings in time for both sponsors and FDA: the sponsor conducts the study design at Stage II during the patient follow-up period, and then simply analyzes outcome data accordingly after study concludes, rather than designing the study as well as performing outcome analysis after study ends; FDA then evaluates the outcome analysis results based on prospectively agreed study design and therefore reduces overall review time. Yue et al (2014) present a straw-man example to illustration the outcome-free study design process as follows.

The first stage design - A clinical study was proposed to demonstrate safety and effectiveness of a medical device through comparison to a control group to be selected from an existing registry. The clinical outcome variable was the success of the treatment, a binary outcome, and associated hypothesis was non-inferiority, with a non-inferiority margin set at 11%. At the first design stage, in total 15 baseline covariates that may affect the treatment assignment and/or the clinical outcome were identified, based on prior knowledge. An applicable registry was selected, and all of these key covariates and the endpoint information had already been planned to be collected in the registry. It was anticipated that at least 500 control subjects would be available. The design and analysis were planned to be based on the propensity score quintiles. The sample size in the investigational study for the treatment group was proposed to be 250. With this sample size and the assumption of equal success rate, $\pi_t = \pi_c$, for the treatment and control, at a significance level $\alpha = 0.025$, a power of about 80% could be obtained even under some imbalanced sample size distribution between two treatment groups across propensity score quintiles, such as the one exhibited in Table 1.

Table 1. Power calculation at the first design stage – based on a hypothetical scenario of 750 subjects (control: 500; treatment: 250)

	Propensity Score Quintile (*k*)					
	1	**2**	**3**	**4**	**5**	**Power**
$n_{c(k)}$	145	130	105	90	30	
$n_{t(k)}$	5	20	45	60	120	
$\pi_{(k)}$	0.87	0.87	0.87	0.87	0.87	82.8%
$\pi_{(k)}$	0.85	0.86	0.87	0.88	0.89	86.7%
$\pi_{(k)}$	0.89	0.88	0.87	0.86	0.85	79.0%

The second stage design – The design was conducted as soon as the enrollment of investigational study was completed. Based on the pre-specified inclusion/exclusion criteria of the investigational study, in total 1000 subjects from the registry were identified to be potential control subjects to be included in the study design and data analysis. The information of all 15 baseline covariates was available for subjects in both treatment and control groups. Based on the 250 subjects in the treatment group and 1,000 in the control, propensity scores were estimated using the logistic regression with all 15 baseline covariates in their linear terms. Multiple models with higher order terms were also tried, but not much difference was noticed in the design result. Thus, the simple model is presented here for illustration of control group selection. Five strata with equal sample size (quintiles) were then formed. The distribution of subjects by treatment group is listed in Table 2. It can be seen that that in the first propensity score quintile there were 250 control subjects but no treated subjects. Considering that the 250 control subjects looked nothing like any treated subjects, they could be reasonably discarded from the investigational study. However, any exclusion of treated subjects should be discouraged in a regulatory setting, because such exclusion may change the intended patient population. Propensity score re-modeling was then performed based on 250 treated subjects and remaining 750 control subjects (see Table 3). It is important to emphasize that the propensity score remodeling is valid only if the process is outcome free, i.e., selecting control subjects without access to any outcomes. Also, in a regulatory setting, an agreement with FDA on the final design is critical. Power estimation was examined (see Table 4) under different assumptions on success rates, all greater than 80%.

Table 2. Distribution of subjects at the five propensity score quintiles – based on 1250 subjects (control: 1000; treatment: 250)

	Propensity Score Quintile					
	1	**2**	**3**	**4**	**5**	**Total**
Control	250	244	234	186	86	1000
Treatment	0	6	16	64	164	250

Table 3. Distribution of subjects at the five propensity score quintiles – based on 1000 subjects (control: 750; treatment: 250)

	Propensity Score Quintile					
	1	**2**	**3**	**4**	**5**	**Total**
Control	196	193	172	128	61	750
Treatment	4	7	28	72	139	250

Table 4. Power calculation at the second design stage

	Quintile (*k*)					
	1	**2**	**3**	**4**	**5**	**Power**
$n_{c(k)}$	196	193	172	128	61	
$n_{t(k)}$	4	7	28	72	139	
$\pi_{(k)}$	0.87	0.87	0.87	0.87	0.87	86.7%
$\pi_{(k)}$	0.85	0.86	0.87	0.88	0.89	82.8%
$\pi_{(k)}$	0.89	0.88	0.87	0.86	0.85	91.1%

4. Summary

Minimizing bias in observational comparative clinical studies should begin in study design stage, particularly in control group selection and study design process. Propensity score methodology could help with the bias reduction, and its critical feature of separating study design and outcome analysis should be well utilized.

References

1. Rosenbaum, P. R. and Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** (1): 41–55.
2. Rosenbaum, P. R. and Rubin D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *JASA* **79**:516-524.
3. Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology* **2:**169–188.

4. Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallel with the design of randomized trials. *Statistics in medicine* **26**: 20-36.
5. Rubin, D. B. (2008). For objective causal inference, design trumps analysis. The Annals of Applied Statistics **2** (3): 808-840
6. Li, H. & Yue, L. (2008). Statistical and regulatory issues in non-randomized medical device clinical studies. *Journal of Biopharmaceutical Statistics*. **18**:20-30.
7. Yue**,** L. Q. (2012). Regulatory Considerations in the Design of Comparative Observational Studies Using Propensity Scores, *Journal of Biopharmaceutical Statistics* **22**: 1272–1279.
8. Yue, L., Lu, N. and Xu, Y. (2014). Designing pre-market observational comparative studies using existing data as controls: challenges and opportunities. *Journal of Biopharmaceutical Statistics* **24**:994-1010.