Statistical Study Design Considerations for Medical Device Clinical Studies: An FDA Reviewer's Perspective

Heng Li[1], Vandana Mukhi[1], Xu Yan[1]
[1]Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993

Abstract

In this paper, study design challenges frequently encountered in the planning stages of medical device clinical studies will be discussed. Issues in several types of study designs will be discussed from a statistical reviewer's perspective. It provides a framework to aid statisticians and preparers of pre-market submissions in deciding what information to include (and not to include) in the statistical sections.

Key Words: Medical device, study design

1. Introduction

In this short note we give some examples of statistical matters to keep in mind in designing clinical studies for investigational medical devices that are reviewed by US FDA Center for Devices and Radiological Health (CDRH). Paying attention to the specific issues that we will discuss could reduce the number of iterations in the whole review process, thereby saving time and resources. In Section 2 we will give an overview of general statistical issues in all types of designs. Sections 3 – 6 will each focus on a particular type of design, with Section 3 on randomized controlled trials, Section 4 on non-randomized comparative studies, Section 5 on single arm studies, Section 6 on adaptive designs. The last Section contains concluding remarks.

2. General Statistical Issues

A key element of designing a pivotal clinical study is selecting appropriate clinical endpoint(s) and pre-specifying the corresponding hypothesis of interest. It is important to clearly write-out the mathematical expression of the null and alternative hypothesis for each of the primary endpoints and explicitly state the test statistic that will be used for hypothesis testing. The verbal statement and mathematical expressions of the hypothesis need to be specified consistently. Let us give an example to explain what we mean: Suppose the primary endpoint for a study was defined to be major adverse cardiac events (MACE) at 1 year (a binary endpoint). The verbal statement of the hypothesis was specified in terms of comparing event free rates between two treatment groups. But the mathematical expression of the hypothesis was specified in terms of comparing survival functions between two treatment groups. The method of analysis for hypothesis testing was further specified as log-rank test. In such situations, it is unclear what the primary hypothesis of interest is: comparing two binary outcomes at 1 year or comparing two survival curves by 1 year? Only when the primary hypothesis of interest is clearly determined, can the appropriate hypothesis test be specified accordingly along with the test significance level.

For the purpose of interpreting study results obtained by pooling across sites, it is important to assess consistency of treatment effect across sites/geographic regions by conducting poolability analysis. In the statistical analysis plan, a plan to assess site poolability needs to be specified. Additionally, analyses that take into consideration variability between investigational sites may be necessary in evaluating a device. Usually an enrollment cap per site or region is also recommended to be specified. This helps eliminate the concern of data from one particular site dominating the study conclusion.

It is important that sponsor develops a plan to prevent or reduce the amount of missing data in a clinical study. The statistical analysis plan (SAP) should include details of missing data analysis methods that will be implemented to establish the validity of study conclusions. An assessment of the reasons for missing data, such as enrollment prior to treatment eligibility assessment, not all primary events were accurately measured due to the flaws of the design, large amount of missing data in the covariates of the historical control, patients unwilling to take the measurement due to the potential risk, can help provide important insights into the nature of missing.

It is recommended to clearly define and pre-specify the analysis populations in the protocol such as intent-to-treat (ITT), as treated (AT), and per-protocol (PP).

3. Randomized Controlled Trials

In general, double-blinded (double-masked), randomized, controlled, multi-center clinical trials provide the strongest level of scientific evidence and are considered the “gold standard”. The clinical protocol should provide adequate details about the randomization and blinding scheme. In some cases it may be appropriate to mask patient and follow-up evaluator to treatment assignment to help minimize bias; but note that this may not be always possible. In general, randomization is not recommended to be stratified by too many stratification factors, since it may result in zero patient enrolled in one treatment arm within some strata.

Sometimes clinical equipoise may not exist to randomize subjects to two treatment groups. In such situations, a non-randomized study with concurrent or non-concurrent/historical control group may be considered.

4. Statistical Considerations for Non-randomized Comparative Studies with Propensity Score Approach

Comparison to a non-randomized control group commonly relies on the use of propensity score methodology. In deciding whether to conduct such a non-randomized comparative study it is important to remember that while the propensity score technique can potentially balance observed baseline covariates, it cannot balance unobserved covariates like physician skills, different clinical practice at different sites, etc. Temporal bias is also difficult to correct for, that is why it is recommended that historical controls be taken from recent studies.

In implementing the propensity score approach it is critical to adopt the "outcome-free" design (Rubin, 2008). Yue, Lu, and Xu (2014) laid out the details of the outcome-free design process in the regulatory setting. Here outcome-free means that those who implement the propensity score methodology are not given access to outcome data of treated and control groups during the "design" stage. Yue et. al. (2014) advocate a two-stage design framework. The first design stage starts with the identification of the source of the non-randomized control group and the covariates to be used in the propensity score modeling. In choosing a control group one needs to make sure that all the pre-specified baseline covariates and clinical outcomes have been collected with minimal missing data, and they should have the same definition and measurement between the treatment groups. Subjects should be enrolled based on the same inclusion/exclusion criteria between the treatment groups. An independent statistician who has no access to outcome data is identified in the first stage to implement the "design" using propensity score methodology, which may include sample size estimation. This statistician needs to remain blinded to any outcome data throughout the entire process of study design. The first design stage should be completed before the investigational study starts.

The second design stage starts as soon as all baseline covariate data are collected, cleaned, and locked. In this stage, propensity score is estimated for each patient. The propensity score distributions in the treated and control groups are compared to see if they have sufficient overlap. Note that it is not a good strategy to "throw away" covariates when the distributions do not sufficiently overlap between the treatment groups, as covariates excluded from the propensity score model may not be balanced. If there is sufficient overlap between the propensity score distributions, matching or stratification can be performed. After that, covariate balance can be assessed. If some covariates are not well-balanced, then it is advisable to adjust the propensity score model by adding interaction and higher order terms. When the second design stage is completed, data may be prepared for outcome analysis. It is good idea to come talk to FDA before unblinding the outcome data.

Once outcome data is unblinded, outcome analysis can start. If propensity score matching is used, there is debate about whether outcome analysis needs to take into account the matched nature of the data (Austin, 2011, Stuart 2010). If propensity score stratification is used, then treatment effect is estimated as a weighted average across strata. One weighting scheme produces an estimate of the average treatment effect (ATE), which is the average effect of moving an entire population from control to the treated. Another weighting scheme produces an estimate of the treatment effect for the treated (ATT), which is the average effect of treatment on those subjects who ultimately received the treatment (Austin, 2011). When choosing a weighting scheme, it should be kept in mind that the estimand needs to be clinically meaningful.

5. Single Arm Studies with a Performance Goal

For a single arm study with a performance goal, the performance goal is used to evaluate whether the investigational device is safe and effective from the clinical perspective. Therefore, usually the performance goal is determined based on clinical judgment, and it is preferred that the performance goal be developed by a medical or scientific society. If the performance goal is dictated by the planned sample size, or developed based on the investigational device's previous own data, then even if at the end of the study this performance goal is met, it cannot help to address the clinical concern on whether the device is safe and effective. Under a single arm study design with a performance goal, it may not be appropriate to claim superiority to the performance goal nor is it appropriate to claim non-inferiority. Instead, it may be more appropriate to claim that the performance goal is met.

Sometimes the treatment effect may be expected to vary across different subgroups. As such, it may be difficult to develop a one-size-fits-all performance goal for the whole target patient population. Alternative approaches to develop the study design may be considered. For example, one may consider a randomized controlled trial if feasible, or alternatively consider narrowing down the target patient population to the subgroup that is clinically most important. If each subgroup of the target patient population can be clearly defined, one may consider analyzing each subgroup respectively with its own performance goal. This approach may provide sufficient information and straightforward interpretation on the performance of the device within each subgroup; however, it may not be cost efficient. If the target patient population consists of two mutually exclusive subgroups and the performance goal within each subgroup can be clearly determined, then a weighted performance goal on the whole target patient population may be considered. In this approach, the overall performance goal is determined by the weighted average of the performance goals from each subgroup. The weight is pre-specified, and usually it is the pre-specified proportion of the subgroup in the whole target patient population. The major concern for this approach is rejecting the null hypothesis does not necessarily imply that the performance goal within each subgroup is met. Therefore, this approach is not recommended.

6. Adaptive Designs

In studies with adaptive design features, the study modifications are recommended to be adequately pre-planned and described generally in the protocol, so that the scientific validity of the study results can be preserved. In order to minimize operational bias, the details of the adaptation algorithms are not recommended to be included in the protocol, but in a separate statistical analysis plan document. Complicated algorithms are not recommended, since they may be implausible to realize in practice.

In general, an unplanned modification to the study may weaken its scientific validity. The following examples illustrate several scenarios to be avoided: Unplanned proposal to increase the sample size after two interim analyses have been performed; Unplanned proposal to change the alpha spending function after the first interim analysis has been performed; Unplanned proposal to drop a treatment group and increase the alpha for the remaining treatment comparisons after the interim analysis has been performed.

7. Concluding remarks

We have sampled some study design considerations for medical device clinical studies submitted to FDA for review. We hope that they may be helpful for statisticians involved in those studies. Continuous dialogue among statisticians from industry, academia, and FDA is necessary to improve the quality of statistical sections of submissions

References

Austin, P. C (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies *Multivariate Behavioral Research*, **46**, 399–424.

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

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

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