

Sequential approach for interim analysis in controlled clinical trials

Approccio Sequenziale per le Analisi Ad Interim negli Studi Clinici Controllati

Maria Grazia Valsecchi

Dipartimento di Medicina Clinica, Prevenzione e Biotechnologie Sanitarie, Università di Milano-Bicocca, Via Cadore, 48, Monza (e-mail: grazia.valsecchi@unimib.it)

Riassunto: Le analisi ad interim sono una parte fondamentale del disegno e della conduzione di una sperimentazione clinica controllata. Si discutono qui i principi dell'approccio sequenziale, se ne mostra un'applicazione e se ne evidenziano vantaggi e limiti, indicando possibili metodi alternativi.

Keywords: Group-sequential design, Stopping rules, Interim analysis

1. Introduction

Interim analyses are analyses done while a controlled clinical trial is still open with the aim of deciding upon the trial continuation or modification based on accumulated data. The need of interim analyses is driven primarily by an ethical concern but scientific and economic issues too play a role. Typically, interim analyses evaluate whether a randomized clinical trial shows, in its course, an unsustainable risk on patients safety or problems with the treatment effect, either because of efficacy or for futility.

Sequential approaches allow to evaluate, based on pre-defined criteria, whether the study should stop prematurely because it shows a large treatment effect (efficacy). They differ on whether they consider a continuous monitoring or a group monitoring, but all define a fixed sample size calculated including an inflation factor that depends on the sequential plan of interim looks at the data. They are devised to control the overall level of type I error in the presence of multiple testing of correlated data.

Methods based on conditional power are usually the choice for interim analyses that face the problem of futility, that is the possibility of terminating the trial because early results suggest there is unlikely to be any treatment difference.

It is common to indicate as adaptive designs (or flexible designs) those which have instead the aim of planning, during the course of the study, analyses that assess whether the study should be modified (not just interrupted) in any of the following aspects: sample size, treatment doses or schedules, initial decisional criteria (power, difference of interest) or trial objectives (for instance, change a non inferiority randomized question into a superiority question).

This work describes how sequential approaches are used for clinical trial monitoring, with a real example on a recent cancer trial. It then points out some of the issues which recently stimulated the application in clinical research of alternative approaches to trial monitoring.

2. Group-Sequential Designs

Group-sequential methods are frequently used in clinical trials in order to monitor for unexpectedly large treatment effects in early results (stopping for efficacy). Formal procedures for interim analyses are required especially in those trials which are relatively large, have lengthy recruitment periods and involve patient survival. Indeed, it would clearly be unethical and unacceptable to continue recruiting patients to the trial if early results provide evidence of a convincing superiority of one of the randomized treatment policies. Interim analyses become a key issue in designing the trial and a group-sequential approach requires that:

- a maximum number of interim analyses is fixed in advance, with a schedule of when they would occur, related to accrual and follow-up;
- a primary test statistic is defined for treatment comparison.

The statistical approach to group-sequential boundaries is to define a critical value for significance at each interim analysis so that the overall type I error criterion will be satisfied. From a pragmatic point of view, the control of the overall type I error is seen as a vital aid to restricting the flood of false positives in the medical literature. In addition, this design is bounded to have a maximum sample size, thus the trial has a closure constraint, based on the specification of the null and of the minimum difference of interest (as a specific alternative), on the type I error and on the desired power. This approach of discrete monitoring is related to the continuous monitoring triangular designs (Whitehead, 1994).

Briefly, the definition of the boundaries for the group-sequential design is based on the joint distribution of the parameter estimates at the K different looks. Let θ be the parameter of interest (treatment effect) and $\hat{\theta}_k$ be the estimate based on data at analysis k , $H_0: \theta = 0$ and δ the (minimum) value of clinical interest in the treatment effect. The information for θ at analysis k is:

$$I_k = \{ \text{Var}(\hat{\theta}_k) \}^{-1} \quad k = 1, \dots, K$$

In many situations, including those related to the large sample theory of maximum likelihood estimation, the canonical joint distribution of $\hat{\theta}_1, \dots, \hat{\theta}_K$ is shown to be approximately multivariate normal

$$\hat{\theta}_k \sim N(\theta, \{ I_k \}^{-1}) \quad k = 1, \dots, K$$

with

$$\text{Cov}(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) = \text{Var}(\hat{\theta}_{k_2}) = \{ I_{k_2} \}^{-1} \quad \text{for } 1 \leq k_1 \leq k_2 \leq K$$

In testing H_0 , the standardized statistics $(Z_1, \dots, Z_K)$, with

$$Z_k = \frac{\hat{\theta}_k}{\sqrt{\text{Var}(\hat{\theta}_k)}} = \hat{\theta}_k \cdot \sqrt{I_k} \quad k = 1, \dots, K$$

are also approximately multivariate normal (Jennison & Turnbull, 2000).

In particular, for survival analysis, the log-rank statistic (or equivalently the score statistic from the Cox model) for testing H_0 , with θ being the natural logarithm of the hazard ratio between two treatment groups, is

$$Z_k \sim N(\theta \sqrt{I_k}, 1) \quad k = 1, \dots, K$$

and

$$\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{I_{k_1} / I_{k_2}} \quad \text{for } 1 \leq k_1 \leq k_2 \leq K$$

where the numerator of Z_k is usually expressed in the metric of observed minus expected number of failures (expected number calculated under H_0 with ranked failure time), namely:

$$Z_k = (O_k - E_k) / \sqrt{V_k} \text{ where } V_k = \text{Var}(O_k - E_k)$$

Suppose at stage k the number of observed deaths is d_k and the distinct failure times (assuming no ties) are $t_{1,k} < t_{2,k} < \dots < t_{d_k,k}$. If we indicate with $r_{iA,k}$ and $r_{iB,k}$ the number of subjects at risk on treatments A and B at $t_{i,k}$ and with r_{ik} the total number at risk, then:

$$E_k = \sum_{i=1}^{d_k} r_{iB,k} / r_{ik}$$

$$V_k = \sum_{i=1}^{d_k} r_{iA,k} r_{iB,k} / r_{ik}^2$$

The variance of $(O_k - E_k)$ is equal to I_k (assuming no ties). The information for θ when $\theta \approx 0$ is approximately equal to one-fourth of the total number of failures (for equal size treatment groups). In general, in survival, observed information is roughly proportional to the number of failures.

The sample size and the boundaries of early stopping depend mathematically on the power requirement $1-\beta$ against a specific δ and on a specified relationship between $(O_k - E_k)$ and its variance.

The commonly used test statistics (including the log-rank) have the “independent increments” property, as the score statistics $S_k = Z_k \sqrt{I_k}$ are also multivariate normal ($k = 1, \dots, K$) and

$$\text{Cov}(S_k - S_{k-1}, S_{k'} - S_{k'-1}) = 0 \quad \text{for } k \neq k';$$

this greatly simplifies the calculation of the interim significance levels.

Based on this theory the boundaries can be constructed of many different shapes, and usually require numerical solutions. The values of the boundaries $B(K, \alpha)$ which ensure

an overall type error probability α , for different values of K and for two different approaches (Pocock, 1977 and O'Brien & Fleming, 1979) are shown in table 1. Interim monitoring has a cost in terms of sample size, as the initial fixed sample needs to be inflated by a factor $R(K, \alpha, \beta)$, which is shown in table 1, and is larger as the number of looks increases.

Table 1: Boundaries for a group-sequential design with K interim analyses according to Pocock (B_P) and to O'Brien & Fleming (B_B) and corresponding inflation factor for sample size.

K	$B_P(K, \alpha)$	$R_P(K, \alpha, \beta)$	$B_B(K, \alpha)$	$R_B(K, \alpha, \beta)$
1	1.960	1.000	1.960	1.000
2	2.178	1.100	1.977	1.007
3	2.289	1.151	2.004	1.016
4	2.361	1.183	2.024	1.022
5	2.413	1.207	2.040	1.026

▪ $\alpha = 0.05$; $1 - \beta = 0.9$ ▪ Stopping rule at interim k: Stop and reject H_0 if: - $Z_k > B_P(K, \alpha)$ - $Z_k > B_B(K, \alpha) \cdot \sqrt{(K/k)}$ $k=1, \dots, K$ (K is the final analysis) ▪ Sample size in a group-sequential design with k looks: First calculate the fixed sample size without interim analysis, then multiply it by $R(K, \alpha, \beta)$
--

In the case of 3 looks ($K=3$) the p-values corresponding to the boundaries in a two-sided test for the two approaches are shown in table 2.

Table 2: p -values corresponding to two types of boundaries with $k=3$ analyses.

No of analyses	Pocock	O'Brien & Fleming
1	0.0221	0.0006
2	0.0221	0.0151
3	0.0221	0.0471

The Pocock boundaries have the undesirable feature that the final test is made with a larger critical value than that conventionally used for fixed sample size. For instance, it can be difficult for investigators to accept that a final result is not significant at the 0.03 level because a sequential procedure was used. The other approach corrects the problem because the boundaries yield nearly conventional levels of significance for the final analysis, at the cost of being less efficient in early analyses, as very extreme results are needed to stop early.

In terms of normal significant levels, in the O'Brien & Fleming, H_0 is rejected at analysis k ($k = 1, \dots, K$) if the two-sided significance level of H_0 is below

$$\alpha'_k = 2 \left\{ 1 - \Phi \left[B_B(k, \alpha) \sqrt{K/k} \right] \right\}$$

Group sequential stopping in its strict interpretation requires equally spaced information levels. However, in survival analysis, increments in number of failures between analyses equally spaced in time (for example yearly) are often unequal and difficult to predict, unless parametric assumptions on the shape of the hazard are made. This leads to consideration of an “error spending” design in which we are able to adjust the timing of interim analyses by “using” varying fractions of the overall type I error for the trial. As originally introduced by Lan & DeMets (1983) a monotone non-decreasing error spending function $f(\tau)$ is defined as a smooth function on the information fraction. At look k the information fraction is $\tau_k = I_k/I$ which, in survival, is approximately d_k/D (number of failures at look k over total number of failures). The boundaries values for stopping, $B_1, \dots, B_K$, are such that, based on the cumulative information I_k , it is:

$$\begin{aligned} \Pr_{\theta=0} \{ |Z_1| > B_1 \} &= f(\tau_1) \\ \Pr_{\theta=0} \{ |Z_1| < B_1, \dots, |Z_{k-1}| < B_{k-1}, |Z_k| > B_k \} &= f(\tau_k) - f(\tau_{k-1}) \quad k = 2, 3, \dots \end{aligned}$$

and, as desired, B_k does not depend on the as yet unobserved $I_{k+1}, I_{k+2}, \dots$. It can be shown that the function

$$f(\tau) = \min \left\{ 2 - 2\Phi \left(z_{\alpha/2} / \sqrt{\tau} \right), \alpha \right\}$$

This method has the potential to allow a great flexibility. It allows one not only to adapt to irregularities in calendar time of analyses, but also to vary the amount of error spent in response to the observed information. This, combined with a sampling rule which aims to continue to a pre-specified maximum information level, can satisfy a power condition quite accurately.

Finally, the theory of group-sequential design supports general comparisons, including adjustment for covariates, if needed.

If a trial is stopped early on the basis of a sequential plan, the estimate of treatment effect is likely to be inflated. The bias tends to be higher the earlier the trial has been stopped.

Adjusted estimates of treatment effect that allow for stopping early can be obtained, but there are no widely agreed approaches to it (Pocock, 1989 and 1993; Parmar, Spiegelhalter, Freedman, 1994).

3. Alternatives Approaches to Trial Monitoring

Problems of inference can arise in the group-sequential methods as the assessment of evidence about treatment depends on the plans for analyses in the future. This problem is relevant when considering, for example, a trial result with a p-value of 0.02. If this is

a final analysis of a fixed sample size study, we might conclude that the treatments have a significantly different effect on outcome. However, if we learn that this is the first of some planned interim analyses, the same evidence will be regarded as not being convincing. This aspect of the frequentist approach is the main ground for regarding the Bayesian approach to interim monitoring as more satisfactory. Here, interim data from the trial are used to estimate the treatment effect and Bayes rule is used to combine the prior distribution and the experimental evidence, in the form of a likelihood, into a posterior distribution. The posterior distribution is the basis for recommendations regarding the status of the trial. The dependence of the process on the choice of a prior distribution may be an obstacle to the introduction of Bayesian methods in practice. Although it is appealing that we are given the possibility to summarize prior knowledge (or ignorance) into a prior distribution, it must be said that a statistical distribution may be a poor representation of sparse evidence or beliefs and that the exact form given to it is somewhat arbitrary. The decision process considers that, if the posterior distribution is shifted far in one direction, the probability that the treatment effect lies in a region of clinical benefit will be high. In this case the inference will favour that particular treatment. On the other hand, neither treatment will be seen as convincingly superior if the posterior distribution is more centrally located. In this case a decision could be made of gathering additional evidence. However, in this situation, we lack a target sample size to aim at and this lack of an expected number (and recruitment time) for trial closure, may be a problem in the conduction of a trial.

The approach based on sequential probability (or likelihood) ratio tests (SPRT) can also be regarded as an interesting alternative to group-sequential designs. SPRT is usually considered as a frequentist approach, as the likelihood ratio was related to formal hypothesis testing in its first development (Wald, 1947). However, its appeal is that, through the likelihood function, it can provide a mean for summarizing evidence from the data. Briefly, suppose we are interested in knowing whether the failure rate estimated during the course of the study favours q_1 or q_2 , which are values of clinical importance chosen to facilitate the decision about early stopping. The relative evidence is summarized in the likelihood, given that a parametric form is chosen for the hazard. Using a hypothesis testing perspective, although we may avoid formal testing here, we are willing to decide between:

$$\begin{aligned} H_0: \{q \geq q_1\} \\ H_1: \{q \leq q_2\} \end{aligned}$$

We assess the evidence through the likelihood ratio, $\Lambda = L(q_1, \mathbf{X}) / L(q_2, \mathbf{X})$, where $\mathbf{X}$ are the observed data. If Λ is much greater than 1, the evidence is in favour of q_1 while, if Λ is much smaller than 1, the evidence is in favour of q_2 . Various types of boundaries can be defined, and a common choice is related to the control of type I and type II error level (although this is unnecessary) as follows:

$$\begin{aligned} \Lambda_1 &= (1 - \alpha) / \beta \\ \Lambda_2 &= \alpha / (1 - \beta) \end{aligned}$$

For simplicity, if d failures are observed and an exponential distribution is assumed, there follows:

$$\Lambda = q_1^d \exp(-q_1 \sum_{i=1}^N t_i) / q_2^d \exp(-q_2 \sum_{i=1}^N t_i) = (q_1/q_2)^d \exp(-[q_1 - q_2]T)$$

where T is the total current exposure time for study subjects. Taking logarithms and solving for d gives

$$d = \frac{\log \Lambda + T(q_1 - q_2)}{\log(q_1/q_2)}$$

which shows that the stopping boundaries for d can be expressed in terms of T . The approach can be extended to statistical models other than the exponential. This simple approach permits, in principle, a continuous monitoring of study outcomes. In practice however, outcome may not be known quickly, as in survival type data. However, since increments in follow-up time contribute to the failure rate estimate, study subjects “continuously” contribute to the statistics needed to monitor the study. Thus the basic methodology of this approach remains applicable. What in this approach can be seen as a strong limitation is that the trial may never favour either parameter, which means that there is no constraint for study closure.

4. Examples

We consider here a controlled clinical trial recently designed on treatment of acute lymphoblastic leukemia (ALL) in infants (children aged less than 1 year at diagnosis), recognized as a very rare variant of pediatric ALL, with a relatively dismal prognosis. This is an international trial (involving 18 collaborative groups), started in January 2006, which stratifies patients into risk groups according to presenting clinical and genetic features, and adopts risk-directed treatment policies. For the large majority of patients (85%), who present with medium-high risk features, a randomized question is asked, while for the remaining standard risk patients an observational study is performed. One major point in study planning was that standard treatment outcome could be estimated from a previous international study run by part of the groups involved here, and the estimated percent disease free survival at 3 years was 41% (SE 4.0). In this international setting, with a complex trial management, it was desirable to have a plan in advance for a target sample size and a calendar for interim analyses. The choice was to design the randomized study, according to a group-sequential approach. An independent Data and Safety Monitoring Committee (with two clinical members and one statistician, all of them not involved in the study) was set up to review the trial progress and the interim analyses.

The study plan includes $K=3$ analyses, according to an O’Brien and Fleming spending function, with the first interim analysis planned on follow-up data updated at the end of 2007, the second at the end of 2009 and the final at the end of 2012. The target sample size is 320 patients (163 events), recruited in 5 years, with overall type I error $\alpha=0.05$ (two-sided test) and power $1-\beta=0.82$, against a difference of 16%. Significance levels for the group-sequential plan are shown in table 3.

Table 3: Significance levels for interim analyses with an O'Brien and Fleming spending function at unequal calendar intervals, with overall type I error $\alpha=0.05$ (two-sided test).

	December 2007	December 2009	Final
p-level	0.00006	0.006	0.048
Years from 1st randomization	2	4	7

For standard risk patients, the observational study adopts a treatment which is similar to that applied in the previous study. Guidelines were developed to monitor the outcome in this group in order to ensure that the failure rate does not exceed that estimated in the previous study.

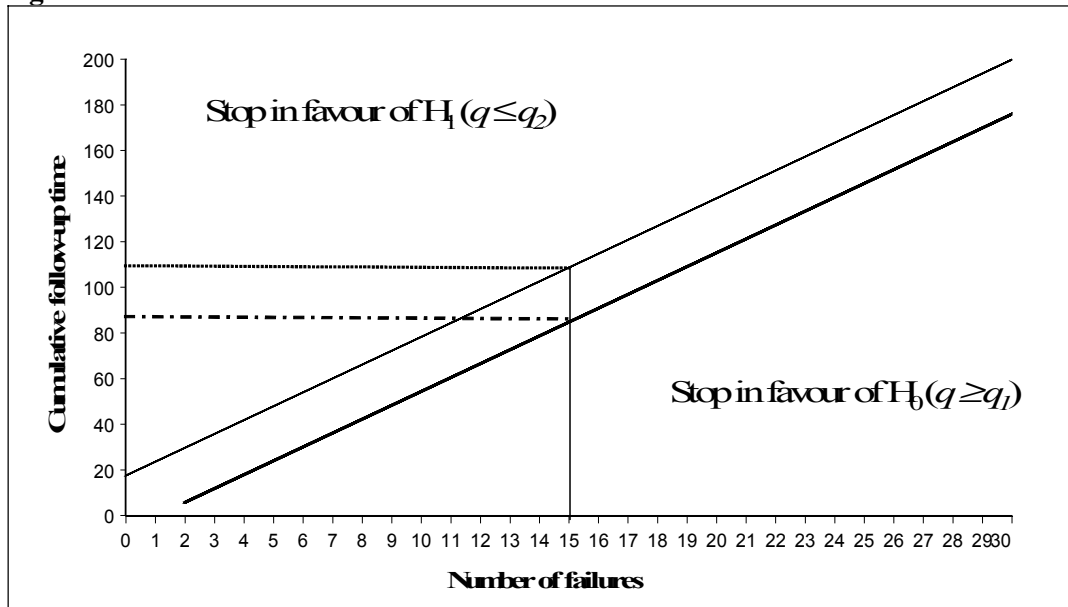
The sequential probability ratio test (SPRT) was applied here, considering the following hypotheses:

$$H_0: \{q \geq 0.32\}$$

$$H_1: \{q \leq 0.07\},$$

where 0.07 and 0.32 are the estimated failure rates in the standard risk and medium risk patients, respectively, when treated in the previous protocol. SPRT boundaries shown in figure 1 are based on the exponential model, with type I error $\alpha = 0.01$ and type II error $\beta = 0.20$, and the Log-likelihood ratio of H_0 versus H_1 .

Figure 1: SPRT boundaries



A nice feature is that the boundaries can be expressed in terms of failures and cumulative observation time (i.e. cumulative years of follow-up), as in table 4, which shows the values related to the first 15 failures. The focus of the application will be on the lower boundary: if, given the number of observed failures, the observed cumulative follow-up time (in years) is inferior to the lower boundary, then the outcome of the standard risk treatment is judged to be poor. In such a case, we conclude that evidence

from accumulating data suggests that the treatment adopted in the observational study needs re-consideration.

Table 4: *SPRT boundaries for the first 15 failures*

No. of failures	Cumulative years of follow-up	
	Lower boundary	Upper boundary
1	-	23.6
2	5.8	29.7
3	11.8	35.8
4	17.9	41.8
5	24.0	47.9
6	30.1	54.0
7	36.2	60.1
8	42.2	66.1
9	48.3	72.2
10	54.4	78.3
11	60.5	84.4
12	66.6	90.5
13	72.6	96.6
14	78.7	102.6
15	84.8	108.7

4. Discussion

Any method for trial monitoring suffers from an oversimplification of a complex process. While we can with no doubt say that the lack of a design that formalizes the decision to stop increases the chance for error, we must also recognize that any decision on a trial has to be made based on a broader judgment.

In this regard, Bayesian approaches may be more appealing than the frequentist view, as they offer a formal way to implement the expert opinion in the monitoring process. Parmar et al (2001) discuss a nice application in a trial, which however had a sample size planned according to a standard frequentist approach. This sample size aspect, which has a great importance in the practice of clinical research, is still handled more satisfactorily within a frequentist approach. A rare example of Bayesian application is in the field of clinical trials for cost effectiveness (O'Hagan et al, 2001) and the issue is also discussed in Lee & Zelen (2000). Certainly this aspect deserves further research, as the use of different priors, for instance to examine the sceptical and enthusiastic view, would lead to different sample sizes.

As for the group-sequential designs, recent research work is focusing on developing boundaries that allow more flexibility for unplanned design changes, as for example the need to adapt to a change in power during the course of the study (Jennison & Turnbull, 2006). This research has been stimulated by work on adaptive designs which however, as defined in the introduction, start from a different standpoint and have a different context of application (Shih, 2006).

While there is no doubt on the need for monitoring the accumulating evidence in a trial, we should not forget that clinical research has also the imperative to gather complete and convincing evidence about all the objectives of a trial. Whatever the method used, a

small number of interim looks (1 to 3) is usually sufficient, at least in studies on survival. Their number is limited by a joint effect of the recruitment period, treatment period and type of response variable. These issues should be kept in mind for a useful application of monitoring rules.

References

- Jennison C., Turnbull B. W. (2000) *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRC, Boca Raton.
- Jennison C., Turnbull B. W. (2006) Efficient group sequential designs when there are several effect sizes under consideration, *Statistics in Medicine*, 25, 917-932.
- Lan K. K. G., DeMets D. L. (1983) Discrete sequential boundaries for clinical trials, *Biometrika*, 70, 659-663.
- Lan K. K. G., Reboussin D. M., DeMets D. L. (1994) Information and information fractions for design and sequential monitoring of clinical trials, *Communications in Statistics - Theory and Methods*, 23, 403-420.
- Lee S. J., Zelen M. (2000) Clinical Trials and Sample Size Considerations: Another Perspective, *Statistical Science*, 15, 95-110.
- O'Brien P. C., Fleming T. R. (1979) A multiple testing procedure for clinical trials, *Biometrics*, 35, 549-556.
- O'Hagan A., Stevens J. W. (2001) Bayesian assessment of sample size for clinical trials of cost effectiveness, *Medical decision Making*, 21, 219-230.
- Parmar M. K. B., Griffiths G. O., Spiegelhalter D. J., Souhami R. L., Altman D. G. (2001) Monitoring of large randomised clinical trials: a new approach with Bayesian methods, *The Lancet*, 358, 375-381.
- Parmar M. K. B., Spiegelhalter D. J., Freedman L. S., et al (1994) The CHART trials: Bayesian design and monitoring in practice, *Statistics in Medicine*, 13, 1297-1312.
- Pocock S. J. (1977) Group sequential methods in the design and analysis of clinical trials, *Biometrika*, 64, 191-199.
- Pocock S. J. (1993) Statistical and ethical issues in monitoring clinical trials, *Statistics in Medicine*, 12, 1459-1469.
- Pocock S. J., Hughes M. D. (1989) Practical problems in interim analyses, with particular regard to estimation, *Controlled Clinical Trials*, 10, 209-221.
- Shih W. J. (2006) Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: A comparison, *Statistics in Medicine*, 25, 933-941.
- Wald A. (1947) *Sequential Analysis*, Wiley, New York.
- Whitehead J. (1994) Sequential methods based on the boundaries approach for the clinical comparison of survival times, *Statistics in Medicine*, 13, 1357-1368.