The Philippine Statistician, 2006
Vol. 55, Nos. 1-2, pp. 1-26

Sample Size Considerations in Clinical Trials Pre-market Approval[1]

Karl E. Peace[2]

ABSTRACT

An important exercise in the design of protocols for new drugs is determining the number of patients required to address study objectives. This is important on a per protocol basis, as well as for the entire clinical development plan. Adequate numbers of patients must be studied within and across the phases of clinical development to justify moving forward with each phase as well as to support regulatory dossier filing. The phases of clinical development and their objectives are reviewed and their connection to labeling discussed. Statistical requirements for sample size determination are presented and recommendations made relative to the size of trials in each phase of clinical development pre-market approval.

Key Words: Exposure Analysis, Long-term CHD Studies, Intent-to-Treat

I. INTRODUCTION

Quality clinical research must be well planned, closely and carefully monitored and conducted, and appropriately analyzed and reported. Greater attentiveness to detail at the design stage argues for greater efficiency at the analysis and reporting stages.

An aspect of good design of protocols for new drugs is determining the number of patients required by the clinical investigation to adequately address the objective. Not only is this important on a per protocol basis, adequate numbers of patients must be studied within and across the phases of clinical development to support regulatory dossier filing.

Although I could go directly to a presentation on computation of sample sizes, similar to what one would present in a Statistics 101 class, a quick review of the phases of clinical trials and their objectives is presented first. Then the clinical development plan and its connection to labeling are discussed. Then statistical requirements for sample size determination are presented; proceeding in this way sets the stage for the recommendations that are made relative to the size of trials in each phase of clinical development pre-market approval. Some philosophical, if not controversial issues then follow as well as concluding remarks.

II. PHASES OF CLINICAL TRIALS AND OBJECTIVES

Anyone who has had any involvement with the clinical development program for a new drug knows that the clinical trials comprising the program are categorized as Phase I, Phase II or

---

[1] The basis for this manuscript was an invited presentation given at the Annual Meeting of the Drug Information Association, June 1991, Washington, DC.

[2] FASA, GCC Distinguished Cancer Scholar, Founding Director, Center for Biostatistics and Professor of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, PO Box 8015, Statesboro, GA USA 30460. Email address: kepeace@georgiasouthern.edu, peacekarl@cs.com

Phase III. Although these categories may not be mutually exclusive (nor in some cases mutually exhaustive), there is general agreement as to what types of clinical studies comprise the bulk of the trials within each phase.

Phase I Trials:

Phase I trials may consist of "early Phase I" trials, early dose ranging trials, bioavailability or pharmacokinetic trials, or mechanism of action studies. Early Phase I trials represent the initial introduction of the drug in humans, in order to characterize the acute pharmacological effect. For most classes of drugs, healthy subjects are enrolled, in an attempt to reduce the risk of serious toxicity and to avoid confounding pharmacological and disease effects. The idea is to introduce the drug to humans without inducing acute toxicity.

Early dose ranging trials, often-called dose tolerance or dose titration trials, are also most often conducted in healthy subjects. Both the effects of single dosing and multiple dosing schemes are studied. The objective of these trials is to determine a 'tolerable' dose range, such that as long as future dosing remains in this range, no intolerable side effects or toxicities would be expected to be seen.

Early Phase I trials and early dose ranging trials don't establish nor quantitate characteristics of a drug. These studies have to be conducted first, so that acute pharmacological effects may be described, and a range of tolerable doses determined, which guide clinical use of the drug for later studies.

The primary objectives of Phase I bioavailability and pharmacokinetic trials are to characterize what happens to the drug once it's injected into the human body. That is, properties such as absorption, distribution, metabolism, elimination, clearance, and half-life need to be described. These trials also usually enroll healthy subjects and are often called 'blood level trials.'

Mechanism of action trials attempt to identify how the drug induces its effects. An example is the class of H2-receptor antagonists, such as cimetidine, ranitidine, famotidine and zanatidine, which by blocking the H2-receptor reduce the secretion of gastrin which in turn leads to a reduction of gastric acid production. Another example is the HI-receptor antagonist, seldane, which by blocking the HI-receptor reduces histamine release.

Bioavailability or pharmacokinetic studies and mechanism of action studies provide additional information so that the drug may be clinically used more effectively and safer in future studies.

Phase II Trials:

Phase II trials represent the earliest trials of a drug in patients. Patients should have the disease under investigation. Patients who enter such trials represent a relatively restricted yet homogeneous population. In some areas of drug development such as oncology drugs, Phase II trials are categorized as Phase IIA and Phase IIB.

Phase IIA trials may include clinical pharmacology studies in patients, and more extensive or detailed pharmacokinetic and pharmacodynamic studies in patients. Phase IIB trials are

controlled and represent the initial demonstration of efficacy and safety of a drug at the doses from the clinical pharmacology studies. Also of interest is to estimate the effective dose range, to characterize the dose response curve, and to estimate the minimally effective dose. Often it is difficult to distinguish between Phase IIB trials and Phase III trials, particularly in terms of objectives. The primary differences are the inclusion/exclusion criteria and the sample size.

Phase III Trials:

Phase III trials may be viewed as extensions of Phase IIB trials. They are larger and the inclusion/exclusion criteria may be less restrictive than those of Phase IIB trials. For a drug to proceed to the Phase III portion of the development program, it must be deemed effective from the Phase IIB program. At this stage, effectiveness has been indicated, but not confirmed.

The primary objectives of the Phase III program are to confirm the effectiveness of the drug in a more heterogeneous population, and to collect more and longer-term safety data. Information from Phase IIB, provides pilot data for the purpose of sample size determination in Phase III.

For the purpose of obtaining more safety data under conditions that better approximate the anticipated clinical use of the drug, relatively large, usually uncontrolled, non-comparative trials may also be conducted in Phase III. Since if the drug is given approval to be marketed, it may be used in the elderly, in the renally impaired, etc., and since such patients are usually excluded from other trials, studies in special populations may also be conducted in Phase III.

III. THE CLINICAL DEVELOPMENT PLAN – PRE-MARKET APPROVAL

As was stated to in section II, the clinical development plan for a new drug includes Phase I, Phase II, and Phase III trials. In viewing the types of trials within each phase of clinical development, it is obvious that the objectives of the trials describe characteristics of a drug that should be known before proceeding sequentially with subsequent clinical use. Further upon the successful completion of the trials through Phase III, sufficient information should exist for the drug to be approved and marketed.

The drug sponsor may wish to include other trials in the clinical development plan particularly to provide a marketing 'hook' for launch. Prior to formulating the clinical development plan the drug sponsor should formulate draft labeling - what is required and realistically desired to be said? The clinical development plan then serves as a blueprint for labeling.

The notion of developing draft labeling prior to embarking on the clinical development plan seems so rational and transparent. Yet at several meetings to review clinical development plans, when I've asked the question: "What do you want the labeling to say?" It surprisingly is answered ambiguously.

Basically, the labeling should communicate characteristics of the drug and give instructions for its use. Usually, the objectives of the trials described in Phase I, Phase II and Phase III, if

met in carrying out the attendant investigations, provide sufficient information to communicate the characteristics of the drug. However, since the population studied pre-market approval is likely to be more homogeneous than the user population post-market approval, and since inferences are based upon group averages, there may be insufficient information from the usual Phase I, Phase II, and Phase III program as to optimal clinical use of the drug, particularly in individual patients.

Therefore, drug sponsors may wish to have a Phase III 1/2 program directed more toward clinical use than toward establishing efficacy as a characteristic of the drug – which in my mind is what the typical pivotal proof of efficacy trials in Phase III provide. Such a targeted program may be unnecessary if more efficient and more optimal designs and methods, such as response surface methodology, and evolutionary operations procedures are incorporated into the clinical development program as early as Phase II. In addition, being proactive in developing an integrated database consisting of all data collected on a compound, so that meta-analysis and other techniques may be used, should enable the drug sponsor to do a better job at labeling.

IV. SAMPLE SIZE REQUIREMENTS

In presenting the statistical features of sample size determination, I assume that a protocol is being developed, and that the project statistician has the responsibility for recommending the appropriate statistical design, determining the sample size requirements, and providing the data analysis section of the protocol -- including appropriate statistical methods. I consider the basis for sample size determination to be a part of design considerations and should be a sub-section of the data analysis section of the protocol. There are other subsections of the data analysis section that need to be considered prior to the basis for sample size determination.

Protocol Objectives as Specific Statistical Questions:

The data analyses section of the protocol should begin by translating, the objectives into specific statistical questions. These should be organized according to whether they address primary efficacy, secondary efficacy or safety.

If inferential decisions regarding the questions are to be made on the basis of hypothesis testing, the questions should be translated into statistical hypotheses. It is desirable from a statistical viewpoint, for the alternative hypothesis (Ha), 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 (Ho). In this framework, failure to reject Ho does not 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 or consumers risk for both 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 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:

$$Ho_1: Pc = Pa \text{ versus } Ha_1: Pc > Pa$$
$$\text{and}$$
$$Ho_2: Pc = Pb \text{ versus } Ha_2: Pc > Pb,$$

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 study objective is the composite hypothesis for which the null is the logical union of Ho_1 and Ho_2, and the alternative is the logical intersection of Ha_1 and Ha_2. It is therefore clear that if a Type I error of 0.05 were required on the experimental objective, then it should be partitioned across the two, separate, univariate hypotheses (questions) using Bonferonni or other appropriate techniques.

Therefore, each question should not be tested at 5% level of significance. The other possible pair wise 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%, providing inference via significance testing is preferred. Ninety-five percent confidence intervals represent a 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 pre-specified 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 that 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.

Endpoints:

After translating the study objectives into statistical questions, the data analyses section should contain a paragraph that 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 is 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 that 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 mm HG. What the baseline is should also be clearly identified and defined.

Statistical Methods:

After specifying the endpoints, the statistical methods that will be used to analyze them should be indicated. The methods chosen should be appropriate for the type of endpoint; for example, 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 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 would 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 should be restricted to the primary efficacy questions (and then only if the study was designed from a power and sample size point of view to provide definitive answers). Otherwise, confidence intervals should be used. 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 that include terms for interaction between treatment and blocks or subgroups. The assessment of generalizability should follow the assessment of average drug effects across design blocks. If the results don't appear to be generalizable with respect to factors such as age, gender, race, etc., this should be reported, and the assessment of average drug effects redone with such factors appearing as covariates. The interactive effects of treatment group and such factors would not appear in this analysis model.

Statistical Design Considerations:

As mentioned previously, I consider the basis for sample size determination to be a part of statistical design considerations. Since the statistical analysis methods to be used to analyze the data to be collected for the protocol should be appropriate for the experimental design, 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 (i.e. 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 well known per group sample size (n) formula:

$$n >= 2\,(s^2/\delta^2)\,[Z_\alpha + Z_\beta]^2\,, \qquad (1)$$

where s is an estimate of the standard deviation, δ is the difference between groups which is clinically important to detect, and Z_α and Z_β are the appropriate critical points of the standard normal distribution corresponding to the magnitudes of the Types I and II errors, respectively. The tables of Fleiss [7] may also be used for hypothesis testing methods. Makuch and Simon [8], Westlake [9], and Bristol [10] provide confidence interval methods.

Hypothesis testing 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 Biostatistics 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 Bernoulli variance.

As is obvious from the previous sample size formula, sample size procedures require a clinical specification of the difference (δ) between two comparative groups of interest that is

clinically important to detect. For confidence interval procedures, the δ 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 δ 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 may 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, and then select the largest as the number to be enrolled, provided that all questions are of equal interest. Otherwise, use the number estimated to be adequate for the primary question, but assess the statistical characteristics for this sample size relative to the other questions.

Most pre-market approval 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 [ll] 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. Other procedures such as Pocock's [12], or Lan and DeMets[13] may also be used. The paper [14] by the PMA Working Group addressing the topic of interim analyses provides a good summary of the concerns about, and procedures for, interim analyses. The sample sizes for early termination, group sequential procedures, such as O'Brien and Fleming's, are determined as per fixed sample size procedures, and then this sample size is spread across sequential groups.

To summarize, the formal statistical basis for sample size determination requires: (i) the question or objective of the clinical investigation to be defined; (ii) the most relevant endpoints reflecting the objective to be identified; (iii) the specification of the difference between groups in terms of the endpoint that is clinically important to detect; (iv) specification of the magnitudes of the Type I and Type II errors; and (v) the mean and variability of the endpoint -- estimated from the literature or from previous studies.

Parenthetically, sometimes instead of estimates of the mean and variance of the endpoint being available, an estimate of the coefficient of variation (CV) is. This may be particularly

true in bioavailability or bioequivalence studies. The CV may be used instead of the mean and variance by expressing δ as a percent (of the mean). Once one has these ingredients, sample sizes may be determined rather easily.

Numbers in Phase I Program:

For clinical trials in the Phase I program, there is no statistical basis for sample size determination. During this phase, only gross estimates of some of the characteristics of the drug are obtained. The number of patients in each Phase I trial is based largely on clinical and/or scientific judgment or comfort. Each trial will usually have from 4 to 24 subjects. The entire program is not likely to have more than 100 subjects. Statistically, it is desirable to have some replication within different dosing levels or under different experimental conditions.

Numbers in Phase II Program:

For some trials in the Phase II program, there may be a statistical basis for sample size determination. For example, the data from the Phase I bioavailability or PK studies and from the acute pharmacology studies may provide pilot estimates for sample size determination for more detailed studies of these characteristics. Also, for some drugs, such as anticancer agents, some Phase IIA studies may be viewed as efficacy screens, in which case the single arm plans of Burdette and Gehan [15], Schultz [16] or Fleming [17] may be used. Further, if pilot estimates exist, it may be possible to statistically determine sample sizes for comparative phase IIB studies, but it is doubtful that one would want to design such trials to have large power.

In Phase II, much is still unknown about a new drug, and there is still the need to proceed cautiously. In Phase II, one is still trying to estimate characteristics of the drug. For single arm Phase II trials, my practice has been to encourage use of the group sequential plans referenced earlier and/or group sequential estimation plans. The sequential use of groups of patients with relatively small numbers in each group is consistent with proceeding cautiously. For comparative Phase II trials, my practice has been to recommend a sample size based upon 50% power, if sufficient pilot information exists. Parenthetically, for dose response or dose comparison studies, δ corresponds to the difference between the target dose (usually the middle dose) and placebo, and the one-sided Type I error critical point is determined similar to Williams' [18, 19] approach. Otherwise, I recommend trying to recruit approximately 50 patients per treatment group arm (parallel design) and assess a priori the statistical 'characteristics' for this number of patients.

The typical Phase II program will recruit only a few hundred patients in the entire phase.

Numbers in Phase III program:

There is a statistical basis for sample size determination in the Phase III program. The Phase II program should provide estimates of dose and frequency of dosing for the Phase III definitive proof of efficacy trials, as well as provide estimates of means and variability for the primary endpoints associated with the dosing regimens.

For the pivotal proof of efficacy trials, my practice has been to determine sample size per treatment group sufficient to provide 95% power to detect the clinically important δ with a one-sided Type I error rate of 5%. For other Phase III studies, such as studies in special populations, I usually recommend a power of 50%.

In determining the sample size, for a protocol with multiple questions, one must be able to decide whether the Type I error is an experiment-wise one which is to cover all questions and whether each question should be tested at some level less than 0.05, or whether each question may be tested at the 0.05 level. If for example there were three separate questions comprising the protocol objective, and it was decided that the results of the study would be positive only if all questions were answered positively, then using a (Bonferonni) Type I error level of 0.0167 instead of 0.05 would provide a conservative basis for sample size estimation. To be even more conservative, if the endpoint is dichotomous, the worst case of the binomial variance may be used. One would probably want to err in this direction if the estimate of variability from previous studies was not very precise.

The typical phase III program will recruit several hundred or a few thousand patients in the entire phase.

Other Sample Size Considerations:

1. Relative Size of Trials and Detectable Differences:

In the previous sections on numbers of patients in Phase II and Phase III trials, a 95% power was recommended for the definitive proof of efficacy trials and a 50% power for other trials. Many have responded when I've suggested using 50% power, "but that is a coin toss, so why does it matter how many patients we have? A power of 50% gives a value of 0 to the quantity Z_β (for standardized symmetric distributions) in the sample size formula (equation 1). So the sample size is basically being determined by the size of the Type I error, the estimate of variability and the size of the difference δ between groups, which is clinically important to detect. By rewriting the sample size formula, it is easy to see that it becomes the pooled-t statistic being greater than or equal to Z_α; that is, the decision rule for declaring δ to be statistically significant.

It is instructive to reflect the relative size of trials with power less than 95% to a trial with 95% power. Relative sizes corresponding to 50%, 75% and 80% power are summarized in Table 1 for a 2-sided and 1-sided Type I error rates of 5%.

Table 1: Size of Trials with Power less than 95% Relative to a Trial with 95% Power

Power Ratio	Relative Size(2-sided)	Relative Size(1-sided)
50/95	29.6 %	25.0 %
75/95	53.5 %	49.7 %
80/95	60.0 %	57.1 %

A trial with 50% power would be one-fourth the size of a trial with 95% power if a 1-sided alternative hypothesis were used, and would be about 30% as large if a 2-sided alternative were used. A trial with 75% power is about one-half the size of a trial with 95% power.

Many use a power of 80% with a 2-sided alternative in determining sample size for pivotal proof of efficacy trials. Sizes of trials corresponding to 80%, 88% and 95% power and a 1-sided alternative relative to a trial with an 80% power and a 2-sided alternative are summarized in Table 2 for a Type I error rate of 5%.

For 80% power, if a 1-sided alternative were used rather than a 2-sided one, 21% fewer patients would be required. Trials of the same size would have a power of 88% rather than 80%, if a 1-sided alternative were used instead of a 2-sided one. A trial with 95% power and a 1-sided alternative would require approximately 39% more patients than a trial with 80% power and a 2-sided alternative.

Table 2: Size of Trials with 1-Sided Ha and Various Power Relative to a Trial with 80% Power and a 2-Sided Ha

Sidedness Ratio	Relative Size
1s80/2s80	79 %
1s88/2s80	100 %
1s95/2s80	139 %

The primary use of statistical power is in planning the necessary size of a study to be conducted to detect a difference between groups that is of clinical importance. Often clinicians are overly optimistic in specifying the clinically important difference at the planning stage. It is instructive to reflect what kind of observed differences between groups, relative to the specified design clinically important difference δ, will be detectable as statistically significant once the trial has completed and we have the data for analysis. Sizes of the observed difference between groups as a percent of the design difference δ, which would be detected as statistically significant appear in Table 3 for various levels of power and for both 2-sided and 1-sided Type I error rates of 5%.

So a trial designed with 50% power to detect a difference δ will be able to detect as statistically significant ($p <= 0.05$) an observed difference of δ (or larger), assuming that the number of patients who provide data for analysis is the same as that from the sample size determination. On the other hand, were the trial designed with 95% power, an observed difference as small as 1/2 δ would be detected as statistically significant. Strictly speaking, these results hold only if the variance in the observed data is the same as the estimate used in the sample size computation. If the variance in the observed data were smaller (greater) than that used for sample size estimation, then smaller (larger) δ's than those in Table 3 would be detected as statistically significant.

In my experience, the design difference δ, is usually larger than the observed difference. It is good practice to design definitive trials with power much larger than 50%. As has been indicated previously, my choice of power for such trials is 95%.

Table 3: Size of Observed Difference between Groups, as a Percent of the Design δ that can be detected as statistically Significant (P<=0.05)

Power	Relative Size(2-sided)	Relative Size(1-sided)
50	100 %	100 %
75	74 %	70 %
80	70 %	66 %
95	54 %	50 %

An exception to this recommendation occurs in the area of trials conducted to support a Supplemental New Drug Application (SNDA). Here, I recommend that two identical trials, each with 75% power, be conducted. Each of these trials is about one-half the size of a single trial with 95% power. Each would be expected to detect as statistically significant a difference between groups as small as 70% δ. So that there is a good chance that statistical significance will be reached in each trial, thereby rigidly satisfying the requirement of substantial evidence from two adequate and well-controlled trials. However, I indicate in the data analysis section of the protocol for each trial that both trials will be analyzed as a single multi-centre trial. Even if not both the individual trials reached statistical significance, but the combined trials did and the individual trials demonstrated reproducibility, this should be sufficient evidence of efficacy for approval of a SNDA. A demonstration of reproducibility in two trials, even though not both trials reach statistical significance, in my mind, is consistent with the scientific basis for requiring two trials. Obviously, the analysis of the combined trials would have to show statistical significance in order for one to claim that a drug effect had been established.

2. Three Arm Efficacy Trial of New Drug, Placebo and Active:

Often a Phase III trial is conducted comparing a new drug at a given dose to placebo and to an active drug that is already on the market. In a placebo controlled, Phase III trial of a new drug, the question is about the (pure) efficacy of the new drug. In an active controlled, Phase III trial of a new drug, the question is about the relative efficacy, or clinical equivalence, of the new drug as compared to the active. Since the difference between the new drug and placebo ordinarily would be expected to be larger than the difference between the new drug and the active agent, thereby requiring fewer patients, the question arises as to what is a reasonable strategy with regard to sample size for the three-arm trial.

Strategy 1:

It is likely that the presence of the active control is to gain direct comparison information that can be used later for planning other studies, or for marketing purposes, and/or for an internal consistency check. The main objective is to prove that the new drug is effective, as compared to placebo. Thus, the sample size per group should be based upon the new drug versus placebo comparison. After the trial is completed, inferences concerning the effectiveness of the new drug compared to placebo should be based p-values whereas inferences concerning the relative effectiveness of the new drug compared to the active should be based confidence intervals.

Strategy 2:

If equal or greater interest is in comparing new drug to the active, then both the number per group Na necessary to compare the new drug to the active, and the number per group Np necessary to compare the new drug to placebo should be determined. The comparison of the new drug to placebo will be clearly over powered if Na patients are enrolled in each of the three arms. On the other hand, enrolling Np patients in each arm, while adequate for the new drug to placebo comparison, will clearly be underpowered for the new drug to active comparison. An alternative is to enroll Na patients into the new drug and active arms, and Np patients into the placebo arm. Since some power is lost in unbalanced allocation of patients to treatment groups, the power of the new drug to placebo comparison may be about what it would be in the balanced case. Of course exact computations can be made, and depending upon the size of Na and Np, it may be possible to enroll fewer than Np patients into the placebo arm and still maintain the same power as Np patients in each of the new drug and placebo arms would provide.

3. Interim Analyses:

For the many years, I've routinely recommended using group sequential, interim analysis procedures, such as those of O'Brien and Fleming[11] and Pocock [12], providing that recruitment relative to treatment, and resource allocation indicate that interim analyses are logistically feasible. Both procedures allow periodic interim analyses of the data while the study is in progress, and permit study termination at an interim analysis providing there is sufficient evidence of effectiveness. The O'Brien and Fleming procedure preserves most of the Type I error for the final analysis upon scheduled study completion - providing there was insufficient evidence to terminate the study after an interim analysis. This means that very little of the Type I error is allocated at earlier interim analyses, and consequently that treatment effects larger than expected would have to be observed for study termination to occur early. Pocock's procedure allocates the Type I error equally across the planned number of interim analyses.

One advantage of group sequential procedures is that on average they will require fewer patients than fixed sample size procedures. This is consistent with use of the procedures in the hope of being able to terminate a trial early. Terminating a trial of a new drug compared to placebo early, when efficacy is established, rather than going to the planned completion has ethical appeal.

Other reasons why I've recommended rather routine use of group sequential procedures are: (i) studies have to be monitored more closely; (ii) data management, including data entry and quality assurance has to occur on an ongoing basis, and data queries resolved quickly; (iii) the efficacy evaluatability assessment criteria have to be determined prior to the case report forms coming in house, and applied in an ongoing manner; (iv) report specifications have to be made prior to the study completing, and (v) statistical analysis programs have to be written and debugged prior to the first scheduled interim analysis. In other words, the technical aspects of good clinical research that one should be doing without incorporating interim analysis procedures have to be done when interim analysis procedures are incorporated. The difference is that we have to be more attentive and proactive; else the main purpose for incorporating interim analysis will be defeated.

There are some disadvantages, which are outweighed by the advantages. One is that greater resources may be required. For example, a trial incorporating 3 interim analyses and 1 final analysis, if not terminated early, will require more analysis and reporting resources than the same trial were interim analyses not incorporated. However interim analyses are usually conducted on one or a few variables and analysis of the full study data performed only if the decision is to terminate the trial early. Second, no inferential, interim analysis procedure would allow the full Type I error of 0.05 to be targeted to the final analysis if interim analyses were performed. Therefore for drugs with marginal effectiveness, statistical significance may not be attained in a trial incorporating interim analyses, when it may be achieved in the same trial were no interim analyses performed. It is good practice, for trials in which interim analyses are to be incorporated, to be designed with relatively large power. Third, interim analysis plans must be well developed and executed; else study objectives may be compromised.

Ideally, the group sequential, interim analysis plan would be included in the protocol. The sample size is determined as a fixed sample size and then spread across the number of analyses. For example, a two-group trial, with 95% power to detect a 20% difference between groups will require approximately 150 patients per group, for a total of 300 patients. If two interim analyses plus the possibility of a final analysis were planned, then the first analysis would occur after 100 patients had completed, the second (if necessary) after 200 patients had completed, and the third and last (if necessary) after 300 patients had completed.

Two issues other than sample size need to be addressed in the interim analysis plan: (i) preservation of the Type I error, and (ii) steps or procedures to minimize bias. The procedures of O'Brien and Fleming and Pocock, among others if followed, preserve the Type I error rate. Minimization of bias can also be achieved. In separate NDA and SNDA clinical development programs, I incorporated interim analysis procedures into the pivotal proof of efficacy trials, and both applications were subsequently approved.

Potential bias was minimized by using an external data management vendor. Identity of investigators and patients was concealed from in house personnel by codes generated by the vendor. Although, of necessity, the data were split into the randomized groups, the groups were identified in random order using labels of A, B, C, and D, whose identity was known only by the vendor. The trials in both programs were dose comparison in nature. Therefore, the vendor, by using a procedure similar to Williams' [18, 19], could assess whether significance was achieved and report this back to the sponsor without revealing the identity of each dose group. These programs will be discussed in greater detail in the next section.

One other type of interim analysis, sample size re-estimation, deserves comment. When one is unsure of the estimate of variability of the primary endpoint that was used to determine the sample size, accumulating data from the trial may be used to assess adequacy of the sample size variability estimate, and appropriate adjustments made to sample size prior to trial completion. Computer programs can be written to perform this exercise so that neither the sponsor nor analyst has to know the identity of the treatment groups. In fact the data should not be separated into treatment groups for the purpose of looking at group averages. As long as this is done, no Type I error penalty needs to be paid. One reason for this is that from normal theory, the sample mean and sample variance are independent. The papers by Shih [20] and Pedersen and Starbuck [21] may be seen for further discussion.

V. EXAMPLES

Four examples of clinical trials illustrating aspects of sample size determination are now considered. The first three trials involve the same formulation of a H2-receptor antagonist. One was a dose comparison trial, one was a bioequivalence trial, and the other was a trial in the elderly. The fourth example represents two identical dose comparison trials of a synthetic PGE2 analogue.

The dose comparison trial and the bioequivalence trial formed the basis o£ approval for a SNDA of a new formulation of the H2~receptor antagonist as a single nighttime dose in the treatment of acute duodenal ulcer. The two identical dose comparison trials formed the basis of approval of a NDA for the synthetic PGE2 analogue in the prevention of NSAID induced gastric mucosal ulcers in osteoarthritic patients. All dose comparison trials incorporated interim analysis procedures.

1. H2-Receptor Antagonist Duodenal Ulcer SNDA Program:

To illustrate the importance of numbers on the length of the clinical drug development process, consider three studies that comprised the major part of a program leading to the approval of a change in dosage (and form) of an already approved anti-ulcer drug. For the first of these studies [22], the original plan at the time I was consulted, called for two separate studies of 300 patients each. One study was to compare dose X to placebo, and the other was to compare dose 1.5 X to placebo. Together, the two studies were to recruit 600 patients at costs for investigators and patients of just over $4.5 million.

My recommendations were (i) to amalgamate the two studies into one, with placebo, dose X and dose 1.5 X groups, each with 164 patients; and (ii) perform an interim analysis at mid-study. The interim analysis would have looked at the two effectiveness comparisons to placebo. If each was effective, then the entire study could be stopped, if effectiveness were the only question. However, if additionally, dose discrimination was of interest, then the placebo arm could be stopped, and the two dose groups run to completion. A conservative estimate of savings would be approximately $1.5 million in investigator and patient costs plus time required to conduct the study.

To make a long story shorter, the final study consisted of 4 groups: placebo, dose 1/2 X, dose X and dose 1.5 X, each with 164 patients per group. The doses used in the trial were multiples of the 1/2 X formulation, which was already marketed. The objectives were:

(i) Prove that dose X is effective;
(ii) Prove that dose X is more effective than dose 1/2 X; and
(iii) Establish that dose 1.5 X is not clinically more effective than dose X.

The primary endpoint was the proportion of patients whose ulcers healed by the end of four weeks of treatment. The sample size of 164 patients per treatment group corresponded to a power of 95% to detect a difference in healing rates of 20% between dose X and placebo (which had an expected healing rate of 50%) with a Type I error rate of 1.67%. Since, the objective of the study consisted of three pair wise comparisons, a Bonferonni approach was taken to split the 0.05 Type I error level across the comparisons. Although, an interim

analysis was planned and performed, which showed very strong evidence of efficacy and dose response, the study continued to recruit, and completed 771 patients from 56 investigational sites. The final analysis was consistent with the interim analysis, in terms of effectiveness and dose response findings. Obviously, twice the amount of safety data was accumulated as would have been the case had the trial stopped early, and other results such as the relationship between healing and smoking habits and healing and ulcer size were better quantitated.

Two interesting aspects of the analyses of this trial should be pointed out. The first is that the interim analysis-stopping rule was based upon comparing the target dose (X) to placebo (the primary objective) using William's [18, 19] methodology. This approach preserves the Type I error and also permits concluding dose response. The second is that analyses were also performed using a model that blocked on 12 cross-classification of baseline ulcer size and smoking habits instead of blocking on investigational site. In studies with large numbers of investigators and small numbers of patients per investigational site, where one has measurements on strongly prognostic factors, it may be better to give up information on investigational site rather than on the prognostic factors.

The second of these studies was a blood level trial comparing the bioequivalence of a new formulation (Y), at dose X, to that of two doses of 1/2X, in the marketed formulation. The study was conducted as a two-by-two crossover with 24 normal volunteers. This number corresponded to an 80% power to detect a 20% difference between the mean AUC of the new formulation and the marketed formulation with a Type I error rate of 5% and a coefficient of variation of 34%.

One volunteer dropped out and was not replaced due to the concern that this would extend the date by which the submission could be made. Based upon the subjects who completed, the relative bioavailability of the new formulation to the marketed formulation was +/- 19% just within the acceptable bioequivalence range of +/- 20%. Had one fewer subject failed to complete, or had the study been designed smaller, may have required the study to be repeated, thereby delaying the submission.

The third and last of these studies was a clinical trial comparing dose X to placebo in 100 elderly patients. During protocol development, I argued for a sample size of 0 for this study, as we would likely have enough elderly patients from other trials to examine clinical response in the elderly population. Prior to completion of the elderly protocol, I subset the existing data base and found that among 101 elderly patients, 42 were on placebo or dose X. The comparison of these two groups revealed 95% confidence limits of 10.3% to 75.6%, in terms of ulcer healing – fairly convincing evidence that dose X was effective in the elderly. The take home message from this example is that a clinical trial may not be needed to answer every question of clinical interest, and conducting unnecessary trials may delay submissions.

2. Two identical studies in the Prevention of NSAID Induced Gastric Ulceration:

A few years ago, I had the responsibility of running a large-scale clinical research program of a synthetic prostaglandin (PGE2) analogue. Clinical and statistical evidence from the program formed the primary basis for NDA approval in the United States, of the drug in the prevention of NSAID induced gastric ulcers in osteoarthritic patients requiring NSAIDs in the management of their arthritic symptoms.

The clinical research program consisted of two identical protocols. Osteoarthritic patients who had upper gastric intestinal (UGI) pain and upon endoscopy were without gastric ulcer were randomized in balanced, double blind fashion to a placebo group, a 100-microgram-drug group, or a 200-microgram-drug group, administered 4 times daily. Patients were to return for follow-up endoscopy and other clinical evaluations after 4, 8 and 12 weeks of study medication administration, The objectives of the protocols were: (i) to demonstrate the effectiveness of the drug in the prevention of gastric ulcers; and (ii) to assess the effectiveness on UGI symptom relief.

The efficacy parameters were: (i) ulcer development, as confirmed by endoscopy at weeks 4, 8 or 12 weeks; (ii) UGI pain relief as derived from pain ratings recorded by the patient in a daily diary; and (iii) relief of other UGI symptoms. Of these, the prevention of ulcer development was primary. UGI pain was rated by the patient according to the following scales:

UGI Day Pain Rating Scale:
0 = None = I had no abdominal pain;
1 = Mild = I had some abdominal pain but it did not interrupt my normal activities;
2 = Moderate = I had some abdominal pain sufficient to interrupt my normal activities; and
3 = Severe = I had severe disabling abdominal pain.

UGI Night Pain Rating Scale:
0 = None = I had no abdominal pain;
1 = Mild = I had some abdominal pain but I was to go back to sleep;
2 = Moderate = I had abdominal pain sufficient to keep me awake for long periods; and,
3 = Severe = I had severe abdominal pain that kept me awake most of the night.

The ratings were recorded on a diary that was provided by the sponsor as part of the case report forms. The diaries were collected at each follow-up visit.

Per protocol sample size determinations revealed 450 evaluatable patients would be needed to address the primary objective. The numbers were determined on the basis of a 5% one-sided, Type I error rate and a 95% power to detect a 15% difference in ulcer development rates, given an expected ulcer rate of 25% in the placebo group.

The primary efficacy endpoint was the proportion of patients with ulcers by 12 weeks. The secondary endpoint was the proportion of patients without daytime or nighttime pain. The Mantel-Haenszel [23] or Fisher's exact test was (to be) used for statistical analyses of the endpoints.

No plans were provided in the protocol for any formal, statistical, interim analyses of the efficacy endpoints. We did however monitor the studies closely, and aggressively computerized the data. We knew on a weekly basis the status of the studies as to entry, completion, and ulcer development, without splitting the data into the three treatment groups. Table 4 summarizes such data at a about the halfway point, during the conduct of the studies.

Table 4: Status of Patients at Approximately Study Midpoint

Protocol	Patients Entered	Patients Completed
1	275	132
2	253	130
1&2	528	262

Ignoring study and treatment group and based upon patient information in the computerized database, we noticed that the incidence of ulcer development might range from a crude rate of 8.4% to a worst-case rate of 27.4% (Table 5).

Table 5: Ulcer Status of Completed Patients in Database

Patients	No Ulcer	Ulcer	Unknown	% Ulcer
215	156	18	41	8.4[1]
215	156	18	41	10.3[2]
215	156	18	41	27.4[3]

[1] Crude or best case estimate (an under-estimate)
[2] Reduced estimate
[3] Worst-case estimate (an over-estimate)

Parenthetically, comparable rates were also observed among patients whose case report form data had not yet been computerized (Table 6). However, all the ulcers could have been in one of the treatment groups. If this were the case, the incidence within that group could have been three times as high, or anywhere from 25.2% to 82.2%. We therefore felt compelled, on ethical grounds, to hold a meeting with the Food and Drug Administration (FDA) to discuss plans for performing an interim analysis of the studies, with the possibility of stopping the studies early.

Table 6: Ulcer Status of Completed Patients not in the Computerized Datadase

Patients	No Ulcer	Ulcer	Unknown	% Ulcer
43	34	5	4	11.6[1]
43	34	5	4	12.8[2]
43	34	5	4	20.9[3]

[1] Crude or best case estimate (an under-estimate)
[2] Reduced estimate
[3] Worst-case estimate (an over-estimate)

We met with the FDA, and discussed the data, our procedures for stopping the trials, collecting any remaining data, and statistical analyses. Among the information we presented at the meeting is that contained in Tables 4, 5 and 6, plus that contained in Tables 7 and 8.

Table 7: Ulcer Status of Completed Patients in Database: Possible Grouping Reflecting Dose Proportionality

Group	Patients	No Ulcer	Ulcer	Unknown	%Ulcer[1]
A	70	58	0	12	0[2]
B	71	53	6	12	8.5[3]
C	74	45	12	17	16.2[4]
All	215	156	18	41	8.4

[1] Crude or best case estimate (an under-estimate)
[2] Worst case = 17.1%; Reduced estimate = 0%
[3] Worst case = 25.4%; Reduced estimate = 10.2%
[4] Worst case = 39.2%; Reduced estimate = 21.1%

Table 8: Ulcer Status of Completed Patients in Database: Possible Grouping Reflecting Dose Proportionality: P-Values and Confidence Intervals

Comparison@	% Difference	Std. Er.	90% C.I.#	P-Value*
$B - A^1$	8.5	0.033	(3.1%;13.9%)	0.015/0.028
$C - A^1$	16.2	0.043	(9.2%;23.2%)	0.000/0.000
$B - A^2$	8.3	0.069	(-2.9%;19.6%)	0.162/0.304
$C - A^2$	13.8	0.064	(3.2%;24.4%)	0.003/0.005
$B - A^3$	10.2	0.039	(3.7%;16.7%)	0.014/0.027
$C - A^3$	21.1	0.054	(12.2%;30.0%)	0.000/0.000

= Normal approximation; * = Fisher's exact test (1-sided/2-sided); [1] = Best case; [2] = Worst case; [3] = Reduced.

Table 7 reflects 215 patients with 18 ulcers being split in a reasonably balanced way across three treatment groups, with numbers of ulcers per group reflecting a reasonable, but perhaps conservative dose response relationship. Table 8 reflects comparative analyses of the data in Table 7 using confidence intervals and Fisher's exact test (expected to be more conservative than the Mantel-Haenszel test).

It should be stressed that Table 7 represents a reasonable distribution of the total number (18) of ulcers under an assumption of dose proportionality. At the time of our meeting with the FDA, the blind had not been broken, nor had we separated the data according to blinded group labels. Since we had not planned to do a formal interim analysis at the protocol development stage, we wanted to make the case to the FDA, that we should perform an interim analysis on ethical grounds, and if dose response was observed, that we may be able to stop the studies early based upon a demonstration of prophylaxis efficacy. We wanted to be convincing that if an interim analysis was done, then it would be performed in a statistically valid, bona-fide manner.

There were three issues that received considerable discussion at the meeting with the Agency. These were: (i) when to terminate the trials? (ii) To what extent should blinding be maintained during the interim analysis? (iii) At what Type I error level should we conduct the interim analysis?

Concerning termination, three possibilities were considered: (i) terminate immediately; (ii) terminate based upon enrollment after 4 additional weeks; or (iii) continue accrual until the interim analysis was completed and then decide on the basis of that analysis. The first two of these possibilities exact no penalty on the Type I error provided we were prepared to live with the results. The third however, would, and is consistent with the philosophy for performing interim analyses.

Blinding considerations consisted of to what extent should investigators, patients and company personnel be blinded as to the results of the interim analysis? The primary concern was that if we failed to terminate the studies on the basis of the interim analysis results, that the study objectives would not be compromised by having performed the interim analysis.

As to the size of the Type I error for the interim analysis, we could take the O'Brien and Fleming approach and use 0.005, and if there was insufficient evidence to stop, allow the studies to continue to completion, and conduct the final analysis at the 0.048 level. Another possibility was to use a two stage Pocock procedure which would allocate a Type I error of 0.031 to each stage. Yet another possibility was to conduct the interim analysis at the 0.01 level with the final analysis being conducted at a level determined as per Lan and Demets [13] or Peace [24], if insufficient evidence existed for termination at the interim analysis.

The FDA was receptive to us performing an interim analysis subject to us providing them with written plans. Such plans should address the three issues noted above, as well as any others that would reflect positively on the scientific and statistical validity of the exercise.

We developed and submitted the plan to the agency. We addressed blinding considerations during the interim analysis so as to minimize bias. We selected a Type I error rate of 0.01. Our stopping rule was: terminate the trial if the P-value for the high dose group compared to placebo was less than or equal to 0.01. Parenthetically, it should be noted that the high dose being effective at the 0.01 level infers dose response via an argument similar to Williams' [18, 19]. In addition, the power of the combined interim analysis was about the same as each individual study at the design stage.

To make a long story shorter, we were able to terminate the trials, perform complete analyses, generate study reports, and compile the submission. Even though the interim analysis was not planned at the protocol development stage through attentive monitoring, and taking a proactive approach to clinical trial/data management, we were able to recognize that an interim analysis was justified on ethical grounds. By working prospectively with the U.S. Regulatory Agency, a bona-fide interim analysis was performed. This led to earlier termination of the program, and consequently, the submission was made and approved earlier than it otherwise may have been.

VI. PHILOSOPHICAL (IF NOT CONTROVERSIAL) ISSUES

In this section, five topics, which may be philosophical if not controversial, are considered. The first is what I've called "axioms of drug development [3]". The second is "sample size: efficacy or ethical imperative?" The third is "whether to have fewer but larger trials or greater but smaller trials [25]?" The fourth is 1-sided versus 2-sided tests [1, 26 - 30]. The

fifth is "amalgamation of Phase IIB and Phase III trials." Some thoughts I've had on these topics are presented in hopes of stimulating further discussion.

1. Axioms of Drug Development:

One of the major goals of clinical research and development of a new drug is to accumulate sufficient evidence of its efficacy and safety. When this has been accomplished, the registrational dossier may be compiled and submitted for a regulatory marketing approval decision. The sequential nature of the phases of clinical development together with the desire to accumulate sufficient evidence of the efficacy (a statutory requirement) and safety of a new drug suggest two axioms [3] of clinical drug development.

Axiom 1: Drugs in Clinical Development are considered inefficacious until proven otherwise.

Axiom 2: Drugs in Clinical Development are considered safe until proven otherwise.

These axioms may be translated into null and alternative hypotheses as follows:

Axiom 1: Hoe: The Drug is not Efficacious
Hae: The Drug is Efficacious,

and

Axiom 2: Hos: The Drug is Safe
Has: The Drug is not Safe.

The clinical development of a new drug will proceed until which time: (i) it is declared unsafe (rejection of Hos); or (ii) until it is declared inefficacious (acceptance of Hoe); or (iii) until it is proven to be efficacious (rejection of Hoe) and it has not been declared unsafe. From the hypotheses constructs, the risk associated with decision (ii) is a Type II error; and the risks associated with decisions (i) and (iii) are Type I errors.

Decision (i) is not likely to be reached based upon statistical analyses, and more often than not, it will be made prior to reaching Phase III. Decision (ii) could be reached in Phase II or Phase III, but most likely it will be reached in Phase IIB. Decision (iii) would be reached in Phase III, and basically represents the goal of Phase III.

So basically, unless the new drug sponsor decides to curtail clinical development on the basis of safety concerns, and/or on the basis of inefficacy in Phase IIB, clinical development programs will proceed into Phase III, and continue until either Hoe is accepted or Hoe is rejected -- decision (iii) is reached. As has been discussed previously in this paper, the way decision (iii) is currently reached is by having two adequate and well controlled trials both of which demonstrate statistical significance of drug effects.

What has greater appeal to me is to develop inferential sequential, statistical procedures that would permit efficacy to be determined based upon the cumulative information on efficacy. If the information on safety at that time does not contradict Hos, then let the regulatory dossier be filed, and hopefully reviewed and approved quickly.

At the time of termination of the development program based upon the demonstration of efficacy, it is unlikely that information would exist as to the optimal use of the drug. As a condition to approval, such studies could be conducted, and the labeling expanded. The attraction of this is that the drug would get on the market more quickly, and sales of the drug could begin funding research to learn more about the drug. The notion that adequate information on every possible characteristic of a drug has to be developed pre-market approval is unrealistic. After all, learning doesn't stop with submission of a regulatory dossier. Safety, for example, needs to be continuously monitored. The cumulative safety information that is available on a drug at one point in time is merely a snapshot of future safety information.

2. Sample Size: Efficacy or Ethical Imperative?

We design the Phase III pivotal proof of efficacy trials with large power. Apart from good science, we do so because it is imperative that we prove efficacy. Therefore, we could think of the determination of sample size as being mandated by an efficacy imperative. However, should there also be an ethical imperative? For example, should anything be said in the informed consent section of the protocol about adequacy of the sample size to address a medically relevant question? How many patients would enter a trial if they knew that there was only a 10% power, say, to detect the minimal, clinically significant difference?

3. Larger versus Smaller Trials:

Clinical development budgets are either fixed (or at least finite). For a fixed budget, particularly for Phase III, a larger number of smaller trials could conceivably be conducted for the same costs, as could a fewer number but larger trials. Which is better? Suppose for arguments sake, that the trials to be conducted will be of a new drug versus a control. Suppose further that whenever it is concluded based upon the results of a single trial, that the new drug is better than the control, the new drug will be added to the treatment armamentarium. The question "Which is better, fewer but larger trials, or greater but smaller trials?" may then be answered [cf. 26; and its references].

To do so, let (i) P denote the probability that a drug deemed superior from a clinical trial is in fact a superior drug; (ii) α denote the probability of concluding a false positive result; (iii) $1 - \beta$ denote the probability of concluding a true positive result; (iv) and r denote the ratio of the average number of false positive results to true positive results. Now r may be written as:

$$r = \text{Exp\#(fPR)/Exp\#(tPR)} = [(1\text{-}P)/P] \times [(\alpha)/(1\text{-}\beta)]. \quad (2)$$

Table 9 reflects values of r for various values of $(1 - \beta)$ and P, for a Type I error rate (α) of 0.05.

Table 9: Values of r for various values of $(1 - \beta)$ (Power) and P for a Type I error of 0.05

$1 - \beta$	P: 0.05	0.20	0.50
40%	2.38(70%)	0.50(33%)	0.13(11%)
80%	1.19(54%)	0.25(20%)	0.06(06%)
100%	0.95{49%)	0.20(17%)	0.05(05%)

The numbers within the parentheses to the right of the ratio values in Table 9, represent the percent of new drugs found superior to the control, which may be false positive results. For example, for new drugs with a intrinsic efficacy value of P = 0.05, which have been deemed superior from controlled trials with a power of 40%, 70% may in fact be false positive results, rather than true positive results, This number becomes 20% for new drugs with a value of P of 0.20 and trials with 80% power.

One notes that r is small whenever P is large and/or power is large. We have no control over P, and cannot easily estimate it [26]. However, we have control over the power of a study. Therefore, in the setting discussed, it is better to have fewer but larger trials, rather than more but smaller trials.

4. 1-Sided versus 2-Sided Tests:

Whether significance testing should be 1-sided or 2-sided has stimulated a lot of debate [1, 27 - 31]. Readers may wish to review these references. Briefly, my position is that the alternative hypothesis should embody the research question, both in substance and direction. Whether the inferences is 1-sided or 2-sided should follow accordingly.

In clinical efficacy trials of a new drug, the research question is "is the drug efficacious?" Therefore, the alternative hypothesis is directional (1-sided), particularly for placebo-controlled trials. If the trial is a confirmatory pivotal proof of efficacy trial, a 1-sided alternative is consistent with the trial being confirmatory. For it to be 2-sided says at the design stage, that you don't know the question you're trying to confirm, To use a 2-sided P-value for inference at the analysis stage, theoretically presents a multiple range test type of problem. Therefore, the results can't logically be viewed as confirmatory.

We should also have internal consistency with respect to directionality. For example, suppose we have a dose response trial with placebo, dose 1 and dose 2 of a new drug. The null and alternative hypotheses are: Hoe: Po = P1 = P2 versus Hae: Po < P1 < P2; where Po, P1, P2 represent the probability of responding while on placebo, dose 1, and dose 2, respectively. If however, the trial could only be conducted with the highest dose and placebo, then the null and alternative hypotheses should be: Hoe: Po= P2 versus Hae: Po < P2, rather than the alternative being Hae: Po =\= P2.

In my view, to operate with a 2-sided 5% Type I error level in placebo controlled efficacy trials of a new drug, is really operating with a 2.5% Type I error level.

5. Amalgamation of Phase IIB and Phase III Trials:

In some areas of drug development, for example in the development of drugs to treat some forms of cancer, the primary response measure in the Phase IIB program is different from the primary response measure in the Phase III program. For example for patients with advanced stages of disease, the usual primary measure of efficacy in phase IIB is response rate, whereas it is survival rate in Phase III. Typically, the Phase IIB and Phase III programs are conducted in different patients.

An alternative to this may be to design a large trial in which the goals of phase IIB and Phase III are amalgamated. We might envision a multi-stage group sequential trial, in which the goals of Phase IIB are addressed in early stages, and goals of phase III are addressed in later stages.

It is unclear whether there has to be a Type I error penalty paid on addressing the Phase III goals for having addressed the Phase IIB goals. A reasonable strategy would be to design the trial as a Phase III trial with respect to sample sizes and allocate the 5% Type I error across the stages in which the goals of Phase III are addressed. Such a plan would appear to save at least the number of patients usually included in Phase IIB.

VII. CONCLUDING REMARKS

The statistician, clinician, and upper management should understand that sample size estimation for pre-market approval studies is an important exercise. It should not be taken lightly nor as 'game playing.' However, one should realize that there is a need to balance numbers with practical considerations; but in so doing all involved need to understand the risks involved by going with smaller rather than larger studies. For example, a truly efficacious drug may be discontinued from further clinical development due to the results from a small trial, when the problem is low power rather than true inefficacy.

In addition, all research should be conducted with a total commitment to quality, and with imaginative and creative research and development teams, who aren't merely satisfied with adhering to status quo, but who will also incorporate innovative approaches, which will lead to the shortest possibly time for safe and efficacious drugs to be marketed and available to patients.

References

[1] PEACE, Karl E:. "The Alternative Hypothesis: One-Sided or Two-Sided?" J Clin Epidemiol; 42: 473-6; 1989.

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

[3] PEACE, Karl E: "Design, Monitoring, and Analysis Issues Relative to Adverse Events." Drug Info J; 21: 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; 52(53): 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] FLEISS, J: "Statistical Methods for Rates and Proportions." 2nd Edition; John Wiley & Sons, New York; 1981.

[8] MAKUCH R, Simon R: "Sample Size Requirements for Evaluating a Conservative Therapy." Cancer Treat Rep; 62: 1037-40; 1978.

[9] 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.

[10] BRISTOL DR: "Sample Sizes for Constructing Confidence Intervals and Testing Hypotheses." Statistics in Medicine, 8: 803-11; 1989.

[11] O'BRIEN PC, Fleming TR: "A Multiple Testing Procedure for Clinical Trials." Biometrics; 35: 549-56; 1979.

[12] POCOCK S: "Group Sequential Methods in the Design and Analysis of Clinical Trials. Biometrika; 64: 191-9; 1977.

[13] LAN KKG, Demets DL: "Discrete Sequential Boundaries for Clinical Trials." Biometrika; 70: 659-70; 1983.

[14] Pharmaceutical Manufacturers Association Biostatistics and Medical Ad Hoc Committee on Interim Analysis: "Issues in Data Monitoring and Interim Analysis in the Pharmaceutical Industry." 1991.

[15] BURDETTE WJ, Gehan EA: Planning and Analysis of Clinical Studies. Charles C. Thomas, Publisher: Spingfield, IL; 1970.

[16] SCHULTZ JR, Nichol FR, Elfring GL, Weed SD: "Multiple Stage Procedures for Drug Screening." Biometrics; 29: 293-300; 1973,

[17] FLEMING TR: "One-Sample Multiple Testing Procedure for Phase II Clinical Trials." Biometrics; 38: 143-51; 1982.

[18] WILLIAMS DA: "A Test for Differences between Treatment Means when Several Dose Levels are compared with a Zero Dose Control." Biometrics; 27: 103-17; 1971.

[19] WILLIAMS DA: "The Comparison of Several Dose Levels with a Zero Dose Control." Biometrics; 28: 519-31; 1972.

[20] SHIH WJ: "Sample Size Re-estimation in Clinical Trials." In: Biopharmaceutical Sequential Statistical Applications, Peace, KE: Editor; Marcel Dekker Inc., New York; 1992.

[21] PEDERSEN R, Starbuck R: "Interim Analysis in the Development of and Anti-inflammatory Agent: Sample Size Re-estimation and Conditional Power Analysis." In Biopharmaceutical Sequential Statistical Applications, Peace, KE: Editor; Marcel Dekker Inc., New York; 1992.

[22] VALENZUELA J, Dickson B, Dixon W, Peace KE, Putterman K, Young MD: Efficacy of a Single Nocturnal Dose of Cimetidine in Active Duodenal Ulcer. Post Grad Med; Nov 1985.

[23] MANTEL N, Haenszel W: "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease." J. Nat. Cancer Inst. 22: 719-48; 1959.

[24] PEACE, Karl E, Schriver RS: P-Values and Power Computation in Multiple-Look Trials." J. Chron. Dis. 40: 23-30; 1987.

[25] PEACE, Karl E "Regulatory or Consumers Risk" (Letter to the Ed.) . J. Clin. Epi. 43(9): 1013-4; 1990.

[26] PEACE, Karl E.: "One-Sided or Two-Sided p Values: Which most Appropriately Address the Question of Efficacy?" J. Biopharmaceutical Statistics. 1(1): 133-8; 1991.

[27] DUBEY S: "Some Thoughts on the One-Sided and Two-Sided Tests." J. Biopharmaceutical Statistics. 1(1): 139-50; 1991.

[28] FISHER L: "The Use of One-Sided Tests in Drug Trials: An FDA Advisory Committee Member's perspective." J. Biopharmaceutical Statistics 1(1) :151-6; 1991.

[29] OVERALL J: "A Comment Concerning One-Sided Tests of Significance in New Drug Applications." J. Biopharmaceutical Statistics. 1(1): 157-60; 1991.

[30] KOCH GG: "One-Sided and Two-Sided Tests and p Values." J. Biopharmaceutical Statistics 1 (1): 161- 9; 1991.