Drug Information Journal, Vol. 31, pp. 1167–1174, 1997
Printed in the USA.
0092-8615/97

STATISTICAL/PRACTICAL ISSUES IN CLINICAL TRIALS

ANNPEY PONG, MS
Research Statistician, Biostatistics Berlex Laboratories, Inc., Montville, New Jersey
SHEIN-CHUNG CHOW, PHD
Executive Director, Biostatistics and Data Management, Covance, Inc., Princeton, New Jersey

For approval of a drug product, the United States Food and Drug Administration (FDA) requires that substantial evidence of the effectiveness and safety of the drug product be provided through the conduct of two adequate, well-controlled clinical trials. To assist the sponsors in preparation of final clinical reports for regulatory submission and review, the FDA and other regulatory agencies and conferences such as the International Conference on Harmonization (ICH) have developed guidelines for the format and content of a clinical report. The FDA and ICH guidelines require that the following statistical issues be addressed in the final clinical report: 1. Baseline comparability, 2. Adjustments for covariates, 3. Dropouts or missing values, 4. Interim analyses and data monitoring, 5. Multicenter studies, 6. Multiplicity, 7. Intention-to-treat subset versus efficacy subset, 8. Active control trials, and 9. Subgroup analyses. This paper provides an overview of these statistical issues. In addition, statistical justification for these issues is also addressed.

Key Words: Statistical issues: Clinical trials; ICH guidelines

INTRODUCTION

FOR APPROVAL OF A drug product, the United States Food and Drug Administration requires that substantial evidence of the effectiveness and safety of the drug product be provided through the conduct of two adequate, well-controlled clinical trials. The characteristics of an adequate, well-controlled clinical trial include a study protocol with a valid statistical design, adequate controls, appropriate randomization and blinding procedures, the choice of clinical endpoints for efficacy and safety, a strict adherence to the study protocol during the conduct of the trial, and a sound statistical analysis. These components are crucial for providing a scientific and unbiased assessment of the effectiveness and safety of the drug product. After the completion of the study, it is extremely important to summarize/interpret the clinical results for regulatory submission.

To assist the sponsors in preparation of final clinical reports for regulatory submission and review, the FDA developed guidelines for the format and content of a clinical report in 1988 (1). In addition, in 1994, the Committee for Proprietary Medicinal Products (CPMP) Working Party on Efficacy on Medicinal Products of the European Community issued a similar guideline entitled "A Note for Guidance on Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products" (2). At the same time, the International Conference on Harmonization (ICH) also signed off on the Step 4 final draft of the "Structure and Contents of Clinical Study

Reprint address: Annpey Pong, Berlex Laboratories, Inc., 340 Change Bridge Road, P.O. Box 1000, Montville, NJ 07045–1000.

Reports" and recommended its adoption to the three regulatory authorities of the United States, European Community, and Japan (3). Therefore, in this paper, the emphasis will be placed on the ICH guidelines for the format and content of a clinical report.

The ICH guidelines require that some critical statistical issues be addressed in the final clinical report. These critical issues include baseline comparability, adjustments for covariates, dropouts or missing values, interim analyses and data monitoring, multicenter studies, multiplicity, intention-to-treat subset versus efficacy subset, active control trials, and subgroup analyses (3). This paper provides an overview of these statistical issues for preparation of the clinical report after the completion of the study. In the following sections, these issues will be addressed separately. In addition, statistical justification for these issues is also discussed.

BASELINE COMPARABILITY

Baseline measurements are those collected during the baseline periods as defined in the protocol. Baseline usually refers to at randomization and prior to treatment. Sometimes, measurements obtained at screening are used as baselines. Basically, the objectives for analysis of baseline data are three-fold:

1. The analysis of baseline data is to provide a description of patient characteristics of the targeted population to which statistical inference is made. In addition, the analysis of baseline data provides useful information regarding whether the patients enrolled in the study are a representative sample of the targeted population according to the inclusion and exclusion criteria of the trial,
2. Since baseline data measure the initial patient disease status, they can serve as reference values for the assessment of the primary efficacy and safety clinical endpoints evaluated after the administration of the treatment, and
3. The comparability between treatment groups can be assessed based on baseline data to determine potential covariates for statistical evaluations of treatment effects.

The ICH guidelines require that baseline data on demographic variables such as age, gender, or race and some disease factors be collected and presented (3). These disease factors include:

1. Specific entry criteria, duration, stage and severity of disease, and other clinical classifications and subgroupings in common usage or of known prognostic significance,
2. Baseline values for critical measurements carried out during the study or identified as important indicators of prognosis or response to therapy,
3. Concomitant illness at trial initiation, such as renal disease, diabetes, and heart failure,
4. Relevant previous illness,
5. Relevant previous treatment for illness treated in the study,
6. Concomitant treatment maintained,
7. Other factors that might affect response to therapy (ie, weight, alcohol intake, and special diets), and for women, menstrual status and date of last menstrual period, if pertinent for the study.

ADJUSTMENTS FOR COVARIATES

For assessment of the efficacy and safety of a drug product, it is not uncommon that the primary clinical endpoints are affected by some factors such as demographic variables, patient characteristics, concomitant medications, and medical history. These factors are referred to as covariates which are also known as confounding factors, prognostic factors, or risk factors. If the covariates are known to have an impact on the clinical outcomes, one may consider stratified randomization according to these covariates. In practice, however, one may collect information on some covariates which may be influential and yet unknown at the planning stage of the trial. In this case, if patients are randomly assigned to receive treatments, the estimated treatment effect is asymptotically free of the accidental bias induced by these covariates. In other words, a randomized trial is asymptotically free of covariates imbalance, even for unknown and unmeasured covariates.

If the covariate is balanced, then the dif-

ference in simple treatment averages would be an unbiased estimate for the treatment effect. On the other hand, if the covariate is not balanced, then the difference in simple average between treatment groups will be biased for estimation of the treatment effect. In this case, it is suggested that the covariates be included in the statistical model such as an analysis of variance (or covariance) model for an unbiased estimate of the treatment effect. In the case where covariates are balanced between the treatment groups, it is still necessary to adjust for covariates for clinical endpoints in order to obtain valid inference of the treatment effect if the covariates are statistically significantly correlated with the clinical endpoints. Note that the key assumption for adjustment of covariates in estimation of the treatment effect is that the treatment does not affect the covariates. This assumption is easily satisfied by the subject-specific covariates such as demographic variables and baseline disease characteristics which are also time-independent.

The ICH guidelines require that selection of, and adjustments for, demographic or baseline measurements, concomitant therapy, or any other covariate or prognostic factor should be explained (3). In addition, methods of adjustments, results of analyses, and supportive information should be included in the detailed documentation of statistical methods.

DROPOUTS OR MISSING VALUES

There are many possible causes for the occurrence of dropouts and missing values. These possible causes include the duration of the study, the nature of the disease, the efficacy and adverse effects of the drug under study, intercurrent illness, accidents, patient refusal or moving, or other administrative reasons. Note that some of these causes for dropouts and missing values are treatment-related and some of them are not. The ICH guidelines suggest that the reasons for the dropouts, the time to dropout, and the proportion of dropouts among treatment groups be analyzed to examine the effects of dropouts for evaluation of the efficacy and safety of the study drug. Little and Rubin (4) classified missing values into three different types based on the possible causes (4). If the causes of missing values are independent of the observed responses, then the missing values are said to be completely random. On the other hand, if the causes of missing values are dependent on the observed responses but are independent of the scheduled but unobserved responses, then the missing values are said to be random. The missing values are said to be informative if the causes of missing values are dependent upon the scheduled but unobserved measurements.

If the missing mechanism is either completely random or random, then statistical inference derived from the likelihood approaches based on patients who complete the study is still valid. The inference is not as efficient, however, as it supposes to be (5). If the missing values are informative, then the inference based on the completers would be biased. As a result, it is suggested that despite the difficulty, the possible effects of dropouts and missing values on magnitude and direction of bias be explored as fully as possible. It should be noted that although some procedures for missing values have been proposed, no satisfactory and well-developed methodology exists for informative missing values. Performing intention-to-treat analysis based on all available data from all randomized patients regardless of whether the patients withdrew from the study is then suggested.

INTERIM ANALYSIS AND DATA MONITORING

Interim analyses and data monitoring are commonly employed for clinical trials in treatment of life-threatening disease or severely debilitating illness with long-term follow-up and endpoints such as mortality or irreversible morbidity. Interim analyses based on the data monitoring can be classified into formal interim analyses and administrative analyses (6,7). The aim of formal interim analyses is to determine whether a decision for early termination can be reached before the planned study completion due to

compelling evidence of beneficial effectiveness or harmful side effects. On the other hand, administrative interim analysis is performed in the pharmaceutical industry for reasons external to the trial such as requests from regulatory agencies or upper management. The administrative interim analyses are usually carried out without any intentions of early termination because of the results of the interim analyses.

The decision for early termination of clinical trials is crucial and cannot be made simply based on an application of one of many statistical procedures for interim analyses. The major concerns/issues regarding interim analyses and data monitoring are the potential bias of the inference of the treatment effect, the inflation of the false positive rate, and the documentation of the process. For example, since the results of interim analyses may change the subsequent conduct of the trial, a nonnegligible bias, which may not be measured or quantified, may be introduced. As a result, the generalizability of the results of the trial are in serious jeopardy. Therefore, it is important to make attempts/efforts to avoid/eliminate both known and unknown biases. An efficient way to avoid potential bias is to maintain blinding throughout the study.

O'Neill (8) classified the issues of interim analyses and data monitoring as planning, reporting, operation, and documentation of the trials. These issues include:

1. Unreported interim analyses,
2. Planned or unplanned interim access to unblinded comparative study results,
3. Failure of assessment of impact of unplanned interim analyses on study results,
4. Bias on the future conduct of the trial caused by unblinded access to study results,
5. The recognition by all relevant parties of the regulatory implications of early termination of trials,
6. Development of efficient, effective communication and information flow between the data monitoring committee and the regulatory authority,
7. Appropriate evaluation of exploratory trials,
8. Planning trials not to stop early for efficacy reasons alone but to balance the need for safety data on longer term exposure with short-term follow-up of early efficacy results, and
9. Establishment of policies regarding access to ongoing data, access to unblinded data, and participation in the decision making chain.

Since the process of examining and analyzing data accumulating in a clinical trial, either formally or informally, can introduce bias and/or increase type I error, the ICH guidelines require that all interim analyses, formal or informal, preplanned or ad hoc, by any study participant, sponsor staff member, or data monitoring group should be described in full, even if the treatment groups were not identified (3). Data monitoring without codebreading should also be described, even if this kind of monitoring is considered to cause no increase in type I error.

MULTICENTER STUDIES

A multicenter study is defined as a single study conducted under a common protocol which involves several centers where the data collected are intended to be analyzed as a whole. A multicenter trial is often conducted to expedite the patient recruitment process. The objective of the analysis of clinical data from a multicenter trial is two-fold. It is not only to investigate whether a consistent treatment effect can be observed across centers but also to provide an estimate of the overall treatment effect. Nevius (9) proposed a set of four conditions under which evidence from a single multicenter trial would provide sufficient statistical evidence of efficacy. These conditions are:

1. The combined analysis shows significant results,
2. There is consistency over centers in terms of direction of results,
3. There is consistency over centers in terms

of producing nominally significant results in centers with sufficient power, and
4. Multiple centers show evidence of efficacy after adjustment for multiple comparisons.

Although all of the centers in multicenter trials follow the same protocol, many practical issues are likely to occur (10). These practical issues include:

1. Some centers may be too small for reliable separate interpretation of results,
2. Some centers may be too big such that they dominate the results,
3. One or more centers look out of line with the others,
4. Some centers may show trends in the wrong direction, and
5. The existence of the treatment-by-center interaction.

As a result, a statistical test for homogeneity across centers is necessarily performed for detection of possible quantitative or qualitative treatment-by-center interaction. Gail and Simon indicated that the existence of a quantitative interaction between treatment and center does not invalidate the analysis by pooling data across centers (11). An overall estimate of the average treatment difference is statistically justifiable which provides a meaningful summary of the results across centers. On the other hand, if a qualitative interaction between treatment and center is observed, an overall or average summary statistic may be misleading and hence considered inadequate. In this case, it is preferable to describe the nature of the interaction and to indicate which centers contribute to the interaction.

MULTIPLICITY

In clinical trials, multiplicity may occur depending upon the objective of the intended trial, the nature of the design, and statistical analysis. The causes of multiplicity are mainly due to the formulation of statistical hypotheses and the experimentwise false positive rates in subsequent analyses of the data. As a result, statistical results should be interpreted with extreme caution because false positive findings may increase as the number of significance tests performed increases. Therefore, the ICH guidelines require that the overall type I error rate be adjusted to reflect multiplicity (2). Basically, multiplicity in clinical trials can be classified as repeated interim analyses, multiple comparisons, multiple endpoints, and subgroup analyses, which are discussed separately below.

Multiple comparison is usually referred to as the comparison among more than two treatment groups. In the interest of an α overall type I error rate, the commonly employed approach is probably the application of the Bonferroni technique. The concept of Bonferroni's technique is to adjust p-values for control of experimentwise type I error rate for pairwise comparisons. Bonferroni's method does not require that the structure of the correlation among comparisons be specified. In addition, it allows an unequal number of patients in each treatment group. Bonferroni's method works well when the number of treatment groups is small. When the number of treatment groups increases, however, Bonferroni's adjustment for p-values becomes very conservative and may lack adequate power for the alternative in which most or all efficacy endpoints are improved. In this situation, as an alternative, one may consider a modified procedure proposed by Hochberg (12). Hochberg's procedure is shown to be more powerful because it only requires one p-value smaller than α to declare one statistically significant comparison. Note that when comparing a number of treatments with a control, the procedure for a one-sided alternative proposed by Dunnett can be used (13). An overview of multiple comparisons in clinical trials can be found in Dunnett and Goldsmith (14).

In clinical trials, it is not uncommon that the efficacy of a test drug in treatment of a certain disease may be characterized through multiple clinical endpoints. Capizzi and Zhang classified the clinical endpoints into primary, secondary, and tertiary categories

according to the biological and/or clinical importance and the study objective (15). Since the sample size of a clinical trial is usually selected to provide a sufficient power for detection of a difference in some or all primary clinical endpoints, the issue of false positive and false negative rates caused by the multiplicity of multiple primary endpoints is of particular concern. Bonferroni's method can achieve its greatest power for the situation where the true treatment effect exists in only one of these multiple endpoints. In some early Phase II or Phase III studies, however, there are usually a large number of clinical endpoints which are highly correlated to one another. In this case, it would be helpful to obtain a single composite index from multiple endpoints to provide an overall summary of the efficacy evaluation. Several methods for construction of a composite index have been proposed (16,17,18).

Although the primary objective of clinical trials is to provide a valid and unbiased inference of the treatment effect for the disease under study, it is of great interest to investigate whether the treatment effect is consistent across some demographic factor such as age, gender, race, baseline disease severity, some prognostic factors or previous medical conditions and concomitant medications. The purpose of subgroup analyses is not for a definitive statistical inference of the treatment effect for each subgroup but rather for an exploratory identification of unusual or unexpected results. The difference between subgroup analyses and multiple comparison and multiple endpoints is that patients in subgroups stratified by different values of a covariate are different and the test statistics obtained from different subgroups are statistically independent. As a result, the adjustment of p-values is straightforward, which is given by $1 - (1 - \alpha)^{1/K}$, where K is the number of subgroups. If the sample size permits, the ICH guidelines suggest that important demographic or baseline value-defined subgroups should be examined for unusually large or small responses and the results presented, for example, comparison of effects by age, sex, or race, by severity or prognostic groups, by history of prior treatment with a drug of the same class, and so forth. Subgroups may suggest hypotheses worth examining in other studies or be helpful in refining labeling information, patient selection, dose selection, and so forth (3). Where there is a prior hypothesis of a differential effect in a particular subgroup, this hypothesis and its assessment should be part of the planned statistical analysis. On the other hand, in a note for "Guidance on Statistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medical Products," CPMP expresses a different view on subgroup analyses. The note suggests that a model including the covariates and the treatment-by-covariate interaction be analyzed to obtain an overall treatment effect rather than multiple separate analyses within strata defined by the covariates.

EFFICACY SUBSET

In clinical trials, despite the fact that there is a thoughtful study protocol, deviation from the protocol may be encountered during the course of the trial. In addition, it is very likely that patients will withdraw from the study prematurely before the completion of the trial due to various reasons. Patients who complete the study might miss some scheduled visits. As a result, which patients should be included in the analysis for a valid and unbiased assessment of the efficacy and safety of the treatment is a legitimate question to ask.

To provide a valid statistical inference, random assignment of patients to treatment is the key. This leads to the concept of the intention-to-treat analysis. In practice, the intention-to-treat analysis is usually considered the primary analysis for evaluation of the efficacy and safety in clinical trials based on all available data obtained from all randomized patients even though they never receive the treatment or they receive a different treatment than they were supposed to. In general, the intention-to-treat analysis is a conservative approach which can reflect the real clinical practice better than any other analyses. A serious bias may be introduced, however, if

no data are collected for patients who discontinued the study prematurely before the end of the study. Therefore, it is recommended that the primary clinical endpoints of the trial be evaluated at the time of withdrawal.

In addition to the intention-to-treat subset, many subsets of the intention-to-treat dataset may be constructed for efficacy analysis. These subsets include:

1. Patients with any efficacy observations or with a certain minimum number of observations,
2. Patients who complete the study,
3. Patients with an observation during a particular time window, and
4. Patients with a specified degree of compliance.

These subsets are referred to as efficacy subsets. The ICH guidelines require that efficacy subsets be analyzed to examine the effects of dropping patients with available data from analyses because of poor compliance, missed visits, ineligibility, or any other reasons. Any substantial differences resulting from the analyses of the intention-to-treat subset and the efficacy subset should be the subject of explicit discussion (3).

ACTIVE-CONTROL STUDIES

An active control trial is often considered an alternative to placebo control study for evaluation of the effectiveness and safety of a test drug with very ill patients or patients with severe or life-threatening diseases based on ethical considerations. The primary objective of an active control trial could be:

1. To establish the efficacy of the test drug,
2. To show that the test drug is equivalent to an active control agent, or
3. To demonstrate that the test drug is superior to the active control agent.

Pledger and Hall pointed out that active control trials offer no direct evidence of effectiveness of the test drug (18). The only trial that will yield direct evidence of effectiveness of the test drug is a placebo-controlled trial which compares the test drug with a placebo. If the test drug has previously been proved to be efficacious against placebo, then the objective of the active control trial is to show either that the test drug is equivalent to the active control agent or it is superior to the active control agent. Note that since the active control agent is known to be an effective agent, it is often assumed that showing that the two treatments are equivalent in an active control study is done by demonstrating that both treatments are effective. This assumption, however, is not necessarily correct and cannot be verified from the data obtained from the active control trial (18).

Temple recommended the following fundamental principle for active control trials: "If we cannot be very certain that the positive control in a study would have beaten a placebo group, had one been present, the fundamental assumption of the positive control study cannot be made and that design must be considered inappropriate" (19). Under this fundamental assumption, the following situations may be considered to support active control studies:

1. Where there is a retrospective review of known placebo-controlled studies of the proposed active control to show that the drug regularly can be shown to be superior to a placebo,
2. Where active control trials are conducted utilizing a similar patient population and similar procedure (eg, dose, dosage regimens, titration methods, response assessment methods, and control of concomitant therapy), and
3. Where there is an estimate of the effect size of the placebo response.

Along this line, the ICH guidelines indicated that if an active control study is intended to show equivalence between the test drug and an active control, the analysis should show the confidence interval for the comparison between the two agents for critical endpoints and the relation of that interval

to the prespecified degree of inferiority that would be considered unacceptable (3).

CONCLUSION

In clinical development of a drug product under investigation, the objective of most clinical trials is to evaluate the effectiveness and safety of the drug product. The FDA and other regulatory agencies require that substantial evidence on the effectiveness and safety of the drug product be provided before the drug can be approved. Substantial evidence can only be obtained through the conduct of two adequate and well-controlled clinical trials. An adequate and well-controlled clinical trial provides a valid and unbiased assessment of the effectiveness and safety of the drug product under investigation. Statistical issues discussed above have an impact on the validity of the statistical design employed and statistical methods used for analysis. The success of a clinical trial depends upon whether these issues, which have an impact on the statistical and/or clinical interpretation of the clinical results, can be addressed successfully.

REFERENCES

1. FDA. Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications. Rockville, MD: U.S. Food and Drug Administration; 1988.
2. Committee for Proprietary Medicinal Products. *A Note for Guidance on Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products*. Brussels, Belgium: CPMP; 1995.
3. ICH. *ICH Harmonized Tripartite Guideline on Structure and Content of Clinical Study Reports.* Recommended for adoption at Step 4 of the ICH Process on November 30, 1995 by the ICH Steering Committee.
4. Little RJA, Rubin DB. *Statistical Analysis with Missing Values.* New York, New York: John Wiley & Sons; 1987.
5. Diggle PJ, Liang KY, Zeger SL. *Analysis of Logitudinal Data.* New York, New York: Oxford Science Publications; 1994.
6. PMA. Issues in data monitoring and interim analysis in the pharmaceutical industry. The PMA Biostatistics and Medical Ad Hoc Committee on Interim Analysis. Pharmaceutical Manufacturing Association; 1989.
7. Williams GW, Davis RL, Geston AJ, Gould AL, Hwang IK, Mathews H, Shih WJ, Snapinn SM, Walton-Bowen KL. Monitoring of clinical trials and interim analyses for a drug sponsor's point of view. *Stat Med*. 1993;12:481–492.
8. O'Neill RT. Some FDA perspectives on data monitoring in clinical trials in drug development. *Stat Med.* 1993;12:601–608.
9. Nevius SE. Assessment of evidence from a single multicenter trial. Proceedings of the Biopharmaceutical Section of the American Statistical Association; 1988:43–45.
10. Lewis JA. Statistical issues in the regulation of medicine. *Stat Med.* 1995;14:127–136.
11. Gail M, Simon R. Testing for qualitative interations between treatment effects and patient subsets. *Biometrics.* 1985;41:361–372.
12. Hochberg Y. A sharper Bonferroni's procedure for multiple tests of significance. *Biometrika.* 1988;75: 800–803.
13. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc.* 1955;50:1096–1121.
14. Dunnett CW, Goldsmith CH. When and how to do multiple comparisons. In *Statistics in the Pharmaceutical Industry.* Buncher CR, Tsay JY, eds. New York, New York: Marcel Dekker, Inc.; 1995.
15. Capizzi T, Zhang J. Testing the hypothesis that matters for multiple primary endpoints. *Drug Inf J*. 1996; 30:949–956.
16. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics.* 1984;40:1079–1087.
17. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics.* 1987;43:487–498.
18. Tang DI, Geller NL, Pocock SJ. On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics.* 1993;49:23–30.
19. Pledger GW, Hall D. Active control trials: do they address the efficacy issue? Proceedings of the Biopharmaceutical Section of the American Statistical Association: 1986:1–7.
20. Temple R. Difficulties in evaluating positive control trials. Proceedings of the Biopharmaceutical Section of the American Statistical Association: 1983:1–7.