Philippine Statistician, 2005
Vol. 54, Nos. 1-4, pp. 1-8

Statistical Section of a Clinical Trial Protocol

Karl E. Peace[1]

I. INTRODUCTION

Developing a good protocol[2] for any research study is imperative to the success of the study. All protocols should contain a statistical section which is statistically and scientifically defensible and which is sufficiently expository to communicate clearly to the reader all statistical aspects of the protocol. No *post hoc* analysis, however clever and thorough, can salvage a study, in the absence of good planning and execution. Over the years, I have organized the statistical section of a protocol into six subsections. These are: **study objectives** as statistical hypotheses, **endpoints, statistical methods, statistical monitoring procedures, statistical design considerations**, and **subset analyses**. General guidelines regarding the content of these subsections follow. Several other topics are addressed in each subsection.

II. STUDY OBJECTIVES AS STATISTICAL HYPOTHESES

The statistical or data analyses section of the protocol should begin by stating the specific questions which comprise the **study objectives**. These should be organized according to whether they address **primary efficacy, secondary efficacy** or **safety**. The specific questions should then be translated into statistical **hypotheses**. It is desirable from a statistical viewpoint for the alternative hypothesis (H_a) to embody the research question, both in substance and direction [1]. For placebo-controlled studies or for studies in which superior efficacy is the objective, this is routinely the case. For studies in which clinical equivalence is the objective, the usual framing of the objective translates it as the null hypothesis (H_o). In this framework, failure to reject H_o does not necessarily permit a conclusion of equivalence. This will depend on a specification of how much the treatment regimens may truly differ in terms of therapeutic endpoints, yet still be considered clinically equivalent, and the power of the test to detect such a difference. Some authors[2] have suggested reversing the null and alternative hypotheses for equivalence studies so that a conclusion of equivalence is reached by rejecting the null hypothesis. An attraction of this specification is that the Type I error is synonymous with the regulatory approval risk for efficacy and equivalence studies.

Separate univariate null and alternative hypotheses should be specified for each question. The reasons for separate specifications are primarily clarity and insight. Clarity because the questions have been clearly elucidated and framed as statistical hypotheses. This sets the stage for appropriate statistical analyses when the data become available. When analyses directed toward the questions occur, it should be clear whether the statistical evidence is sufficient to answer them. Insight is gained from the univariate specifications, as

[1] FASA, GCC Distinguished Cancer Scholar, Senior Research Scientist and Professor of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, PO Box 8148-01, Statesboro, GA USA 30460. Email address: kepeace@georgiasouthern.edu, peacekarl@cs.com

[2] See appendix for a brief outline

to the significance level at which the tests should be performed. This is true even though the study objective may represent a composite hypothesis.

As an example, suppose that there are three randomized groups in a duodenal ulcer study of a H2-receptor antagonist (X): placebo(A), 150 mg(B), and 300 mg(C) group. Further suppose that the objective of the study is to prove that 300 mg is effective and that it is more effective than 150 mg. There are two separate efficacy questions comprising the study objective: (i) Is 300 mg effective? (ii) Is 300 mg more effective than 150 mg? These two questions translate into the two univariate hypotheses:

$$\mathbf{H_{o1}: Pc = Pa \text{ versus } H_{a1}: Pc > Pa}$$

$$\mathbf{\text{and } H_{o2}: Pc = Pb \text{ versus } H_{a2}: Pc > Pb}$$

where Pa, Pb, and Pc represent the true proportions of patients treated with placebo, 150 mg of X, and 300 mg of X, respectively, whose ulcers would heal by the end of four weeks of treatment. The **primary study efficacy objective** is the composite hypothesis for which the null is the logical union of H_{o1} and H_{o2}, and the alternative is the logical intersection of H_{a1} and H_{a2}. It is therefore clear that if a Type I error of 0.05 were required on the experimental objective, then it would have to be partitioned across the two, separate, univariate hypotheses (questions) using Bonferonni or other appropriate techniques. Therefore, each question could not be tested at the 0.05 level of significance. The other possible pairwise comparison: 150 mg of X versus placebo, is not a part of the study objective. It may be investigated (preferably using a confidence interval), but it should not invoke a further penalty on the Type I error of the experiment. Further the global test of the simultaneous comparison of the three regimens is not of direct interest.

Secondary efficacy objectives should not invoke a penalty on the Type I error associated with the primary efficacy objectives. It may be argued that each secondary objective can be addressed using a Type I error of 5%, provided inference via significance testing is preferred. Ninety-five percent confidence intervals represent more informative alternative. Since the use of confidence intervals implies interest in estimates of true treatment differences, rather than interest in being able to decide whether true treatment differences are some prespecified values, confidence intervals are more consistent with a classification of secondary.

Safety objectives, unless they are the primary objectives, should not invoke a penalty on the Type I error associated with the primary efficacy objectives. It is uncommon that a study conducted prior to market approval of a new drug would have safety objectives which are primary. This does not mean that safety is not important. The safety of a drug, in the individual patient, and in groups of patients, is of utmost importance. Questions about safety are very difficult to answer in a definitive way, in clinical development programs of a new drug. There are many reasons for this [3]. There may be insufficient information to identify safety endpoints and/or the target population, and inadequate budgets or numbers of patients. Clinical development programs of new drugs should be aggressively monitored for safety within and across trials, but designed to provide definitive evidence of effectiveness. This position is entirely consistent with the statutory requirements [4] for new drug approval in the United States.

III. ENDPOINTS

After translating the study objectives into statistical hypotheses, the data analyses section should contain a paragraph which identifies and discusses the choice of **endpoints reflecting** the **objectives**. It should be clearly stated as to which endpoints reflect primary efficacy, which reflect secondary efficacy and which are safety related. An endpoint may be the actual data collected or a function of the data collected. Endpoints are the analysis units on each individual patient, which will be statistically analyzed to address study objectives. In an antihypertensive study, actual data reflecting potential efficacy are diastolic blood pressure measurements. Whereas it is informative to describe these data at baseline and at follow-up visits during the treatment period, inferential statistical analyses would be based upon the endpoint: change from baseline in diastolic blood pressure. Another endpoint of interest is whether patients experienced a clinically significant reduction in diastolic blood pressure from baseline to the end of the treatment period. Clinically significant is usually defined as a decrease from baseline of at least 10 mmHg. What constitutes the baseline measurement should also be clearly defined.

IV. STATISTICAL METHODS

After specifying the endpoints, the **statistical methods** which will be used to analyze them should be indicated. The methods chosen should be appropriate for the type of endpoint; *e.g.* parametric procedures such as analysis of variance techniques for continuous endpoints, and nonparametric procedures such as categorical data methods for discrete endpoints. Analysis methods should also be appropriate for the study design. For example, if the design has blocking factors, then statistical procedures should account for these factors. It is prudent to indicate that the methods stipulated will be used to analyze study endpoints, subject to actual data verification that any assumptions underlying the methods reasonably hold. Otherwise alternative methods will be considered. The use of **significance tests** is encouraged for **primary efficacy** questions, and **confidence intervals** for other questions. The method for constructing confidence intervals, particularly how the variance estimate will be determined, should be indicated.

Unless there are specific safety questions as part of the study objectives for which sample sizes with reasonable power to address them have been determined, it is usually sufficient to use descriptive procedures for summarizing **safety data**. Again this position is consistent with statutory requirements [4, 5]. If inferential methods are to be used, Edwards *et. al.* [6] provides a large variety including examples.

The last portion of the statistical methods section should address what methods will be used to **address generalizability** of results across design blocking factors or across demographic or prognostic subgroups. Most clinical trials require several investigational sites or centers in order to recruit enough patients. Randomization of patients to treatment groups within centers is the standard practice. Therefore, centers represent a design blocking factor. Age, gender, and race, for example, if not stratification factors, would not be design factors. However, it is usually meaningful to explore the extent to which response to treatment is generalizable across such subgroups. **Methods for generalizability** include descriptive presentations of treatment effects across blocks or subgroups, a graphical presentation of confidence intervals on treatment differences across blocks or subgroups, and analysis of variance models which include terms for interaction between treatment and blocks or subgroups.

V. STATISTICAL MONITORING PROCEDURES

Following the discussion of statistical methods, the data analyses section of the protocol should indicate what **monitoring procedures**, if any, will be followed in the study. This should include **data management procedures**, and any **procedures for monitoring** accumulating safety and/or efficacy data - regardless of whether they are formal, statistical, early termination procedures.

The **data management procedures** subsection should be brief. Basically, one wants to know the data trail and what procedures will be used to ensure quality of the data collected and to be analyzed. The quality of the data recording process at the investigational site is usually the responsibility of a field clinical monitoring group. The quality of the computerization of the data from the case report forms, after they get in house, to a study data base, and the extraction there from of data sets for statistical analyses, is usually the responsibility of a clinical data management group. The biostatistician is concerned with the quality of the data he/she analyzes and therefore should be familiar with company procedures for ensuring quality of the data so that a brief sketch of them can be included in the data analyses section of the protocol.

Successful clinical development programs, require **good clinical trial management.** An aspect of clinical trial management is staying current with enrollment, dropout, and completion rates of each clinical trial. This permits taking corrective action early on, if such is needed. Beyond interest in the progress of clinical trials, there is often strong interest in knowing safety and efficacy outcomes prior to study completion. These interests are genuine, and may represent a concern for patient safety, a need to plan future studies, or a desire to stop the study early to permit earlier filing of the registrational dossier. Whatever the reason, early looks at the data, particularly if unplanned and/or if sorted into treatment groups, may result in failure of the study to adequately address the objectives.

All studies should be monitored for safety. Ideally, this should be done on a patient by patient basis without knowledge of the treatment to which the patient was assigned. Group monitoring can be done if this is important to make a clinical decision as to whether the study should be stopped for safety reasons. In this case, it is often sufficient to separate the safety data into treatment groups without revealing group identity [3]. Since the design of studies of a new drug is almost always based upon efficacy considerations, it is not likely that monitoring for safety while a study is ongoing, will in itself, compromise (efficacy) study objectives. However, it is good practice to indicate in the protocol, what procedures will be used to monitor safety.

As indicated previously, most studies of new drugs are designed to provide answers to questions of efficacy. Therefore monitoring for efficacy while the study is in progress, particularly in an unplanned, *ad hoc* manner, will almost always be seen to compromise the answers. **If it is anticipated that the efficacy data will be looked at prior to study termination, for whatever reason, it is wise to include in the protocol an appropriate plan for doing this**. The **plan** should **address Type I error penalty** considerations, what steps will be taken to **minimize bias,** and permit **early termination**. The early termination procedure of O'Brien and Fleming [7] is usually reasonable. It allows periodic interim analyses of the data while the study is in progress, while preserving most of nominal Type I error for the final analysis upon scheduled study completion - providing there was insufficient evidence to terminate the study after an interim analysis. The paper [8] by the PMA Working Group addressing the topic of interim analyses provides a good summary of the concerns about, and procedures for, interim analyses.

VI. STATISTICAL DESIGN CONSIDERATIONS

The penultimate portion of the data analyses section of the protocol should reflect **statistical design considerations**. First of all, justification for the choice of experimental design, should be given. For example, if a crossover design was chosen, why is it appropriate for the disease under study? Then a thorough presentation of the basis for determining sample sizes should ensue. **Statistical inferences** (decisions with regards to whether the study objectives have been demonstrated) may be provided via **hypothesis tests** or via **confidence intervals**. These may require different **sample size determination** methods. Appropriate methods should be used. The tables of Fleiss [9] are usually sufficient for hypothesis testing methods. Whereas Makuch and Simon [10] or Westlake [11] provide confidence interval methods.

Hypothesis testing methods and confidence interval methods, require **estimates of endpoint** means and variances of the control group. These estimates may be obtained from the literature or from previous studies. It is good practice for the Biometrics Department to develop a file of such information from all studies of company compounds.

In obtaining such information, care should be taken to make sure that the information is on a population similar to the target population of the study protocol. If no such information exists, it may still be possible, particularly for dichotomous endpoints, to determine sample sizes by using the worst case of the variance.

Sample size procedures also require a **clinical specification of the difference ("delta")** between two comparative groups of interest which is clinically important to detect. For confidence interval procedures, the "delta" may be thought of as the bound on the allowable clinical difference. Hypothesis testing procedures require the **Type I error and the Type II error or the power** of the test to detect "delta" to be specified. Confidence interval methods require the confidence level (the complement of the Type I error) to be specified. They also require specification of either the maximum, allowable length of the interval or the degree of certainty of the coverage of the allowable clinical difference.

Sample size determinations yield the estimated numbers of patients required for analyses of efficacy. As such they represent the number of patients expected to be efficacy evaluatable. The numbers of patients who should be enrolled into the clinical trials, are obtained by dividing the numbers required for the efficacy evaluatable analyses by the expected proportions of those who enroll, who will be evaluatable for efficacy. In many clinical trials, the primary objective represents more than one question. Consequently, there will be more than one primary endpoint. To ensure that adequate numbers of patients will be enrolled, it is good practice to compute the sample size required for each question, or endpoint, then select the largest as the number to be enrolled.

VII. SUBSET ANALYSES

The last portion of the data analyses section of the protocol should address any **subset analyses** which will be performed. Clinical trials of efficacy may be considered as bioassays in patients to assess efficacy as a characteristic of the drug [12]. Questions about efficacy are best answered using data from that subset of patients who are efficacy evaluatable.

However, there is usually merit in performing analyses based upon all patients who entered the trial and who were randomized to treatment - provided they are in the study long enough to contribute follow-up data. This analysis is usually requested by regulatory agencies to see whether there were preferential reasons for excluding patients from the efficacy evaluatable analysis. A maxim is "analyze what you randomize." This analysis is commonly called the **intent-to-treat** analysis. It is good practice to include in the data analysis section of the protocol: (i) that the intent-to-treat analysis will also be performed; (ii) the identification of what endpoints will be analyzed in the intent-to-treat analysis; and (iii) the definition of what group of patients will contribute to this analysis. There should not be more than one intent-to-treat analysis, and not all endpoints need to be included in this analysis. Usually it is sufficient to include only the primary endpoints. It may be permissible to exclude some patients who were randomized from the intent-to-treat analysis. Patients who entered the trial in error, for example, those who did not have the disease under study, may be legitimately excluded, since "they were not intended-to-be treated."

Performing *post hoc* inferential analyses of demographic or prognostic subgroups of patients is hardly ever justified, statistically. However, if the analyses of generalizability reveal that treatment effects do not generalize across meaningful subgroups of the treated population, there may be a desire to perform inferential analyses within subgroups in an "effort to save the study." These may be performed with a caveat that they should not be viewed as confirmatory evidence of real treatment effects within subgroups.

References

1. Peace KE: "The Alternative Hypothesis: One-Sided or Two- Sided?" **J Clin Epidemiol**; Vol 42, pp. 473-6; 1989.

2. Hauck WW, Anderson S: A New Procedure for Testing Equivalence in Comparative Bioavailability and Other Trials." **Commun Stat Theor Meth**; Vol 12, pp. 2663-92; 1983.

3. Peace KE: "Design, Monitoring, and Analysis Issues Relative to Adverse Events." **Drug Info J**; Vol 21, pp. 21-8; 1987

4. Food and Drug Administration. "New Drug, Antibiotic, and Biologic, Drug Product Regulations; Final Rule." **21 CFR** Parts 312, 314, 511, and 514; Vol 52, No 53, pp. 8798-8857; Thursday, March 19, 1987.

5. Food and Drug Administration. "Guidelines for the Format and Content of the Clinical and Statistical Sections of New Drug Applications." **Center for Drugs and Biologics, Office of Drug Research and Review**, Rockville, MD; 1988.

6. Edwards S, Koch GG, Sollecito, WA: "Summarization, Analysis, and Monitoring of Adverse Events." In: **Statistical Issues in Drug Research and Development**, Peace, KE: Editor; Marcel Dekker Inc., New York; pp. 19-170; 1989.

7. O'Brien PC, Fleming TR: "A Multiple Testing Procedure for Clinical Trials." **Biometrics**; Vol 35 pp. 549-56; 1979.

8. The PMA Biostatistics and Medical Ad Hoc Committee on Interim Analysis. "Issues in Data Monitoring and Interim Analysis in the Pharmaceutical Industry." **Pharmaceutical Manufacturers Association**, Washington, DC; 1989.

9. Fleiss, J: "Statistical Methods for Rates and Proportions." 2nd Edition; John Wiley & Sons, New York; 1981.

10. Makuch R, Simon R: "Sample Size Requirements for Evaluating a Conservative Therapy." **Cancer Treat Rep**; Vol 62, 1037-40; 1978.

11. Westlake WJ: "Bioavailability and Bioequivalence of Pharmaceutical Formulations." In: **Biopharmaceutical Statistics for Drug Development**, Peace, KE: Editor, Marcel Dekker Inc., New York; pp. 329-52; 1988.

12. Peace KE: "Intention to Treat - What is the Question." In: **Statistical Issues in Drug Research and Development**, Peace, KE: Editor; Marcel Dekker Inc., New York; pp. 347-9; 1989.

APPENDIX: BRIEF CLINICAL PROTOCOL OUTLINE

I. BACKGROUND/ RATIONALE

II. OBJECTIVES

III. PLAN OF STUDY

3.1. Study Population
- 3.1.1 Demography
- 3.1.2 Criteria for Patient Inclusion
- 3.1.3 Criteria for Patient Exclusion

3.2. Study Design
- 3.2.1 Type of Study
- 3.2.2 Treatment Assignment
- 3.2.3 Blinding, Dosage, and administration of Study Drugs
- 3.2.4 Concomitant Medication
- 3.2.5 Procedures
  - 3.2.5.1 Pre-treatment period
  - 3.2.5.2 During-Treatment Period
  - 3.2.5.3 Post-Treatment Period
  - 3.2.5.4 Observers
  - 3.2.5.5 Data Recording
  - 3.2.5.6 Dropouts

3.3. Problem Management
- 3.3.1 Adverse Reactions
- 3.3.2 Criteria for Discontinuing Study Drug

IV. STATISTICAL OR DATA ANALYSIS

4.1 Study Objectives as Statistical Hypotheses
- 4.2.1 Criteria for defining Efficacy Endpoints
- 4.2.2 Criteria for defining Safety Endpoints

4.3 Statistical Methods
4.4 Statistical Monitoring Procedures, including Early Termination Plans
4.5 Statistical Design Considerations (including Types I, II errors, 'delta', variability estimate, experimental design)
4.6 Subset Analyses

V. ADMINISTRATION

5.1 Review and consent requirements
5.2 Record Keeping
5.3 Monitoring

VI. BIBLIOGRAPHY