STUDY DESIGNS

IN BIOMEDICAL RESEARCH

STATISTICAL PLANS

(Statistical Analysis Plan & Interim Analysis Plan)

This lecture covers some preliminary or early analysis issues, those we have to deal with even before the start of the trial. These design-stage analysis issues include:

(1) Statistical Analysis Plan (SAP) which is often required, and

(2) Interim Analysis Plan which is sometimes desired, sometimes required

EARLY REQUIRED PLANS

Statistical Analysis Plan (SAP) for a clinical trial describes, in a written document, the planned statistical analysis of the trial in details; this is often required before the start of the trial.

Interim Statistical Analysis Plan describes the planned statistical interim analysis (or analyses) of the trial (and therefore needs to address handling of partial un-blinding issues in case of double blind trials). Details are included as part of the Trial Protocol.

STATISTICAL ANALYSIS PLAN (SAP)

Clinical Trials are complex scientific experiments designed to provide evidence to answer questions regarding the safety and efficacy of products. Furthermore data generated as part of these clinical studies are used for regulatory applications and/or communications of study results in manuscripts, marketing materials, or other symposia.

Study Protocol is the most important document describing all the details of the trial. The other companion document is a Statistical Analysis Plan (SAP).

The Statistical Analysis Plan (SAP) is one of critical importance. The SAP provides the details on the scope of planned analyses, population definitions, and methodology on how prospective decisions are to be made for presenting study results.

The Study Protocol has to be completed before the start of a Clinical Trial.

The Statistical Analysis Plan presents a more technical and detailed elaboration of procedures for executing the statistical analysis of variables involving in the primary and secondary specific aims. This document, the SAP, could be written later in the process but should be finalized before breaking the blind.

The SAP is critically important for documenting all the planned statistical analysis. The SAP is also stored in the trial master file and it is used during audits to check if statistical programming followed exactly the descriptions in the SAP. The SAP is meant to be a stand alone document. Besides the technical statistical details it should contain brief descriptions and summaries of the protocol. It should not only refer to the protocol.

Usually the SAP is written by the trial or project statistician by using a template. In most pharmaceutical companies the SAP will be written according to an available template which contains a standardized structure and which usually standardized and used in all clinical trials sponsored by the company. In general the SAP should give more details about the planned statistical analysis than the study protocol.

The SAP is meant to be a stand alone document. There may be a variety of templates but, in general, most templates have similar contents. We present below a typical temple as used by pharmaceutical firms; some of the sections are optional

SAP CONTENTS

1. INTRODUCTION

This includes a brief summary of the trial; being a stand alone document, it should has enough details without referring to sections of the Study Protocol.

2. DATA SOURCE

In this section, describe the data set or sets to be analyzed.

3. ANALYSIS OBJECTIVES

Briefly state the overall scientific objectives of the analyses, including the key unanswered questions that these analyses are designed to address. If necessary, provide additional detail to formulate the objectives in statistical terms. Include a brief summary of how each objective or specific aim will be addressed in the analyses.

SAP CONTENTS

4. ANALYSIS SETS/ POPULATIONS/SUBGROUPS

Include a brief definition of analysis sets/populations to be used including criteria for inclusion/exclusion for the population. Subgroups/subsets should be clearly defined and related back to the objectives stated above.

5. ENDPOINTS AND COVARIATES

Provide a brief definition of each type of endpoint, if different from those defined in the original protocol(s), indicating any use of visit windows and definition of baseline, as appropriate. In general, an endpoint should be defined by both a variable and a time point (eg, HIV-RNA viral load, change from baseline to week 24). If covariates are to be included in the statistical analyses, provide brief definitions/derivation rules.

SAP CONTENTS

6. HANDLING OF MISSING VALUES

Describe how missing values will be handled in the statistical analyses, and justify the methods used.

7. STATISTICAL METHODOLOGY

7.1 Statistical Procedures

Provide here the types of statistical tests to be used, with methods of stratification, types of sums of squares (if applicable), etc. If a formal meta-analysis is to be performed, then the model should be specified, including which terms are to be considered as fixed effects and which are to be considered as random effects. Sub-structure may be added in this section to break out methods, for example, by parametric or non-parametric/binary endpoints, or by types of data.

SAP CONTENTS

7.2 Measures aimed at Causal Inference & Adjustments for Multiplicity

Briefly describe and justify these measures if applicable. For example, in the setting of observational data analyses, one possible adjustment measure might be propensity scoring.

8. SENSITIVITY ANALYSES

If any sensitivity analyses are planned, then these analyses should be described and briefly justified here. Sensitivity analysis is the study of how the uncertainty in the outcome of the trial can be apportioned to different sources of uncertainty in its inputs. For example, how different sample size estimates or variance estimates affect statistical power.

SAP CONTENTS

9. RATIONALE FOR ANY DEVIATION FROM PREVIOUS SAP

If these analyses differ from those that were already propose d in previous version, provide brief rationale for the changes. This section is not needed for first version

10. QUALITY CONTROL (QC) PLAN

Provide a brief description of the QC Plan.

11. PROGRAMMING PLANS

Provide algorithms for generating tables and results to be executed by a Statistician.

12. REFERENCES

13. APPENDICES

INTERIM ANALYSIS PLAN

EARLY TERMINATION

- Typically, a Clinical Trial is designed as a single stage study in which patients are enrolled, treated, and are observed for outcome responses
- Accrued data are analyzed and recommendation is made at the end of the trial.
- However, for ethical reasons, the conduct of the trial sometime should allow for early termination if early results are extreme.

If early estimate of response to the new treatment is “very high” compared to Placebo control, the trial should be stopped, new treatment adopted, so that more patients could benefit from this good treatment (one with high efficacy)

The termination of a trial involving a poor agent, due to its low response rate, can be accomplished by a proper study design – as seen later in lectures on early phases clinical trials.

If early estimate of response is “very high”, trial should also be stopped, but this cannot be achieved by a design.

The current conventional approach is to reach early termination for poor efficacy by study design and to reach early termination for high/excellent efficacy by data analysis, interim analysis.

In order to reach a decision “early”, we would need to analyze data early – that is analyzing data more than once; at least once before the (planned) end of the study.

Each analysis leads to a decision by a statistical test of significance. For binary outcomes, such as “response”, the test is “Chi-square”; and one can apply the statistical test once or more than one times. Of course, we are all aware of the “multiple-decision problem”.

THE CONCERNS IN MULTIPLE DECISIONS

The central objective is to essentially preserve the “size” and the “statistical power” (involved in the decision to adopt or not to adopt the new treatment); those commonly pre-set for the conventional single-stage procedure

The most obvious/serious concern is the "size". Why?

To perform many tests increases the probability that one or more of the comparisons will result in a Type I error (test is significant but null hypothesis is true); For example, suppose the null hypothesis is true and we perform 20 tests---each has a 0.05 probability of resulting in a Type I error; then 1 of these 20 tests would be wrongly statistically significant simply as the results of Type I errors (false positives).

The main focus is the typical two-arm randomized, placebo controlled clinical trials. We also briefly show the procedure for Phase II trials which are often one-arm open-label.

RANDOMIZED TWO-ARM TRIALS

(O’Brien and Fleming, 1979)

USUAL SETTING FOR TWO-ARM PHASE II TRIALS

- Randomized clinical trials for comparing two treatments; phases II and III are treated similarly.
- Response is dichotomous and immediate.
- Single-phase, with sample sizes fixed in advance.
- At the end of the trial, compare “success rates” - i.e. proportions- using a formal test of significance based on the usual Pearson Chi-square test.

Aim:

A multiple testing procedure which provides the investigators with opportunity to conduct periodic reviews of the data as they accumulate and, thereby, offers the chance for early termination should one treatment prove superior to the other early on while continuing to use essentially the single-phase decision rule should early termination not occur.

O'BRIEN & FLEMING PROCEDURE

- Investigators plan to "test" N times, including the final comparison at the end of the trial
- Data are viewed periodically with m_1 subjects receiving treatment 1 and m_2 treatment 2 between successive tests; total $N(m_1 + m_2)$ subjects.
- Want to maintain an overall size α, say $\alpha = .05$
- **Rule:** After the nth test, $1 \leq n \leq N$, the study is terminated and H_0 is rejected if $(n/N)X^2 \geq P(N,\alpha)$ where X^2 is the usual Pearson Chi-square statistic.

RESULTS

Using the theory of “Brownian motion”, O’Brien and Fleming obtained the values for $P(N,\alpha)$ but, more importantly, they concluded that they are approximately the $(1-\alpha)$th percentile of the Chi-square distribution with 1 degree of freedom - <u>almost</u> independent of N, the number of tests.

EXAMPLE #1

- Take N=2 (one "interim" analysis) and $\alpha = .05$, P(2,.05)= 3.928 ("exact")
- Rule: reject the Null Hypothesis after the interim analysis if: $X^2 \geq (2)(3.928) = 7.86$ which is equivalent to having the p-value less than .005 and reject the Null Hypothesis after the final analysis if: $X^2 \geq (1)(3.928)$ which is equivalent to having the p-value less than .045

.

H_0 is rejected if $(n/N)X^2 \geq P(N,\alpha)$

A Simple Application of Procedure:

(1) Use approximate value (the 95th percentile of the Chi-square distribution with 1 degree of freedom, i.e. 3.84) instead of $P(N,\alpha)$ (i.e. 3.928)

(2) Calculate "cut-point for p-value" for the interim analyses (In application of this rule, one would assign .5% to the interim analysis) and subtract them out of the planned size (say 5%) to obtain "cut-point for p-value" for the final analysis.

(3) In other words, we can use usual Chi-square tests at .5% and 4.5% respectively.

EXAMPLE #2

- Take N=3 (two "interim" analyses) and $\alpha = .05$, Rule: Reject the Null Hypothesis after the first interim analysis if: $X^2 \geq (3)(3.84) = 11.52$ which is equivalent to having the p-value less than .001; after the second if $X^2 \geq (3/2)(3.84) = 5.76$ which is equivalent to having the p-value less than .014
- In application of this rule, one would assign .1% and 1.4% to the interim analyses and 3.5% to the final analysis for an overall 5% size.

H_0 is rejected if $(n/N)X2 \geq P(N,\alpha)$

The problem and the rule were formulated assigning constant (m1 + m2) subjects accrued between successive periodic tests. However, O'Brien and Fleming's simulations showed that their conclusions/results virtually remain valid in the more general settings. In other words, the only major factors affecting the rule is the number of tests N and the overall size α. The number of test N affects the rule but not P(N,α).

In practice most use N=2 (one interim analysis) or N=3 (two interim analyses), most often with just one interim. And the rules have been adopted for use with other statistical tests, for example, two-sample t-test instead of Chi-square test. The p-value distributions are:

(1) One interim: 0.5% and 4.5%

(2) Two interims: 0.1%, 1.4%, and 3.5%

ONE-ARM TRIALS

(Schultz et al., 1973; Fleming, 1982)

USUAL SETTING FOR ONE-ARM PHSE II TRIALS

- A (small) group of patients, all receive the same dose in an one-arm, open-label trial
- The investigator is first asked to specify the largest response rate, π_0, which if true the treatment does not warrant further investigation.
- Secondly, the investigator is asked to judge the smallest response rate, π_A, which if true would imply that the treatment has adequate therapeutic efficacy to warrant further investigation.

It seems it would be easier to understand if we set $\pi_0 = \pi_A$; however, the "gap" would also allow for consideration of other factors: cost, ease of applications, safety, etc… It is also similar to the case of monitoring for toxicity or adverse effects (where π_0 is the baseline rate but the study is only stopped if the toxicity rate exceeds π_A).

ONE-SAMPLE "TEST"

- After acquiring needed components, we state the (one-sided) hypotheses to be tested as: $H_0: \pi = \pi_0$ versus $H_A: \pi = \pi_A$.
- In addition, we should specify the size (α) and the power ($1-\beta$) from to determine the sample size n:

$$\alpha = \Pr(\text{reject } H_0 \mid H_0 \text{ is true})$$

$$\beta = \Pr(\text{reject } H_A \mid H_A \text{ is true})$$

SINGLE-STAGE DECISION

- The number of responses "x" is distributed as Binomial Bin(n,π); but if $n\pi \geq 10$, say, the distribution of Y is approximately normal.

$$Y(\pi) = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}$$

- Using the normal approximation, the Null Hypothesis H_0 is rejected when

$$Y(\pi_0) = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} > z_{1-\alpha}$$

$$x > x_r = n\pi_0 + z_{1-\alpha}\sqrt{n\pi_0(1-\pi_0)}$$

SINGLE-STAGE DECISION

$$Y(\pi_0) = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} > z_{1-\alpha}$$

$$x > x_r = n\pi_0 + z_{1-\alpha}\sqrt{n\pi_0(1-\pi_0)}$$

The rejection rule preserves the size α but we need some "continuity correction"

REJECTION DECISION

In order to main the power (may result from the normal approximation), the single-stage procedure should more appropriately reject H_0 whenever

$$x \geq x_r = [n\pi_0 + z_{1-\alpha}\sqrt{n\pi_0(1-\pi_0)}]^* + 1$$

Where x* denotes the nearest integer to x (round up and add "1"); Note: No need to specify the power)

EXAMPLES

Example #1: $\pi_0 = .20$, $\pi_A = .40$, and n = 25; Null Hypothesis is rejected if there are 9 responses or more.

$$x_r = [n\pi_0 + z_{1-\alpha}\sqrt{n\pi_0(1-\pi_0)}]^* + 1$$
$$= [(25)(.2) + (1.65)\sqrt{(25)(.2)(1-.2)}]^* + 1 = 9$$

Example #2: $\pi_0 = .25$, $\pi_A = .50$, and n = 25; Null Hypothesis is rejected if there are 11 responses or more.

Note: Use "1.65" because one-sided

In single-stage plan, the decision is made at the end; if the Null hypothesis is rejected, i.e. $x \geq x_r$, the agent/drug is recommended for further investigation.

As in any statistical test of significance, the Alternative Hypothesis and the statistical power are only needed for sample size determination; those values are not needed in making decision to reject or not to reject the Null Hypothesis.

MULTIPLE TESTING DESIGN

Suppose one decides to perform k tests (usually N = 2 or 3) and to allow n_i patients accrue between the (i-1) th and ith tests, so that $n = n_1 + n_2 + \ldots + n_k$.

Let $x_1, x_2, \ldots, x_k$ represent the number of responses among the $n_1, n_2, \ldots, n_k$ evaluable patients (so that $x = x_1 + x_2 + \ldots + x_k$). In addition, denote the set of (cumulative) rejection points (of H_0) by $\{x_{r1}, x_{r2}, \ldots, x_{rk}\}$.

CUMULATIVE SET OF REJECTION POINTS

Note that:

x_{ri} are determined after i tests, $(n_1 + n_2 + \ldots + n_i)$:

$$x_{ri} = [(\sum_{j=1}^{i} n_j)\pi_0 + z_{1-\alpha}\sqrt{(\sum_{j=1}^{i} n_j)\pi_0(1-\pi_0)}]^* + 1$$

RULE BY SCHULTZ ET AL.

After test #j, 1 ≤ n ≤ N; Stop and reject H_0 if:

$$\sum_{i=1}^{j} x_i \geq x_{rj}$$

A PROBLEM & SOLUTION

Fleming observed that the true significance level of a multiple testing procedure can be considerably higher than the nominal level

To preserve the nominal significance level in multiple testing, Fleming proposed to “inflate” the variance used in determining the rejection points

FLEMING'S RULE

$$x_{ri} = [(\sum_{j=1}^{i} n_j)\pi_0 + z_{1-\alpha}\sqrt{(\sum_{j=1}^{i} n_j)\pi_0(1-\pi_0)}]^* + 1$$

$$x_{ri} = [(\sum_{j=1}^{i} n_j)\pi_0 + z_{1-\alpha}\sqrt{n\pi_0(1-\pi_0)}]^* + 1$$

The method is presented in general but in practice, because of the small size of most Phase II Clinical Trials, there is often just one interim analysis.

EXAMPLE

- Let take: $\pi_0 = .20$, (& $\pi_A = .40$), and n = 30
- Let use k=2 (1 interim analysis)
- After the first 15 patients: $x_r = 7$
- After all 30 patients: $x_r = 10$

(It can be seen that there must be stronger evidence at interim analysis to stop the trial for reason of excellent efficacy)

Suggested Task:

Search and find a complete SAP sample.

Suggested Exercise:

Consider one-arm trial with 2 planned interim analyses with: $\pi_0 = .20$, $\pi_A = .40$, and n = 45

Find the threshold for trial termination after the first 15 patients