Practical Issues on the Observational Comparative Study Design Using Propensity Score Methodology in Pre-Market Medical Device Clinical Studies from the Regulatory Perspectives

Nelson Lu, Lilly Yue, Yunling Xu
Center for Devices and Radiological Health, US Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993

Abstract

Observational (non-randomized) comparative studies have often been utilized in the pre-market safety/effectiveness evaluation of therapeutic medical devices due to ethical or practical reasons. The comparators may come from data collected in earlier investigational device exemption studies or registry. To address the possible imbalance in patient characteristics between the investigational device group and the control group, propensity score methodology has been widely used to design and analyze these studies. In this paper, some practical issues and challenges will be discussed from regulatory perspectives. Topics include separation of design and analysis, types of estimands, selection of subjects, sample size estimation, and diagnostic checking in covariate balance.

Key Words: Observational comparative study, propensity score

1. Introduction

The randomized controlled trials (RCT) are viewed as a gold standard for causal inference. The data yielded from a well-designed and well-conducted RCT may provide the strongest evidence in evaluating the effectiveness and safety of an investigational device in the premarket setting. However, due to feasible or ethical reasons, conducting a RCT is not always practical. Alternatively, observational (nonrandomized) comparative studies may be utilized in device evaluation. The controls served in such studies may be the concurrent or non-current (e.g. historical). A commonly seen control is formed based on subjects whose data have been collected from a previous investigational device exemption (IDE) study. A control could also be formed from data collected in a high-quality registry database.

Unlike what can be expected in an RCT where distributions of all observed and unobserved baseline covariates are balanced between two treatment groups, same cannot be expected in an observational study. Usually, treatment selection and study outcomes would be influenced by subject baseline characteristics. Treatment effect may be estimated with large bias from an observational study. Proper adjustment needs to be performed to remove the bias.

A commonly used approach to remove the bias is propensity score methodology, which is introduced by Rosenbaum and Rubin (1983). The propensity score (PS) is the probability

of treatment assignment conditional on observed baseline characteristics. It is a one-dimension summary of the observed covariates. Propensity scores are often estimated using a logistic regression model where the response is the treatment assignment and predictors are baseline covariates. Subjects from both groups can then be matched up based on similar estimated PS. Examples of matching include K:1 matching and sub-classification. The sub-classification method is used in this paper in all of the illustrating examples. In such an approach, the subjects are first ranked based on their estimated PS, then they are divided into several strata (usually five or more) with balanced size. Treatments are compared on outcomes within each stratum, and the overall treatment effect is pooled across strata.

For comparative observational studies in the regulatory setting, separation of design and analysis is essential to produce valid evidence and to make study results more interpretable. Section 2 presents current recommendation on study development based on this principle. Section 3 discusses some practical issues, including what subjects should be included in the analysis set, selection of treatment effect estimands, sample size consideration, and diagnostic checking on covariate balance. The paper is concluded with summary in Section 4.

2. Current Recommendation on Study Development

One critical feature of RCT is that the study is designed without access to any outcome data. We believe that the very same feature should be maintained in an observational study in the regulatory setting. This is in line of what is advocated by Rubin (2001, 2007, and 2008): study design and outcome analysis should be separated. Based on this principle, the design should be carried out in two stages in the pre-market regulatory setting (Yue et al. 2014). The design in Stage 1 is conducted before the initiation of the investigational study. One main task is to plan the sample size of the investigational device group. The design in Stage 2 should be conducted as soon as subject enrollment is concluded and baseline covariates data are collected. An independent statistician should perform the task to estimate the propensity scores and to match patients from both groups based on the propensity scores. This may involve an iterative process until the balance is reached in the covariates distribution. Meanwhile, the statistical analysis plan should be finalized. Note that, in the study design phase, only the treatment assignment and baseline covariate data are needed. The outcome data are not required nor should be accessed.

3. Some Practical Issues

3.1 Subjects to be Included in Analysis Set

In the second stage design when the analysis method needs to be finalized, so is the analysis set. That is, the plan should identify what subjects are included in the analysis set. In an observational comparative study, oftentimes not all subjects from two treatment groups are comparable in the distribution of baseline characteristics. As a result, the estimated propensity scores are not well comparable. Therefore, subjects from both groups may be thrown away if the propensity scores distribution is quite different

between the two treatment groups. While such a practice is common and well accepted for general research and/or exploratory purposes, it may be inappropriate from the pre-market regulatory perspective where the study is confirmatory. The main regulatory concern, by leaving out any subjects treated with investigational device based on their estimated propensity scores, is the difficulties in defining the intended population. It would be difficult to identify the new intended population when subjects are deleted based on the unmatched propensity scores. On the other hand, not all the subjects in the control group need necessarily to be included in the analysis set. The intended treated population may not be affected by leaving out some control subjects.

Because of the regulatory constraint discussed in the previous paragraph, it would be problematic if there are no comparable control subjects for some subjects in the treatment group. It is then really important that control subjects provide good matches to the device group. That is, in terms of propensity score, the range of control group should cover the range of the subjects treated with the investigational device.

3.2 Treatment Effect Estimands

Following the framework set by Rubin (1974), each subject has two potential outcomes, $Y(T = 0)$ and $Y(T = 1)$, where $T = 1$ denotes that the subject is treated with investigational device and $T = 0$ denotes that the subjects is treated with control. The sample size is denoted as N_1 for the investigational device group and N_0 for the control group.

One commonly used estimand in pre-market medical device observational comparative clinical studies is average treatment effect (ATE), which can be expressed as

$$\text{ATE} = E[Y(1) - Y(0)].$$

Another commonly used estimand is the average treatment effect on the treated (ATT), which can be expressed as

$$\text{ATT} = E[Y(1) - Y(0)|T = 1].$$

In an observational study, ATT and ATE do not necessarily coincide, since the population for subjects treated with investigational device may not match with the overall population in practice. Further discussions of these estimands can be found in the literature such as Imbens (2004).

The point that the estimated ATT and ATE may differ is illustrated in the following hypothetical example.

Example: A total of 250 subjects treated with investigational device and 500 with control treatment are available, and are all included in the analysis set. Using the sub-classification method based on the estimated propensity scores, the sample size distribution and the observed treatment effect in each stratum are presented in Table 1. For example, a total of 11 subjects in treatment group and 139 subjects in control group have the lowest 20% estimated propensity score, as they are in the first stratum (quintile). The difference in the observed treatment effect based on these subjects is 0.3.

Table 1: Sample Size Distribution Among Strata With Observed Treatment Effect

Stratum	1	2	3	4	5	Total
N_1	11	19	39	67	114	250
N_0	139	131	111	83	36	500
Total	150	150	150	150	150	750
ATE weight	150/750	150/750	150/750	150/750	150/750	
ATT weight	11/250	19/250	39/250	67/250	114/250	
Observed Trt effect (δ)	0.3	0.25	0.2	0.15	0.1	

Based on this information, it can be computed that ATE = $\sum_{i=1}^{5} w_{ATEi}\delta_i$ = 0.2, and ATT = $\sum_{i=1}^{5} w_{ATTi}\delta_i$ = 0.15.

It can be observed that any large difference in the two estimands is likely due to the unbalanced distribution of sample size per arm and the heterogeneity in observed treatment effects across strata. The unbalanced distribution of sample size per arm across strata indicates that the distribution of treated population is distinct from that of overall population. This concludes the example.

As ATT may differ from ATE, it is important to indicate which one to be used in the second stage design under the regulatory setting prior to unblinding of outcome data. The main consideration in selecting between ATT and ATE is based on the objective, or the question intended to be answered.

The ATE should be considered if the interest is to get an answer to a question such as "What is the treatment effect on outcomes if all patients (eligible to both treatments) are only offered investigational device?" One common situation is that the older generation of a device is intended to be replaced by the newly developed version where the previous study was relatively recently conducted . Another possible situation occurs when the control treatment, serving as the current medical practice, will be potentially replaced with the investigational device.

A common question that is intended to be addressed is the following: "What is the treatment effect on outcomes in patients who select investigational device?" For such a situation where the investigational device is viewed as an alternative treatment option, the ATT may be more appropriate.

Selection of subjects (subjects to be included in the analysis set) may be closely related to the issue of selection. Under the regulatory requirement that all subjects treated with investigational device are needed to be included in the analysis set, ATT should always be able to be estimated. Certainly, this is under the assumption that the sample from the investigational study represents the treated population reasonably well.

On the other hand, it is possible that ATE may not always be reasonably estimated, depending on how the control subjects are selected into the analysis set. If control subjects are selected such that they cannot well represent the population of the control treatment, the combined subjects from both treatment groups in the analysis set may not well represent the overall population. Consequently, the validity of ATE estimate may be doubtful.

3.3 Sample Size Considerations

In first design stage, the sample size for the investigational device group needs to be proposed for the regulatory purpose. The sample size planning could be fairly challenging, and careful consideration may be demanded due to some uncertainties.

The major challenge is due to the uncertainty regarding the degree of comparability in subjects' characteristics between both groups. Poorer comparability may result in the requirement of a larger sample size to achieve a certain power. Generally speaking, reasonable good comparability may not be expected if there is significant time lag between the investigational study and the study where the control is obtained likely due to the evolve of medical practice and technology. In addition, it is not uncommon that the distributions of baseline characteristics from subjects treated with device differ greatly from those from subjects treated with medical management, even the two groups are conducted in the same time frame.

Another uncertainty is the unsureness of the matching method and statistical plan. The sample size determination generally depends upon the design and analysis methods. However, such plan may not be finalized in this stage; it usually is modified in the second stage design. Therefore, various potential designs and analysis methods may need to be considered in planning the sample size in this stage.

The sample size of control group is usually known for the historical control. However, when the control subjects are selected from a concurrent study, the sample size in the control group is unknown at this design stage. This may also add some complexity in the sample size planning.

To deal with these uncertainties, a more conservative approach is recommended. Many scenarios, in terms of different study designs and analysis methods, may need to be considered. The control sample size (from a concurrent study) may be somewhat underestimated. A larger sample size may be desirable in order to safeguard against the (unexpected) poor comparability and allow for greater flexibility in the second stage design.

A hypothesized example is provided to illustrate the point regarding the comparability between two groups.

Example: Suppose that a sponsor intends to demonstrate the non-inferiority of an investigational device to the medical management. As a registry of the medical management is available and can be served as a control, a comparative observational study is planned. It is expected that at least 350 control subjects will be available.

The outcome measure is assumed to be normally distributed. The non-inferiority hypothesis is listed in the following:

$$H_0: \mu_1 \leq \mu_0 - 0.2$$
$$H_1: \mu_1 > \mu_0 - 0.2.$$

This will be tested at the significance level at 0.05. The stratification method based on the propensity scores approach will be used. The estimand ATT will be used. The power is calculated based on the assumption that means of both treatment groups are equal.

In planning the sample size, a good starting point is to calculate the sample size based on a RCT design. With 350 subjects in each group, 84% power can be obtained. That is, a total of 700 subjects need to be enrolled in a 1:1 RCT.

Note that it is unknown, at this stage, regarding the distribution of subjects across strata, which is an indication of the comparability of subjects' baseline characteristics. A less balanced distribution generally yields a less power of the test. As powers vary with subject distributions, they should be evaluated accordingly with any proposed sample size in the device arm.

With the expected 350 control subjects in the intended observational comparative study, the highest power that can be achieved based on 350 subjects treated with the investigational device is 84%. Any distribution deviated from the even distribution yields a power less than 84%. Therefore, the sample size may need to be increased from 350 unless it is strongly believed that the subjects from two groups are highly comparable.

For illustration purpose, 350 subjects in device arm are doubled up to 700 subjects so that the number matches with the total sample size needed in the abovementioned RCT design. The powers are calculated based on three different distributions of subjects among quintiles. Table 2 presents the calculated powers based on an evenly distributed case in panel A, a mildly unevenly distributed case in panel B, and a greatly unevenly distributed case in panel C.

Table 2: Power Based on Different Sample Size Distribution Among Strata
(calculated based on $\mu_1 \leq \mu_0$)

Strata	1	2	3	4	5	Total	Power
A: Evenly distributed							
N_0	70	70	70	70	70	350	0.92
N_1	140	140	140	140	140	700	
B: Mildly unevenly distributed							
N_0	140	90	60	40	20	350	0.74
N_1	70	120	150	170	190	700	
C: Greater unevenly distributed							
N_0	160	110	50	20	10	350	0.51
N_1	50	100	160	190	200	700	

It can be observed that, although the power reaches 92% when distributions of subjects are even between arms, the power suffers greatly otherwise. It is to a point that, if subjects of two groups are greatly incomparable, a larger sample size may be required in the comparative observational study than that in a RCT.

This example illustrates that, when the sample size in the control group is somewhat limited, it is important that the control subjects need to provide relatively good matches. Otherwise, the study may be underpowered. On the other hand, if there are abundant control subjects, the lack of comparability may be less critical.

3.4 Covariate Balance Diagnosis

In second design stage, a propensity score estimation model needs to be developed. For a selected estimation model, the subjects can be matched up with the study design. It is important to assess whether the observed baseline covariates are balanced based on the selection. If the balance is not reached, alternative propensity score estimation model is needed. An iterative process between PS estimation/design and covariate balance assessment may need to be conducted until an appropriate PS model is identified. In submitting the selected propensity score estimation models and the grouping method to the FDA, it is important to illustrate the covariates balance is reached.

The covariate balance should be assessed in the same way as the design/planned analysis, as pointed out by, for example, Stuart (2010). When a 1:1 matching method is used, the comparison should be made between distributions from two groups provided that the matched pair feature is not accounted for. When the data analysis is based on matched pairs, the distribution of the differences in covariates may need to be checked. If the PS design is based on the sub-classification, the balance in covariates is established if the distributions within each stratum are similar between two groups.

Ideally, in assessing the covariate balance, the joint distributions of covariates should be examined between groups. However, it is burdensome and oftentimes impracticable. Therefore, from the regulatory perspective, the minimal requirement should include the assessment of the marginal distribution of every observed covariate that is identified in the first stage design.

Several diagnostic methods have been proposed in the literature and have been used by applicants. Here we present some of such methods.

Graphical presentation, such as Q-Q plot, box-plots, histograms, and Love plots based on the estimated propensity scores for each covariate are useful tools and can facilitate in the assessment of balance. However, it may require some judgmental call in a borderline case.

Some applicants present the c-statistic of the PS model in justifying the proposed PS estimation. The c-statistic is the area under the ROC curve, and it is to measure how well the PS model discriminates between two treatment groups. However, it is not a measure on whether the covariate is balanced. Therefore, it is not a valid measure and thus not recommended.

A popular approach is to use the hypothesis testing, such as testing for equivalence of means, to demonstrate the covariate balance. However, such an approach is debatable. Some researchers, such as Stuart (2010), Imai et. al. (2008), and Austin (2009), point out that that the balance is inherently an in-sample property, without reference to any broader population or super-population. The property that a test statistic and hence p-value is explicitly affected by the sample size also makes it less desirable to some researchers. An alternative approach is to use the statistic standardized (absolute) mean difference, which is expressed as the following:

$$|d| = \frac{|\bar{x}_1 - \bar{x}_0|}{\sqrt{(N_1 s_1^2 + N_0 s_0^2)/(N_1 + N_0)}},$$

where $\bar{x}_0$ and $\bar{x}_1$ denote sample mean of the baseline covariate for the control and device group, respectively; and s_0 and s_1 denote sample standard deviation for the control and device group, respectively. An advantage of using such statistics is that, with fixed sample means and standard deviations, it is not affected by the sample size.

If the covariate is balanced, the standardized absolute mean difference should be close to 0. In practice, a threshold value is set such that the balance is demonstrated if $|d|$ is less than the threshold value. In the literature, the popular choice of threshold values range from 0.1 to 0.25 (Rosenbaum and Rubin, 1985; Austin, 2009). Such threshold values appear to be reasonable when the sample size is relatively large. However, as the sample size for a typical medical device pre-marketing clinical study is relatively small, such threshold values may be too strict.

To see this, Table 3 lists some percentiles for the large sample distribution of $|d|$, assuming samples from both groups are independent and are drawn from the same distribution. These can be obtained based on the fact that d converges to $N(0, (N_0 + N_1)/N_0N_1)$, as pointed out in Hedges and Olkin (1985). It can be observed that, when the sample size is 1000 per group, it is highly unlikely to observe a value such as 0.2 for $|d|$. However, when the sample size is 50 or 100 per group, it is not uncommon to observe a value as high as 0.25.

Table 3: Percentiles for Large Sample Distribution of Standardized Absolute Difference With Different Sample Sizes

N_0	N_1	90^{th} percentile	95^{th} percentile	99^{th} percentile
50	50	0.33	0.39	0.52
100	100	0.23	0.28	0.36
1000	1000	0.07	0.09	0.12

A percentile of the distribution of $|d|$, such as 95^{th} percentile, can be proposed to serve as threshold value as illustrated in the following example.

Example

A clinical study was proposed to demonstrate safety and effectiveness of a cardiovascular device through comparison to a control group to be selected from an existing registry. More details of this straw-man example can be found in Yue et.al (2014).

A total of 250 subjects per group are included in the analysis. Sub-classification with five quintiles was used to match the subjects from both groups. Within each stratum, the standardized absolute difference $|d|$ for each baseline covariate are calculated. Meanwhile, the distribution of $|d|$ based on the assumption that all subjects from same normal distribution can be derived for each stratum. The 90^{th}, 95^{th}, and 99^{th} percentile of $|d|$ for each stratum are presented in Table 4.

Figure 1 plots the standardized absolute difference of each baseline covariate for five strata side by side. Within each stratum, the 95^{th} percentile of $|d|$, serving as the threshold, is vertically plotted. Since all $|d|$'s are below the respective threshold, the balance of covariates appears to be improved after the propensity score adjustment.

Table 4: Percentiles for Distribution of $|d|$ for Five Strata

Stratum	N_0	N_1	90th percentile	95th percentile	99th percentile
1	61	39	0.34	0.40	0.53
2	70	30	0.36	0.43	0.56
3	57	43	0.33	0.40	0.52
4	41	59	0.33	0.40	0.52
5	21	79	0.40	0.48	0.63

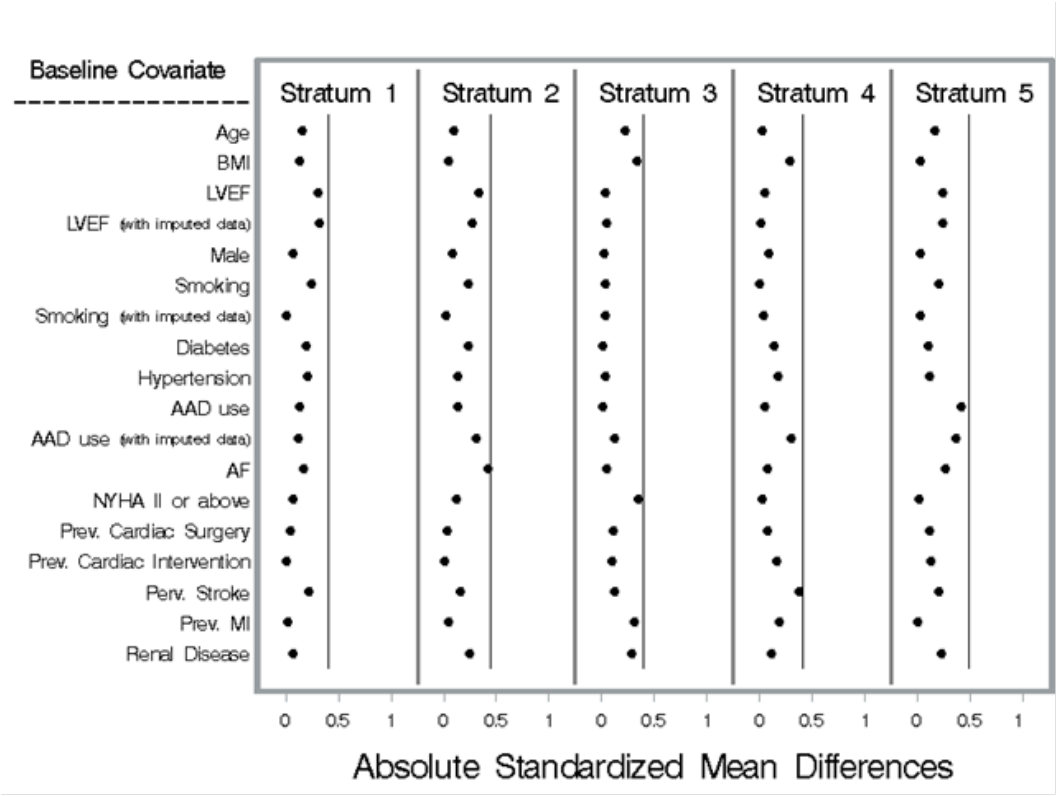

Figure 1: Standardized absolute mean difference of all baseline covariates for each stratum. Vertical line in each stratum is placed at the 95th percentile of distribution of |d|.

As the objective is to assess the similarity of the distribution of the covariate between two groups, only investigating equality of means for the continuous variable may not be adequate. Comparisons in higher order moments may be needed. The variance ratio is often used to compare the variance between groups. Refer to Austin (2009) and Imai et al. (2008) for further details.

4. Summary

Observational comparative studies have been utilized to support the premarket medical device approval. The propensity score methodology has often been used to address the

issue of covariate imbalance in such studies. Efforts should be put forth such that such studies should be designed to mimic RCTs. They should be prospectively designed following the principal, set by Rubin, that the analysis and design should be separated. That is, no outcome can be accessed during the design stage.

Several issues that may be encountered in the practice are discussed in this paper.
Average treatment effect (ATT) and average treatment effect on the treated (ATE) are two commonly used estimands for treatment effect. The selection regarding which to use should mainly depend on the objective and should be specified in the second stage design.

All subjects in device group need to be included in the analysis set, and control subjects to be included in the analysis set may not. However, when ATE is chosen to be the estimand, the control subjects should be selected such that the combined subjects should reasonably well represent the treated population

Sample size estimation in first stage design may be challenging due to various uncertainties such as level of comparability between control and investigational device groups, final analysis/group matching method, and control sample size. If the control subjects do not provide good matches and/or control sample size is limited, the sample size needed in the single-arm investigational study may even be more than what is needed in an RCT.

The true propensity scores are never known, so the selected propensity score estimation model need to be justified. Diagnostic checking in covariate balance is essential in demonstrating the appropriateness of the propensity score estimation model.

A RCT should remain the top choice in design selection. Although observational comparative studies using propensity score methodology may address some issues and provide a means to mimic RCT, there are certain limitations that cannot be overcome. It is possible that no appropriate control group can be found. Compared to RCT, much more design effort would be needed. The evidence based on such studies is generally not as strong as that based on RCTs.

References

Austin, P. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*: 28:3083-3107.

Hedges LV and Olkin I. (1985). Statistical Methods for Meta-Analysis. *Academic Press: San Diego, CA.*

Imai K, King G, Stuart EA. (2008) Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 171:481-502.

Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86: 4-29.

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

Rosenbaum, P. R., Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 39:33 – 38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688-701.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to thetobacco litigation. *Health Services & Outcomes Research Methodology* 2:169-188.

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.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics* 2(3):808-840.

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

Yue, Q.L., Lu, N., and Xu, Y. (2014). Designing premarket observational comparative studies using existing data as controls: challenges and opportunities. *Journal of Biopharmaceutical Statistics* 24:994-1010.