

Information-based monitoring of clinical trials

Anastasios A. Tsiatis^{*,†}

Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A.

SUMMARY

When designing a clinical trial to compare the effect of different treatments on response, a key issue facing the statistician is to determine how large a study is necessary to detect a clinically important difference with sufficient power. This is the case whether the study will be analysed only once (single-analysis) or whether it will be monitored periodically with the possibility of early stopping (group-sequential). Standard sample size calculations are based on both the magnitude of difference that is considered clinically important as well as values for the nuisance parameters in the statistical model. For planning purposes, best guesses are made for the value of the nuisance parameters and these are used to determine the sample size. However, if these guesses are incorrect this will affect the subsequent power to detect the clinically important difference. It is argued in this paper that statistical precision is directly related to *Statistical Information* and that the study should continue until the requisite statistical information is obtained. This is referred to as information-based design and analysis of clinical trials. We also argue that this type of methodology is best suited with group-sequential trials which monitor the data periodically and allow for estimation of the statistical information as the study progresses. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: group-sequential test; statistical information

1. INTRODUCTION

For ethical as well as practical reasons, data from a clinical trial are monitored periodically during the course of a clinical trial with the possibility of early stopping for

- serious toxicity,
- established benefit,
- no trend of interest,
- design problems, logistical issues too serious to fix.

*Correspondence to: Anastasios A. Tsiatis, Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A.

†E-mail: tsiatis@stat.ncsu.edu

Contract/grant sponsor: National Institute of Allergy and Infectious Disease
Contract/grant sponsor: National Cancer Institute

Although all the issues above are important, we will focus attention on the early stopping of a clinical trial in the case where we observe large treatment differences during the course of the study. The methods we discuss will be in the context of what is called a 'group-sequential study'. Basically, a group-sequential study is a study where data are collected over time for the purpose of testing a particular research hypothesis. During the course of the study the data are monitored periodically and, if sufficient evidence emerges in support of the hypothesis, the study may be terminated early.

The way this is generally done is, at each interim analysis, a test statistic is computed using all the accumulating data and the test statistic is compared to some prespecified stopping boundary. If the test statistic exceeds the boundary value, then this is taken as evidence that a significant treatment effect has been achieved and the study is stopped at that point; otherwise, the study continues accumulating data until the next monitoring time or until the prespecified end of study at which point a final analysis is conducted.

Some care and a great deal of thought is necessary in constructing the stopping boundaries in order to maintain the desired operating characteristics such as type I error and the power to detect clinically important treatment differences. Often, the group-sequential designs are based on certain assumptions about the data that are based on best guesses of what we expect to occur. It may be the case, however, that the initial guesses may not be accurate and the question is what should be done as the study progresses and the data itself suggest that some of the assumptions that the study was based on were incorrect. We will argue that this issue can be ameliorated by working with the idea of 'statistical information'.

In this paper, we will define what we mean by statistical information and how we can design and analyse group-sequential studies using statistical information and argue why this is preferable to the designs that are commonly used based on sample size. We will illustrate the methods using a simple example which compares the probability of a dichotomous response between two treatments.

2. NOTATION AND PRELIMINARIES

As statisticians, we view data from a clinical trial as realizations of random variables which has a distribution from a statistical model that is described through population parameters. It is often the case that one of the parameters in the statistical model, which we denote by δ , is the parameter of interest that describes the research hypothesis and the remaining parameters that are necessary to describe the model are referred to as nuisance parameters.

To be concrete, consider a simple example where we want to compare the response rate for a dichotomous endpoint (response/non-response) between two treatments say, treatment *A* versus treatment *B*, where treatment *B* may be the standard treatment currently used (control group) and treatment *A* is some new experimental treatment. A two-arm randomized study is conducted where the primary aim is testing whether the probability of responding to treatment *A* is the same as the probability of responding to treatment *B*. In such a trial, the data are monitored periodically, possibly by a data-monitoring committee and, at each monitoring time, a test statistic is computed comparing the observed response rate of treatment *A* to treatment *B* using all the accumulating data. If the difference in response rates is sufficiently large then the study may be terminated early.

For this problem, the statistical model can be defined by the population parameters π_A, π_B which denote the population probability of response on treatments *A* and *B*, respectively. The data that are collected in such a study, at some interim analysis, can be summarized as X_A, X_B , the number of individuals responding among the n_A, n_B patients randomized to treatments *A* and

B , respectively. Because of randomization, it may be reasonable to assume that the number of responses X_A and X_B follow independent binomial distributions; namely, $X_A \sim \text{bin}(n_A, \pi_A)$ and $X_B \sim \text{bin}(n_B, \pi_B)$.

The parameter of interest in such a study is denoted by the treatment difference $\delta = \pi_A - \pi_B$ and specifically, the primary focus of the clinical trial is to test the null hypothesis $H_0 : \delta = 0$. Although δ is the parameter of primary interest, it does not suffice in describing the probability distribution of the data. In addition, we need to define a nuisance parameter say, the parameter π_B (the population probability of response in the control group). As we will demonstrate, although the nuisance parameter is not of primary interest for inference and decision making, it plays an important role in the design of the clinical trial.

3. SINGLE ANALYSIS ONLY

Before discussing the issue of group-sequential designs, let us first review some of the issues in the design of a clinical trial when only a single (final) analysis is to be conducted. We also use this opportunity to introduce the notion of *Statistical Information*. As we already mentioned, the primary goal is to test the null hypothesis $H_0 : \delta = 0$. In designing the trial we also need to define the clinically important difference; δ_A ; i.e., the minimum treatment difference, if it exists, that is important to detect with high probability (power). For example, in our dichotomous response problem suppose after discussions with the clinical investigators it is determined that an increase in the response rate of 0.15 is deemed clinically important. This example will be carried out throughout the paper for illustrative purposes.

Decisions to reject or accept the null hypothesis depend on the magnitude of a test statistic T which is based on the estimated treatment difference, properly standardized. We denote the estimated treatment difference by $\hat{\delta}$ and the standard error by $\text{se}(\hat{\delta})$, in which case, the test statistic is given by

$$T = \frac{\hat{\delta}}{\text{se}(\hat{\delta})}$$

Specifically, for a two-sided test, we reject the null hypothesis if $|T|$ is sufficiently large.

For the example above, the decision to reject or accept the null hypothesis is based on the proportions test; namely

$$T = \frac{\hat{p}_A - \hat{p}_B}{[\bar{p}(1 - \bar{p})] \left\{ \frac{1}{n_A} + \frac{1}{n_B} \right\}^{1/2}}$$

where $\hat{p}_A = X_A/n_A$, $\hat{p}_B = X_B/n_B$ and $\bar{p} = (X_A + X_B)/(n_A + n_B)$.

In designing the study, a critical question is how large should the study be? The answer is 'when we have sufficient evidence to prove the scientific/clinical question being addressed'. For statistical models, the amount of evidence can be summarized by the amount of statistical information in the data regarding the parameter of interest δ . Roughly speaking, statistical information is related to the standard error of the estimator $\hat{\delta}$ of the parameter δ . That is, $I \approx [\text{se}(\hat{\delta})]^{-2}$, where I denotes information. The greater the information the smaller is the standard error hence the more precise the estimator.

Statistical information depends on both sample size and the value of the nuisance parameters. In our example, the information will depend on the sample size and the probability of response in the control group, π_B . In more complex problems, information can depend on factors such as accrual rate, event rate, length of accrual, length of follow-up, drop-outs, number of measurements per individual in a longitudinal study, etc. Much of this cannot be predicted precisely during the design stage.

Therefore, the appropriate question should be how much information is necessary to answer the research question? In hypothesis testing problems we are often interested in having a certain degree of statistical precision which is measured by the type I error and the power to detect a clinically important treatment difference δ_A . Specifically, in order to have sufficient evidence that will enable us to detect a clinically important treatment difference δ_A with probability (power) $1 - \beta$ using a test at the α (two-sided) level of significance, we need the standard error of the estimate to be

$$\text{se}(\hat{\delta}) = \frac{\delta_A}{z_{\alpha/2} + z_{\beta}} \quad (1)$$

or

$$\text{Information} = \frac{(z_{\alpha/2} + z_{\beta})^2}{\delta_A^2} \quad (2)$$

where $z_{\alpha/2}$ denotes the $(1 - \alpha/2)$ th quantile of a standard normal distribution. Thus, for example, if we wanted to detect a clinically important difference δ_A with 90 per cent power using a test at the 0.05 (two-sided) level of significance, we would need $\text{se}(\hat{\delta}) = \delta_A/(1.96 + 1.28)$ or $\text{Information} = 10.5/\delta_A^2$.

Therefore, the study should be analysed when the standard error is sufficiently small (or information is sufficiently large), as given by (1) and (2), in order to meet the objectives of the study. This strategy will provide correct level and power independent of the nuisance parameters (nuisance parameters are built into the standard error computations).

Of course, we cannot launch into a study without some idea of the physical requirements that may be necessary; i.e., sample size, study duration, etc. Therefore, before launching into a clinical trial, we must convert information into a physical study design. That is, using initial guesses for nuisance parameters, we conduct theoretical calculations or simulations to compute the number of subjects and other relevant design parameters (e.g. accrual rate, length of follow-up, etc.) needed to attain the targeted information. For our dichotomous response example, if we wanted to detect a clinically important difference $\delta_A = \pi_A - \pi_B$ with power equal to $1 - \beta$ using a two-sided test at the α level of significance in a two-arm randomized study with sample sizes $n_A = n_B = n$, then

$$I = \frac{n}{\pi_A(1 - \pi_A) + \pi_B(1 - \pi_B)} = \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta_A} \right]^2$$

$$n = \left[\frac{\{\pi_A(1 - \pi_A) + \pi_B(1 - \pi_B)\}(z_{\alpha/2} + z_{\beta})^2}{(\pi_A - \pi_B)^2} \right] \quad (3)$$

where π_B , the probability of response for the control group, is determined by the investigators based on some initial guess and $\pi_A = \pi_B + \delta_A$. For our example, suppose the best initial guess for the probability of response in the control group $\pi_B = 0.30$. If we want to detect a clinically

important increase of $\delta_A = 0.15$ (i.e., $\pi_A = 0.45$) with 90 per cent power using the two-sided proportions test at the 0.05 level of significance, then using (3), we deduce that the sample size needed is about 214 patients per treatment arm or a total of 428 patients.

If we base our sample size calculations on such an initial guess, then, of course, the power of the ensuing test will not be that which is desired if, in truth, our initial guess was incorrect.

4. INTERIM MONITORING

Suppose, instead of analysing the data only once at the end of the trial, we compute the test statistic at various interim times t_1, \dots, t_K with the possibility of stopping the study early if a large treatment difference emerges at one of the interim analyses. In such a case, a strategy for stopping the trial may be to reject the null hypothesis H_0 at the first interim time that the test statistic

$$|T(t_j)| \geq b_j, \quad j = 1, \dots, K$$

where $T(t_j)$ denotes the value of the test statistic using all the data that have accumulated up to and including the j th interim analysis and the values b_1, \dots, b_K are referred to as boundary values.

Some thought must be given to the choice of boundary values because if data are being monitored many times there is an increased chance of rejecting H_0 by chance alone. For example, if we took a naive point of view and rejected the null hypothesis at the first interim analysis that was significant at the nominal 0.05 level (i.e., if we rejected the null hypothesis the first time that the p -value was less than 0.05), then the overall type I error would be inflated. The increase in the type I error would depend on the number of interim looks as given in the following table:

| K | False Positive Rate |
|----------|---------------------|
| 1 | 0.050 |
| 2 | 0.083 |
| 3 | 0.107 |
| 4 | 0.126 |
| 5 | 0.142 |
| 10 | 0.193 |
| 100 | 0.374 |
| 1000 | 0.530 |
| ∞ | 1.000 |

For example, if we monitored the study five times and decided to reject the null hypothesis if the result at any of the five interim analyses was significant at the 0.05 level of significance, then the overall type I error would be 0.142 rather than the desired 0.05.

Consequently, stopping boundaries and/or alpha spending functions are used to adjust for the multiple comparisons. These are constructed in such a way that we need greater evidence to declare a true treatment difference at the different decision points and the resulting test has the desired overall significance level. Examples of such boundaries include the triangular boundaries, [1–4] boundaries. Also Lan–DeMets [5] proposed the use of spending functions that allow the construction of data-dependent stopping boundaries for more flexibility in the number and timing of the interim analyses to be conducted.

All of these methods assume a certain structure in the joint distribution of the sequentially computed test statistics. Namely, that the test statistics derived at interim times are approximately

normally distributed, and that the data between interim analyses used to construct test statistics are uncorrelated (Independent Increments). Although such results may not be expected to apply broadly, it was shown by Scharfstein *et al.* [6], that any *efficient* test (most tests used in practice), when computed sequentially over time, has asymptotically a normal independent increments process whose distribution depends only on *Statistical Information*. Consequently, the group-sequential tests and procedures described in the literature apply to almost all situations.

A key element in the design of a group-sequential test is the determination of the maximum sample size; i.e., the sample size that would be necessary at the K th and final analysis. This is important in determining the power of the test to detect a clinically important alternative. As for the single analysis, the power of the test is related directly to the statistical information which depends both on sample size as well as the value of the nuisance parameters. As we will see, it is useful to study the relationship of the maximum information necessary for a group-sequential test to detect a clinically important difference to the information necessary for a single analysis test to detect the same difference with the same power.

When data are monitored several times with the possibility of early stopping, there will be a loss of power compared to a study which uses a single analysis. Consequently, in order to compensate for this loss of power, the final analysis should be conducted when the Information (i.e., $[\text{se}(\hat{\delta})]^{-2}$) is equal to

$$\text{MI} = \left(\frac{10.5}{\delta_A^2} \right) \text{IF}$$

where IF denotes an inflation factor; i.e., some multiplicative constant greater than one.

The inflation factor IF depends on the shape of boundary values, the number of interim analyses to be conducted, as well as the type I error and power of the test. Otherwise, the inflation factor does not depend on the actual analysis to be conducted or the endpoint that is being analysed as long as the statistical structure is that given by Scharfstein *et al.* [6] (which applies very broadly). In the following table we give some examples of inflation factors for two commonly used group-sequential tests (i.e., Pocock and O'Brien–Fleming boundaries).

| K | Spending function | $\alpha = 0.05$ | | |
|-----|----------------------|-----------------|------|------|
| | | Power | | |
| | | 0.80 | 0.90 | 0.95 |
| 2 | Pocock | 1.11 | 1.10 | 1.09 |
| | O-F | 1.01 | 1.01 | 1.01 |
| 3 | Pocock | 1.17 | 1.15 | 1.14 |
| | O-F | 1.02 | 1.02 | 1.02 |
| 4 | Pocock | 1.20 | 1.18 | 1.17 |
| | O-F | 1.03 | 1.03 | 1.03 |
| 5 | Pocock | 1.23 | 1.21 | 1.19 |
| | O-F | 1.03 | 1.03 | 1.03 |

The above considerations suggest the following information-based strategy (algorithm) for designing and analysing group-sequential clinical trials that applies broadly to many classes of problems.

1. Monitor the data whenever necessary.
2. Compute $se(\hat{\delta})$ using all the accumulating data.
3. Translate this into proportion of total information

$$t = \frac{[se(\hat{\delta})]^{-2}}{MI}$$

4. Compute appropriate stopping boundary, say $b(t)$. For example, we might use the Lan–DeMets spending function which can be implemented by using the software package EAST [7].
5. Stop if boundary is crossed (i.e., if $|\hat{\delta}/se(\hat{\delta})| \geq b(t)$), or continue to next monitoring time.
6. Final analysis, if necessary, is made at time that standard error yields maximum information; i.e., $t = 1$.

We now illustrate how this procedure can be implemented in the example where we compare the probability of response between two treatments. The design specifications are as follows: The null hypothesis is given as $H_0 : \delta = \pi_A - \pi_B = 0$. The clinically important alternative hypothesis which we want to detect with power = 90 per cent at the 0.05 (two-sided) level of significance is given by $\delta_A = \pi_A - \pi_B = 0.15$.

Note: We have not specified nuisance parameters, nor do we have to when using information-based monitoring.

For planning purposes, however, we expect, as an initial guess, that the probability of response in the control group will be about $\pi_B = 0.30$. For a single analysis, we computed in the previous section a sample size of about 214 patients per treatment arm or a total of 428 patients. In addition, suppose we expect 20 patients per month to be accrued into this study and that an interim analysis will be conducted about every six months. The study will be monitored using an O'Brien–Fleming type boundary as adapted by using a Lan–DeMets spending function. Roughly, this will entail conducting up to four analyses. If all the design specifications were correct, then the maximum sample size would be $428 \times IF$, which, for this design, is approximately $428 \times 1.03 = 442$ patients, or 221 patients per arm (rounding up).

Although the calculations above are for planning purposes, in an information-based design we do not necessarily assume that the *a priori* guesses are correct. Instead we work with statistical information. We first compute the maximum information, which for this problem is

$$MI = \left\{ \frac{1.96 + 1.28}{0.15} \right\}^2 (1.03) = 480$$

As the data are monitored, we then compute the proportion of the maximum information and the corresponding boundary values as indicated by the information-based monitoring algorithm given earlier.

As an example, consider the following scenario where the data were generated using a simulation experiment to be described later. Here the data are analysed after every 120 patients (60 per treatment arm) and the results are summarized in the following tables. The boundary values were obtained using the software package EAST [7].

First analysis:

| | | | |
|---------|-----------|----|-----------------------------|
| | Treatment | | $\hat{\delta} = 0.167$ |
| | A | B | $se(\hat{\delta}) = 0.0781$ |
| respond | 14 | 15 | $I = 163.8$ |
| not | 46 | 45 | $t_1 = 163.8/480 = 0.34$ |
| | 60 | 60 | $b_1 = 3.48$ |
| | | | $T_1 = 0.21$ |

Second analysis:

| | | | |
|---------|-----------|-----|-----------------------------|
| | Treatment | | $\hat{\delta} = 0.10$ |
| | A | B | $se(\hat{\delta}) = 0.0583$ |
| respond | 29 | 41 | $I = 294$ |
| not | 91 | 79 | $t_2 = 294/480 = 0.61$ |
| | 120 | 120 | $b_2 = 2.62$ |
| | | | $T_2 = 1.72$ |

Third analysis:

| | | | |
|---------|-----------|-----|-----------------------------|
| | Treatment | | $\hat{\delta} = 0.11$ |
| | A | B | $se(\hat{\delta}) = 0.0471$ |
| respond | 41 | 61 | $I = 450$ |
| not | 139 | 119 | $t_3 = 450/480 = 0.94$ |
| | 180 | 180 | $b_3 = 2.06$ |
| | | | $T_3 = 2.34$ |

The boundary is crossed at the third analysis.

4.1. Some remarks

1. Based on the rate of information, the study would have terminated after 190 patients (per arm) entered.
2. This is in contrast to the 221 patients (per arm) that we initially guessed.
3. This is because the sample response rate of the control group is less than 30 per cent (estimated at design).
4. In actuality, the data were simulated using 20 and 35 per cent response rates for the two treatment arms rather than the 30 and 45 per cent response rates guessed initially.

5. CONCLUDING REMARKS

One of the disadvantages of information-based monitoring is that it would require studies to be of variable length. Although this is administratively inconvenient, we believe it is scientifically important. Although information-based methods can be used even when the plan is to conduct a single analysis, we believe it can be implemented naturally in conjunction with a study using a group-sequential design monitored by an independent data safety monitoring board (DSMB).

We suggest that one starts by giving a time frame for the study based on best guesses during the design stage for planning purposes. As the data emerge during the monitoring process, the monitoring committee should address the issue of whether the information necessary to achieve the scientific goals of the study is likely to be obtained in the original time frame. If not, recommendations should be made to extend, or change the study in some other fashion to meet the information goal.

For the reader interested in studying more details in conducting information-based design and monitoring of clinical trials, we recommend the papers by Scharfstein and Tsiatis [8] and Mehta and Tsiatis [9].

ACKNOWLEDGEMENTS

This research was supported by Grants from the National Institute of Allergy and Infectious Disease and the National Cancer Institute.

REFERENCES

1. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Ellis Horwood Limited: Chichester, 1983.
2. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.
3. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
4. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987; **43**:193–199.
5. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
6. Scharfstein DO, Tsiatis AA, Robins JM. Semiparametric efficiency and its implication on the design and analysis of group sequential studies. *Journal of the American Statistical Association* 1997; **92**:1342–1350.
7. EAST. *Software for Group Sequential Inference*. Cytel Software Corporation: Cambridge, MA, 2000.
8. Scharfstein DO, Tsiatis AA. The use of simulation and bootstrap in information-based group sequential studies. *Statistics in Medicine* 1998; **17**:75–87.
9. Mehta CR, Tsiatis AA. Flexible sample size considerations using information based interim monitoring. *Drug Information Journal* 2001; **35**:1095–1112.