Statistical approaches to interim analysis : a critical appraisal

Thesis

UNRESTRICTED

STATISTICAL APPROACHES TO INTERIM ANALYSIS: A CRITICAL APPRAISAL

Thesis submitted for the degree of Doctor of Philosophy
at the Open University, UK

Discipline of Life Sciences

by

Irene Claudia Floriani, Degree in Biological Sciences

Istituto di Ricerche Farmacologiche "Mario Negri", Milan Italy

July, 2005

DATE OF SUBMISSION: 22 JULY 2005
DATE OF AWARD: 15 NOVEMBER 2005

Abstract

Several important questions have been raised about decision of stopping a trial early and on what basis to reach such a decision. It seemed therefore of interest to investigate the forms of monitoring used in cancer clinical trials and to gather information on the role of interim analyses in the data monitoring process of a clinical trial.

The project addressed the following issues:

- what is the performance of different interim analysis approaches;
- how often interim analyses are used in cancer clinical trials;
- which types of statistical analyses are more frequently adopted;
- how the data monitoring is organised and which is the weight of statistical analyses in the decisional process.

Analysis of performance of different statistical analysis approaches has been conducted by comparing the probability of stopping and the estimation bias on clinical scenarios based on real data of trials performed in ovarian and colorectal cancers. The project also focused on the prevalence of different types of interim analyses and data monitoring for both safety and efficacy in cancer clinical trials.

Sources of investigation were the literature data and the protocols of cancer clinical trials included in the in the Italian registry of clinical trials.

Results of our research indicate that the more widely used statistical approaches reduce the risk of "incorrect" early stopping, compared with the adoption of no stopping rule, with similar performance. Analysis of protocols and early reports suggests that the implementation of these procedures in a monitoring strategy is not satisfactory. Use of interim analyses is still limited to the frequentist approach of the alpha-spending function, while the Bayesian is not considered.

Interim analysis plans are still scarcely described, even in more recent protocols, denoting a not yet sufficient attention to this issue not only by the researchers, but also by regulatory bodies.

ai Gladiatori di tutti i tempi,,,,,

Acknowledgements

I am grateful to my supervisors, Carlo La Vecchia and Mahesh Parmar, for their invaluable advice, with a special thank to Valter Torri for his ever present support from before the beginning and throughout the thesis writing process.

I would like also to thank all the people, who helped me in these years: Antonella Bodini and Fabrizio Ruggeri for their stimulating discussion on Bayesian approach and critical review of the statistical aspects of the manuscript; Angelo Tinazzi and Enzo Bagnardi for their help in developing SAS macros; Andrea Tosato and Simone Mangano for the patient assistance with the use of LaTeX system; Filippo Debraud for facilitating the access to the database of Italian registry of clinical trials; Elena Albertazzi and Nicole Rotmensz for their kindness and precious help in the extraction of relevant data from protocols; Vanna Pistotti for providing supervision on literature search strategy.

A particular mention to Eugenio Muller, Paolo Mantegazza and Mario Nespoli, Presidents of Ethics Committees of Istituto Neurologico "Carlo Besta", Milan, San Paolo Hospital, Milan and Sant'Anna Hospital, Como, respectively, for their kind permission to examine clinical protocols submitted to these bodies for authorization.

I would like to thank the authors for their kindness in providing me PEST software.

Finally, I would like to express my gratitude to the "Mario Negri" Institute for funding my project.

Contents

List of Figures

List of Tables

Glossary of symbols and abbreviations

α	Working significance level
θ	Unknown parameter of interest representing the difference between treatments
θ_R	Reference improvement
Φ	Cumulative normal probability
τ	Information fraction
CI	Confidence interval
CP	Conditional power
df	Degree of Freedom
DSMC	Data and Safety Monitoring Committee
EC	Ethics Committee
H_0	Null hypothesis
H_1	Alternative hypothesis
HR	Hazard ratio
HR$_1$	Hazard ratio under the alternative hypothesis
$\sim N(\mu, \sigma^2)$	indicates "is distributed as a normal distribution with mean μ and standard deviation σ

OsSC	Osservatorio Nazionale sulla Sperimentazione Clinica dei Farmaci (National Monitoring Centre for Clinical Trials)
OBF	O'Brien and Fleming
OS	Overall Survival
p	Observed significance level
PFS	Progression free Survival
RCI	Repeated confidence interval
RCT	Randomized clinical trial
RFS	Recurrence Free Survival
SAS	Statistical Analysis System
SD	Standard Deviation
SPRT	Sequential Probability Ratio Test
V	Measure of the amount of information contained in Z about θ
Z	Cumulative measure of the observed evidence on the difference between treatments (unless otherwise specified)

Summary

Several important questions have been raised about decision of stopping a trial early and on what basis to reach such a decision. Differences between treatments may be larger than expected or unanticipated adverse effects may occur, and either of these may justify early termination of a trial. Early stopping is sometimes suggested because continuing a trial would not provide sufficiently useful information to warrant continuation. Existing evidence from outside the trial, such as meta-analyses of data from comparable trials, other existing evidence external to the trial itself and the nature of the condition and its alternative treatments may be also taken into consideration when deciding to stop or continue a trial.

Since data from the trial may later constitute the base of management of future patients, the evidence should also be sufficiently convincing to the wider clinical and patient communities to determine future practice.

In chronic life-threatening diseases, like cancer, evidence of early therapeutic benefits may be even more compelling than in other diseases and the detection of the best trade-off between obtaining results as earlier as possible and getting good estimation of the magnitude of treatment effect is particularly important.

A range of formal statistical approaches can be used as a basis for judging at what point such differences are so extreme as to be sufficiently unlikely to reflect the play of chance. These analyses help to control for errors in decision making and estimation; however, although interim analysis approaches in clinical trials are widely

known, information on explicit adoption of some form of planned monitoring, even in long-term trials, is still scarce and basically driven by the published reports of studies, which rarely include details on the strategies for data monitoring and interim analysis plan. More research is needed looking at actual protocols on ongoing studies.

It seemed therefore of interest to investigate the forms of monitoring used in cancer clinical trials and in particular to gather information on the role of interim analyses in the data monitoring process of a clinical trial.

More specifically, the project addressed the following issues:

- what is the performance of different interim analysis approaches;
- how often interim analyses are used in cancer clinical trials;
- which types of statistical analyses are more frequently adopted;
- how the data monitoring is organised and which is the weight of statistical analyses in the decisional process.

Analysis of performance of different statistical analysis approaches has been conducted by comparing the probability of stopping and the estimation bias on clinical scenarios based on real data of trials performed in ovarian and colorectal cancers.

The project also focused on the prevalence of different types of interim analyses and data monitoring for both safety and efficacy in cancer clinical trials. Source of investigation was the literature data and the protocols of cancer clinical trials included in the Italian registry of clinical trials.

The reason for using a protocol registry is that the quality of published reports is generally not sufficiently high and details of statistical analysis are seldom reported. Moreover, even when reported, statistical analysis of published trials belongs to protocols designed years before the study publication, and may not be appropriate for estimating the up to date prevalence of utilisation and kind of interim analyses. Italian registry of clinical trials gave an unique opportunity to get this information and allow a critical appraisal of the statistical designs utilized in current cancer clinical trials.

Similar data have been retrieved from the clinical studies published between 2000 and 2005 in order to compare older strategies to the more recent approaches. The impact of results of interim analyses on the decision of modifying the planned study conduction has been investigated in two separated ways: first, to provide information on how often an early termination of a trial is caused by results of an interim analysis. For this proposal, causes of early interruption have been explored searching published papers of preliminary or early clinical trial results. Second, to obtain information from the literature on the modalities of implementation of interim analyses in the data monitoring process, by investigating which are the more reported forms and rules for data monitoring.

The most important findings of our research can be summarised as follows:

- The more widely used statistical approaches reduce the risk of "incorrect" early stopping, compared with the adoption of no stopping rule. Performance of restricted procedures and alpha-spending function with O'Brien and Fleming boundaries are very similar, while triangular test obtains values that are similar to what achieved using Bayesian approach;

- If no approach is used, the probability of interrupting at early stages increases, with a higher probability of incurring an estimation bias. Since stabilization of the estimates appears to happen when a substantial amount of events has occurred, it seems appropriate to conduct interim analyses only after at least half of the expected events occurred, in order to reduce bias. With respect to this, alpha-spending function with O'Brien and Fleming approach and restricted procedure, which are more protective against early termination at the beginning of the study, favour a reduction of the magnitude of estimation bias;

- The number of analyses has a moderate impact on estimation, when some approach is adopted, but it can be important when no criteria for making allowance for multiple analyses are used.

Analysis of protocols and early reports suggests that, although the field of methodology of interim analyses of clinical trials is largely covered and various approaches are available, the implementation of these procedures in a monitoring strategy is still uncommon.

According to the sources of data investigated, analysis of statistical aspects of randomised clinical trial protocols in oncology, systematically collected in the National Monitoring Centre for Clinical Trials, reveals that the most recent trend, based on the analysis of the international and national trials with participation of Italian centres, is still not completely satisfactory.

The most important figures derived by this project indicate that only sixty-four percent of the protocols incorporate statistical interim analysis plans. Despite of the large availability of statistical methods for interim analysis, the almost only used approach is the frequentist method, with O'Brien and Fleming boundaries. A data and safety monitoring committee is present in 58% of protocols, but there is lack of information on their composition and on rules to be implemented.

When looking at the data derived from early reports, the adoption of a formal process of interim analysis affects only a minority (13%) of published trials and slightly more than half (55%) of early publications based on interim analysis. Again, the largely preferred approach is the frequentist method, generally with O'Brien and Fleming boundaries. Explicit use of a data and safety monitoring committee is reported only in 20% of reports of early publications explicitly based on interim analysis,

and the lack of information regarding its rules and composition is confirmed.

The most important 'take-home' message is that interim analyses play a fundamental role in the balance between the need of timely information regarding the treatment effect and the control of false positive errors and estimation biases. The most discussed and popular approaches appear to have good performance. However, the use of interim analyses is still limited basically to the frequentist approach of the alpha-spending function, while the Bayesian is not considered at all. Moreover, interim analysis plans are still scarcely described, even in more recent protocols, denoting a not yet sufficient attention to this issue not only by the researchers, but also by regulatory bodies.

Chapter 1

Introduction

1.1 Issues in monitoring clinical trials

The simplest approach for evaluating results of a clinical trial is to plan only one statistical analysis at the end of the study, using a fixed-sample size design: planning and conduction are easy, and correct methods for estimation are available.

This approach, natural when all observations are available in a short period of time, is less appropriate when data become available sequentially. This is the case of studies on chronic diseases, like cancer, in which recruitment may last many years, so that the first outcomes can be observed when the accrual is still ongoing: in such situations there might be ethical, practical and economical reasons for looking at the data before the planned end of the study.

Data monitoring conducted during a still ongoing study focuses on the following issues:

- Performance;

- Data integrity;

- Safety;

- Treatment effect.

The assessment of study performance in terms of quality of data, protocol adherence, recruitment rate is normally performed periodically by the study sponsor in an informal way, adopting modalities which can be grouped under the definition of "internal monitoring" (Armitage, 1991).

On the contrary, tasks of external monitoring are to evaluate data integrity, safety and efficacy of treatment and to provide advice on continuing the study as originally planned, on suggesting changes in the conduction, or on stopping the study.

The advice is mainly based on trial results, but should take into account the context of information currently available at the moment of the analysis.

This process, named "interim analysis" is usually conducted by a data and safety monitoring committee (DSMC), composed by a group of experts in the involved fields (biostatistician, clinical researcher, epidemiologist, clinician with expertise in the disease under investigation). The committee should be preferably independent, in the sense that people taking part on it have no involved interests in the study and do not directly participate in the trial.

Undoubtedly, ethical reasons play a major role on decision to stop a trial, since one should minimise the number of subjects treated with an unsafe, uneffective or clearly inferior treatment: in this sense interim analyses make the process more efficient. The increase in efficiency has repercussion also on economic side, since in case of early stopping, the study size and duration are obviously shortened. Another goal is to make available a beneficial treatment as soon as possible. A further positive role of interim analyses is that they may increase interest among study participants, sometimes revitalising the accrual and study participation.

However, there are also disadvantages in conducting an interim analysis, thus influencing study conduction: immature results may provide imprecise or even biased point and interval estimates of the treatment effect, increasing the error in inferential process (Hughes and Pocock, 1988). In fact, when a clinical trial is closed because a treatment difference has been detected, the estimate of the magnitude of that difference will overstate the "true" value (Armitage *et al.*, 1969). Finally, it is important to emphasize that trials stopped early are likely to be of small size, and as a consequence their results may lack of both statistical precision and credibility, since medical community might give them a sceptical welcome, even in case of highly significant results.

Therefore, while informal reviews are necessary, the process of repeatedly evaluating data must be done with caution, especially early in the course of a trial when the number of both participants and events related to safety and efficacy are relatively small.

All sequential designs have some common features. First, they relate patient accrual (or occurred events) to when analyses are performed. Then, they define a statistic to test the null hypothesis, as well as a statistical stopping rule, which specifies at each interim analysis the difference between groups that will result in stopping the study. Even so, it should be stressed that the decision to stop a trial before the pre-specified final analysis should not be guided only by statistical, but also by practical (toxicity, ease of administration, costs, etc.), as well as clinical considerations: for this reason it is prefereable to refer to them as guidelines, rather than rules.

There is no simple formula for how often data should be retrieved: monitoring activities should be commensurate with the nature, size, and complexity of the trial. Therefore whereas generally for phase I and early phase II trials a DSMC may not be appropriate due to the small size and short duration of these studies, in late phase II, phase III and phase IV trials more frequent and rigorous looks at the data become necessary.

A monitoring plan should also consider the severity of the disease, the nature of the intervention, and the characteristics of the target population.

All of these factors need to be taken into account in deciding on the frequency and intensity of the activities.

1.2 Summary of statistical methods for interim analyses

Many quantitative methods have been developed to monitor the proceeding of randomized clinical studies, even if since the beginning of its development the theory of the sequential analysis has been the waterfront of debates between Bayesian and frequentist statisticians (Barnard, 1949; Anscombe, 1963; Birnbaum, 1964; Armitage, 1963, 1967; Cornfield, 1966a, 1966b; Dupont, 1983; Brown, 1983; Canner, 1983), also because in this field the differences between frequentist and Bayesian approaches are particularly evident.

The main matter of controversy is if the knowledge of the previously carried out or planned for the future analyses should somehow influence the approach to the analysis of the data.

Frequentists, who follow the principle of the repeated sampling, support the necessity of adjustment in the analysis phase to make allowance for the multiple tests carried out.

On the other side, Bayesan statisticians, who are "supporters" of the principle of likelihood (Berger, 1985; Berry, 1987) consider that the inference should only be based on the function of likelihood and that the design has no role in the analysis. They underline that, unlike frequentist, Bayesian inference adhers to the likelihood principle, according to which all information contained in the data relevant to hypothesis testing, is captured by the ratio of the likelihoods under those hypotheses, and that inference not based on this ratio is not as relevant.

Surely the process of trial monitoring involves rules often complex and subjective, based on both statistical and non statistical aspects. Furthemore, none of the proposed monitoring methods seems optimal in every different circumstances and since they have been developed to solve specific issues, some aspects of the problem seem better solved by some approaches and some other aspects by other approaches.

With no doubt in the practice frequentist approaches are the most widely used, even if in theory there are no reasons to prefer them to Bayesian.

Statistical monitoring methods can be classified according to two factors (Freedman *et al.*, 1994):

- whether the method is frequentist or Bayesian;
- whether the method uses the current evidence, available at the moment of the analysis or data predicted by supposing for the future observations a certain trend until the achievement of the planned sample size.

Current frequentist approaches include fully sequential and group sequential methods, alpha-spending functions, and repeated confidence intervals, whereas stochastic curtailment is a frequentist method that uses predictions. Bayesian counterparts are based on either the posterior or the predictive distributions.

Among current and predictive methods, Armitage (1989), as well as Freedman *et al.* (1994), suggested to use "current" ones, since they are based on real data and not on the projection of what is anticipated but has not yet occurred. Regardless the specific method used, a key issue to keep in mind is that statistical rules are only a side of the question, all the more so because they tend to oversimplify the information relevant to the decision to take and the way it is taken through.

Comprehensive reviews of statistical aspects of monitoring can be found in Whitehead (1992), Jennison and Turnbull (1990) and Piantadosi (1997).

1.2.1 General statistical aspects

We will consider the case of trials aimed at comparing the efficacy, in terms of prolonged survival, of an experimental treatment with a control (represented by the reference standard or no treatment), testing the null hypothesis of no treatment difference, against an alternative hypothesis that the effect of the treatments differs, namely that one is greater than the other.

Let θ be the natural logarithm of the unknown parameter of interest, the hazard ratio (HR), measuring the relative difference in efficacy between treatments:

$$\theta = \ln\left(\frac{h_e(t)}{h_c(t)}\right), \tag{1.1}$$

assumed to be constant over all t, where h_e and h_c are the hazard functions in the experimental and control group, respectively. It represents the limiting probability that the event (assumed to be undesirable) occurs at time t, conditional on it not occurring before t.

The *proportional hazards model* under which $h_e(t) = e^{\theta}h_c(t)$ for all t is being assumed.

Values of θ lower than zero represent an advantage for the experimental treatment, values greater than zero an advantage for the control treatment, whereas the value zero evidence in neither directions. θ_R denotes the difference considered of clinical interest (reference improvement) between the two treatments.

An equivalent form to (1.1) is:

$$\theta = \ln\left(\frac{\ln S_e(t)}{\ln S_c(t)}\right), \tag{1.2}$$

where S_e and S_c are the survivor functions in the experimental and control group, respectively, representing the probability that the event occurs after time t.

The hypotheses to be tested are:

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta \neq 0 \tag{1.3}$$

Two sample statistics can be used in the investigation of θ (Piantadosi, 1997): the first, denoted by Z^*, is a cumulative measure of the observed difference between treatments. For survival data it is called the *logrank*, equal to the observed number of events in the experimental group minus the expected number under the null hypothesis. The second, denoted by V, measures the amount of information contained in Z^* about θ. It is approximately equal to one quarter of the total number of observed events, thus increasing as the trial progresses. When θ is small and study sample of moderate or large size, then Z^* is approximately normally distributed with mean θV and variance V. Z^* can be also used as test statistic and the test based on it is called the logrank test: Z^{*2}/V is calculated and then referred to a χ^2 distribution on one degree of freedom (df). Z^* is negative if the experimental treatment is better than control, otherwise is positive.

In order to facilitate the presentation of results, from now then $Z = -Z^*$ will be used. Therefore, Z is positive if the experimental treatment is better than control, otherwise Z is negative.

- If $Z > k$, with $k \equiv (\alpha, \beta)$, then the null hypothesis will be rejected at the level of significance α and it will be concluded that the experimental treatment is superior to the control;

- If $Z \leq -k$, then the null hypothesis will be rejected concluding that the experimental treatment is inferior.

The requirements for the test are thus:

$$P(Z \geq k; 0) = \frac{\alpha}{2} \qquad \text{and} \qquad P(Z \geq k; \theta_R) = 1 - \beta \tag{1.4}$$

where the term appearing within the brackets after the semicolon represents the true value of θ.

When $\theta = 0$, Z is normally distributed with mean 0 and variance V. When $\theta = \theta_R$, Z is normally distributed with mean $\theta_R V$ and variance V.

The requirements (1.4) are satisfied by a suitable choice of V and k, namely:

$$V = \left[\frac{\eta_{\alpha/2} + \eta_\beta}{\theta_R}\right]^2 \qquad \text{and} \qquad k = \frac{(\eta_{\alpha/2} + \eta_\beta)\eta_{\alpha/2}}{\theta_R} \tag{1.5}$$

where η_γ denotes the upper 100(1-γ) percentage point of the normal distribution for any value of γ between 0 and 1.

The information required increases as the reference improvement decreases, as the working significance level decreases, and the power increases.

At each analysis, the actual value of V observed should be used. As Z is approximately normally distributed with mean θV and variance V, $\widehat{\theta} = Z/V$ provides a simple estimate of θ. When sample are large, $\widehat{\theta}$ is approximately equal to the maximum likelihood estimate of θ and (Z/V-1.96/$\sqrt{V}$, Z/V+1.96/$\sqrt{V}$) provides an approximate 95% confidence interval (CI) for θ.

1.2.2 Frequentist approach

Frequentist sequential procedures are of two types: those derived from the boundary approach and those derived from the repeated significance test approach.

In the first approach the two statistics Z and V, described in section 1.2.1, are plotted one against the other and at each analysis the identified point is compared to a prefixed stopping boundary; in the other a certain number of interim analyses are performed with significance levels adjusted to make allowance for repetition.

Earlier designs implied frequent looks of the data, even at each new available observation, so that the monitoring could be considered as continuous. This is rarely feasible due to practical problems, thus later designs, the so called "group sequential trials", tried to solve this gap by involving a fewer number of analyses and for their easier conduction they are of common use.

1.2.2.1 Boundary approach

According to this approach, at the i-th inspection, Z_i and V_i are calculated, as well as an upper u_i and a lower l_i stopping limits, $(l_i < u_i)$.

- If $Z_i \geq u_i$ then the trial will be interrupted, with the rejection of the null hypothesis and the conclusion that experimental treatment is better than control;

- If $Z_i \leq l_i$ again the trial will be stopped with either the rejection of the null hypothesis and the conclusion that the experimental treatment is worse than the control, or the acceptance of the null hypothesis and concluding that there is no evidence of any difference;

- If $l_i < Z_i < u_i$ then the study will continue until the next look.

If a boundary has not been crossed, the process will continue endlessly (open design), or up to a pre-specified maximum number of enrolled patients (or events occurred). It is not necessary to define in advance the timing of the analyses, as well as the amount of new information between analyses is not required to be of constant size. Accordingly, the values of V_i, u_i and l_i have not to be fixed in advance, too. On the contrary, the identification of a prespecified rule for the calculation is very important. As a general principle, u_i and l_i are determined as a function of $V_1, \dots V_i$, but they must have no relationship with $Z_1, \dots Z_i$.

Due to the previously mentioned difficulties in following continuous monitoring, a certain form of adjustment should be introduced to account for the discrete way of monitoring, which makes more difficult to detect the crossing of the boundaries. This can be done by bringing boundaries nearer, yielding a "Christmas tree" shape.

Sequential designs were introduced for the first time in the context of industrial quality control at the end of twenties of the last century (Dodge and Romig, 1929). In such experiments, the components are classified as effective or defective and a batch is accepted only if its proportion of defectives is acceptably low. These procedures were drew on and developed during II World War in connection with military quality control in the same time and in a similar way by Wald in the United States and by Barnard in England. Considered military secrets, these methods were widespread only after the end of the war, disguised as a book (Wald, 1947) and as an address to the Royal Statistical Society (Barnard, 1946).

Particularly the work of Wald was of considerable speculative impact on clinical research, even if of limited practical application, since addressed toward sampling inspection more than comparative trials. We owe to the same author the development of sequential stopping rules and of the Sequential Probability Ratio Test (SPRT) for testing between to simple hypotheses:

$$H_0 : \theta = \theta_1 \qquad \text{versus} \qquad H_1 : \theta = \theta_2 \tag{1.6}$$

Suppose that based on the test result, one out of two decisions, D_0 or D_1 is taken, the first favouring H_0, the second favouring H_1, with α and β the probabilities of the errors made choosing D_1 when H_0 is true or D_0 when H_1 is true.

According to Wald's method, at the occurrence of each event the ratio L_1/L_0 of the likelihoods of H_1 and H_0 is calculated. The study is closed with the decision D_1 if at any stage:

$$\frac{L_1}{L_0} \geq \frac{1-\beta}{\alpha} \tag{1.7}$$

or with the decision D_0 if

$$\frac{L_1}{L_0} \leq \frac{\beta}{1-\alpha}, \tag{1.8}$$

otherwise the study is continued.

These values, expressed through the likelihood ratio, are constant during all study period and correspond to those that should have been obtained using a fixed-sample design. This approach leads to open plans and in the simpler cases to graphical methods with linear boundaries. The actual error probabilities are only approximately equal to the specified values, since it is easier that the likelihood ratio value crosses one boundary rather than taking exactly the same value. Solutions for calculating exact probabilities have been supplied by many authors (Barraclough and Page, 1959; Manley, 1970; Dhuang-Zhen, 1990). Various forms of open designs have been described by Armitage (1954, 1975).

In Figure 1.1 the open SPRT in the case of continuous monitoring is showed. The plot of Z against V, called the "sample path" is updated at each inspection of the data. If it crosses the green line, the null hypothesis is rejected concluding that the experimental treatment is superior, whereas if the red line is crossed, again H_0 is rejected and the conclusion is that control treatment is superior. Blue line crossing causes on the contrary the acceptance of the null hypothesis. These colours will be used with the same meaning through all this chapter.

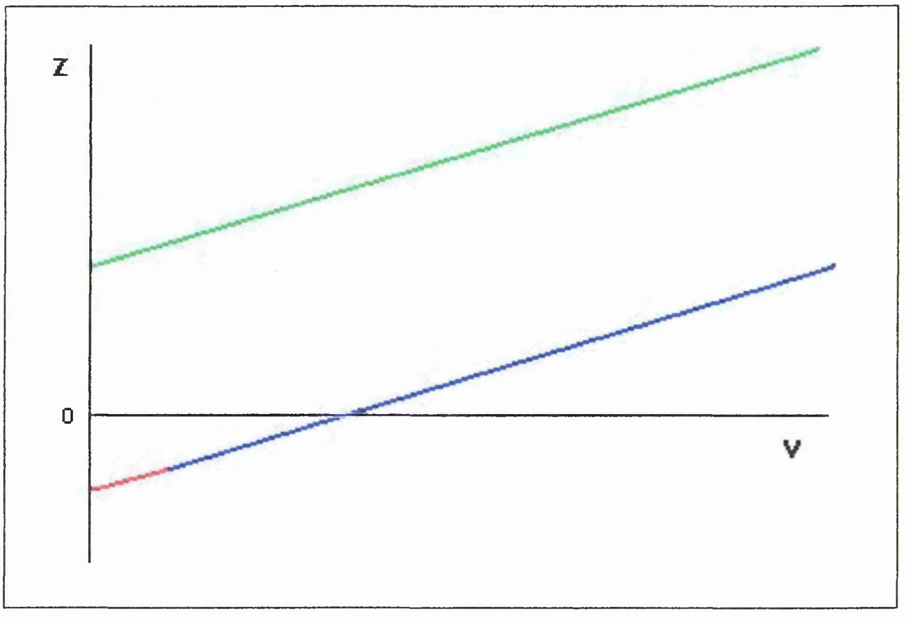

Figure 1.1: **Open sequential probability ratio test**

Theoretically, several reasons make the open SPRT designs appealing for constructing stopping rules for study monitoring. First they are based on the likelihood function, an efficient summary measure, second they are easy to use and interpret and third allow to continuously monitor accumulating data (Piantadosi, 1997).

In practice, such intensive surveillance and prompt action are rarely feasible: it is often more reliable to analyze data at periodic intervals, say every few months.

Furthermore, the plan is open, i.e. there is the chance that a boundary will never be crossed and the study will not have an end.

A possibility of solving this latter problem is to define a maximum number of patients (or events): once reached, the study is interrupted either with evidence in neither direction or with the conservative approach of accepting the null hypothesis. Otherwise, it is also possible to adopt a close design, truncated at certain point, as shown in Figure 1.2 in which we are sure to cross a stopping boundary.

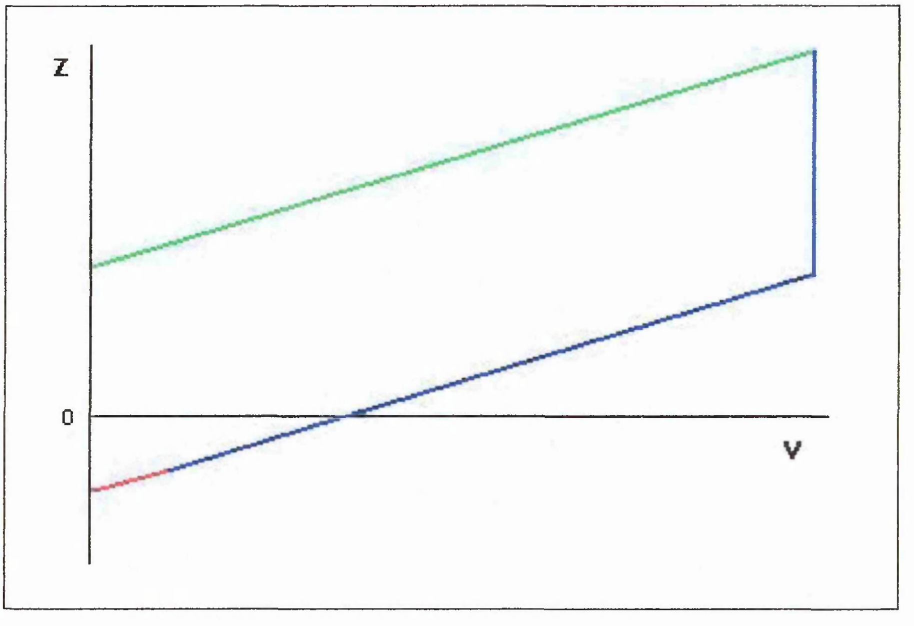

Figure 1.2: **Restricted sequential probability ratio test**

The application of sequential methods to clinical studies was first attempted by Kilpatrick and Oldham (1954), Bross (1952, 1958) and Armitage (1957).

Bross was escribed to suggest the use of two plans, deliberately closed and therefore distinct from arbitrarily truncated open plans.

Restricted procedures, shown in Figure 1.3 and introducted by Armitage (1957), are similar to Bross' design, even if they provide a much wider choice. Such a design offers scarce chances to early interrupt the study and in this sense it does not always satisfy the ethical requirements of early detecting an important difference in efficacy between treatments.

Other sequential designs are the triangular, the reverse triangular and double triangular tests (Whitehead, 1983, 1992; Whitehead and Stratton, 1983), that as well as the restricted procedure are special cases of the modified sequential probability ratio test introduced by Anderson (1960). The way to reach a conclusion is the same of that presented for the open SPRT.

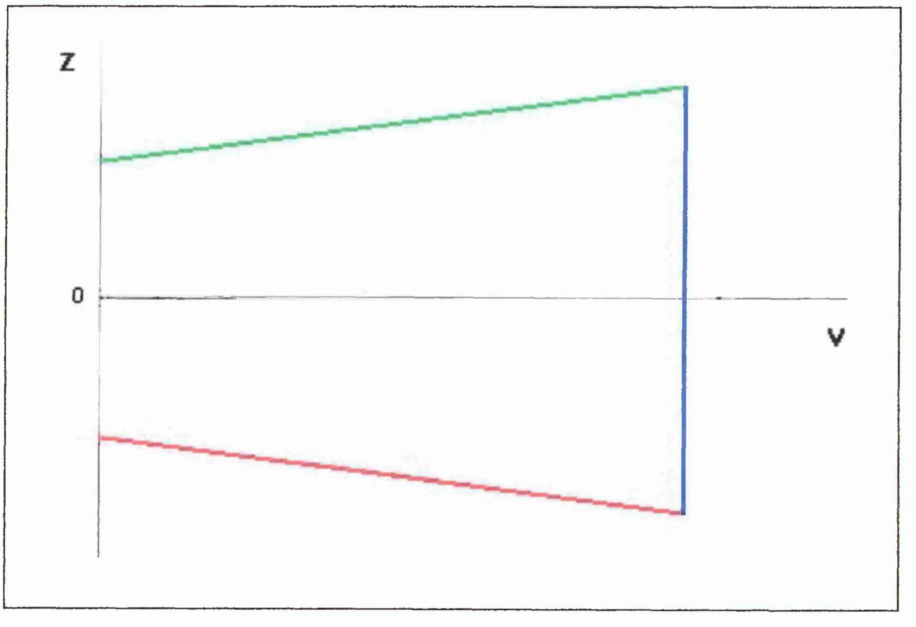

Figure 1.3: **Restricted procedure**

The triangular test, shown in Figure 1.4, has convergent boundaries, giving an asymmetrical triangular continuation region, which is finite. The study is continued until the sample path stays within the two boundaries and is stopped when one of the boundaries is crossed. The conclusion of the study depends on which boundary is crossed: experimental better than control for upper boundary, experimental non different or inferior to control for the lower boundary.

The reverse triangular test, shown in Figure 1.5 is particularly suitable for non-inferiority trials, in which even if the expected efficacy of the experimental treatment is equivalent or a little lower than the control one, it holds some known advantages in other endpoints (tolerability, costs, feasibility). In such a situation, we are not interested in distinguishing between superiority and equivalence, since in either case there is a benefit in using the new treatment: therefore this design has a high power to detect inferiority of the experimental treatment and a lower power to detect its superiority.

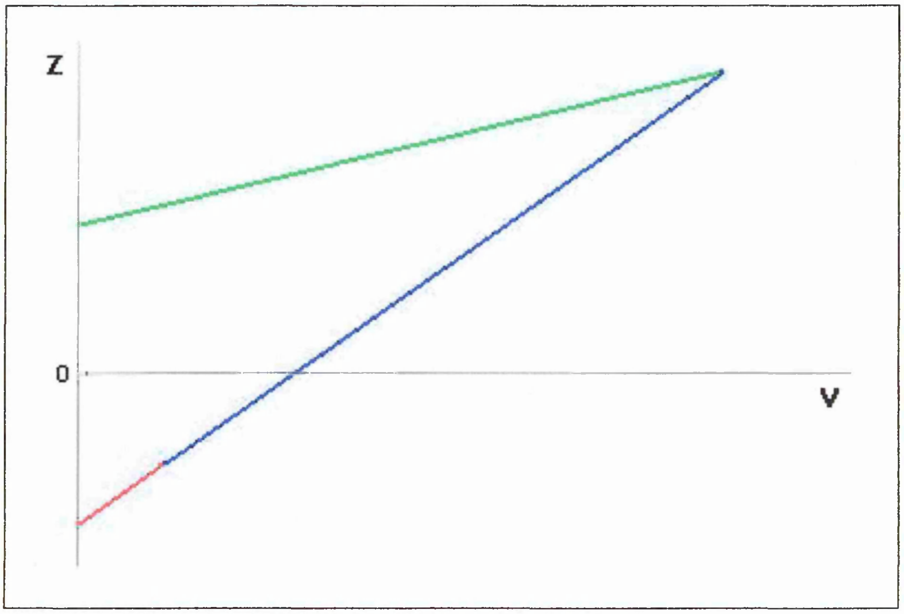

Figure 1.4: **Triangular test**

A double triangular test combines a triangular test and a reverse triangular test. In most cases this symmetric design, that guarantees a high probability of detec-

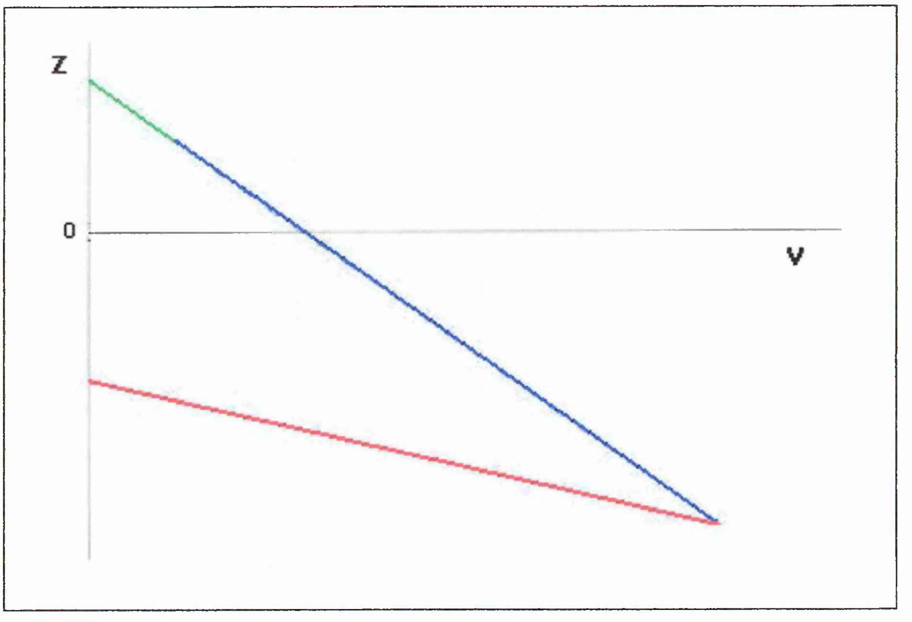

Figure 1.5: **Reverse triangular test**

ting either superiority or inferiority of the experimental treatment, leads to the same conclusions of the triangular test. The only difference is that when the sample path crosses either the lower or the upper broken line, the triangular test would stop immediately the study, while the double triangular test would continue in order to decide whether experimental treatment is not different or worse/better than the control treatment. The design of the double triangular test is reported in Figure 1.6.

1.2.2.2 Repeated significance testing procedures

From a clinical point of view, the optimal design is one that allows to point out a difference between treatments as soon as it reveales itself with an acceptable grade of certainty: this can be achieved with frequent analyses of data as they become available.

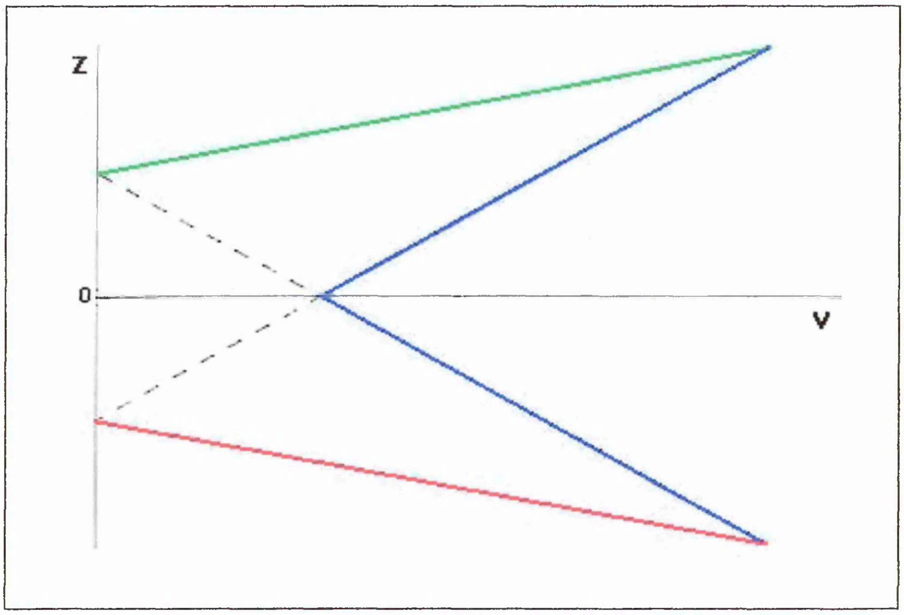

Figure 1.6: **Double triangular test**

The gap is that repeated use of significance tests on accumulating data increases the overall significance level, that is the probability of having at least one statistically significant difference, even when the hypothesis of no treatment difference is true. The greater the number of looks, the greater the possibility of observing a false positive result.

This phenomenon is called "optional stopping", because it was showed by Feller (1940) as a possible explanation of particularly favourable results in experiments on extra-sensory perception.

Previously, Kintchine (1924) underlined that by repeated analyses of accumulating data one can be absolutely sure to obtain an extremely low level of nominal significance, i.e. $p < 0.000001$, even under the null hypothesis.

The same concept, couched also by Cornfield (1966b), who the term "sampling versus a foregone conclusion" is due to, is extremely important in the decision process, since if we are certain to achieve any level of nominal significance even when the null hypothesis is true, we cannot rely on even highly significant results obtained. Robbins (1952) and Anscombe (1954) provided further discussion on this point.

Armitage *et al.* (1969) were the first to quantify the extent to which the type I error probability is increased over its nominal level, if a standard hypothesis test is conducted in a series of interim analyses. They studied the case of testing a normal mean with known variance and set the significance level at 5%. As an example, if a total of two analyses (one interim and one final) are performed the error is inflated to 8%. If a total of five analyses (four interim and one final) are performed this figure is 14%.

Many authors investigated the problem of "optional stopping" for different distributional forms of the endpoint variable (Armitage *et al.*, 1969; Armitage, 1971; McPherson, 1974) and in particular for survival data (Canner, 1977). These calculations have been also done when the alternative hypothesis is true, therefore focusing on the power of clinical trials providing for interim analyses of the data (McPherson and Armitage, 1971). It is interesting to note that apart from the distributional form of the response variable, if the analyses are performed at equally spaced intervals the frequency properties of repeated significance tests are extremely similar.

A possibility of controlling the probability of type I error is to adopt at each analysis a more stringent nominal significant level, thus keeping the overall alpha error at some suitable level (usually 5%). Since more and more conservative nominal significant levels must be adopted as the number of looks increases, this implies that planning in advance the maximum number, N, and the frequency of interim analyses becomes not only worthwile, but also necessary in order to apply a valid repeated significance testing procedure.

At the i^{th} inspection, Z_i and V_i are calculated, and the observed fixed-sample significant level compared with η_i, called the nominal significance level. Using the approximate normal distribution of Z_i the possibilities are:

- $l_i < Z_i < u_i$ for $i = 1 \ldots N-1 \Longrightarrow$ the study continues;
- $l_i < Z_i < u_i$ for $i = N \Longrightarrow$ the study closed, with evidence in neither direction;
- $Z_i \geq u_i$ for $i = 1 \ldots N \Longrightarrow$ the study is interrupted, concluding that the experimental treatment is superior to control;
- $Z_i \leq l_i$ for $i = 1 \ldots N \Longrightarrow$ the study is interrupted, concluding that the experimental treatment is inferior to control

where

$$l_i = -k_i\sqrt{V_i}, \qquad u_i = k_i\sqrt{V_i} \tag{1.9}$$

and k_i is the $100(1-\frac{1}{2}\eta_i)$ percentage point of the standard normal distribution. As already underlined, the values of $\eta_1 \ldots, \eta_N$ are chosen to maintain the overall significance level equal to α.

Repeated significance testing were introduced by Armitage (1958) and extensively explained by the same author (1975).

Earlier designs imply equally spaced inspections in terms of information available, $V_i = iV_1$, $i = 1, \ldots, N$ and therefore equal nominal significance levels $\eta_i = \eta'$, $i = 1, \ldots, N$ and $k_i = k$, $i = 1, \ldots, N$. The number of inspections can be high, even after every individual patient or pair of patients.

Group sequential approaches were proposed for the first time by Cutler *et al.* (1966), followed by McPherson (1974, 1977) and Pocock (1977, 1982) as a reasonable compromise between fully sequential and fixed-sample designs, in which plans are made for a small number of interim analyses, in contrast to fully sequential methods in which analyses are performed after the recording of even each outcome.

This alternative method is motivated partly because a little additional increase in efficiency is added by undertaking more than five analyses during the course of a clinical study, unless an extremely large treatment difference may be anticipated (Pocock, 1982) and because data management constraints usually do not allow for the continual availability of good quality data.

Many sequential stopping boundaries have been proposed to guide early stopping of clinical studies.

Pocock (1977) used a constant nominal level for all analyses.

O'Brien and Fleming (OBF)(1979) proposed for equally spaced analyses a significance level so that $k_i = k_i/\sqrt{i}$, $i = 1, \ldots, N$. Since for equal spacing, $V_i = iV_1$, from equation (1.9) the boundaries are $l_i = -k_1/V_1$ and $u_i = k_1/V_1$. The graphic of Z_i against V_i is compared with a constant horizontal line, as for the restricted procedure with horizontal boundaries.
With $\alpha = 0.05$ and a total number of five analyses, Pocock procedure uses significant levels $\eta_1, \ldots, \eta_5 = 0.0158$, while that of O'Brien and Fleming $\eta_1 = 0.00001, \eta_2 = 0.0013, \eta_3 = 0.0084, \eta_4 = 0.0225, \eta_5 = 0.041$.

Pocock approach was criticized, since offers high probability of early stopping, causing lack of accuracy in estimation of treatment effect. Another problem is that it undertakes the last analysis at a p-value considerably smaller than the conventional value of 0.05 (Geller and Pocock, 1987).

OBF method was the first attempt to overcome the above mentioned problems, with tests of gradually decreasing stringency, even if it offers little chance of early stopping and it is perhaps too conservative at the first analyses: for this reason the same authors suggested a little change of their design (Fleming *et al.*, 1984). This approach in case of time to event data preserves the sensitivity to late occurrence of survival difference.

Pocock (1982) suggested intermediate schemes between these two extremes, that minimize the sample size in order to detect the alternative hypothesis with a certain power.

Even more conservative requirements were suggested by Haybittle (1971) and Peto *et al.* (1976). According to what proposed by the first author, the study is stopped if the chi-square statistic on one degree of freedom is greater than nine at an interim look or greater than 3.84 at the final analysis.

More particularly, if $\alpha_1 = \alpha_2 \ldots \alpha_{N-1} \sim 0.0027$, it is possible to maintain $\alpha_N = 0.05$ and the overall type I error of about 0.05.

Peto suggested a similar way to proceed: it is very simple ($p < 0.001$ for stopping the study early) and can be appropriate when there is the need of having some flexibility in analysis timing.

Other group sequential families have been proposed by Koepcke *et al.* (1982), Wang and Tsiatis (1987) and Pampallona and Tsiatis (1994). Armitage (1975) and Jones and Whitehead (1979) suggested that continuous sequential boundaries could be applied to the logrank statistic by plotting the logrank score against the Mantel-Haenszel variance.

Gail *et al.* (1981) showed throughout simulations that the group sequential boundaries proposed by Pocock and O'Brien and Fleming could be used with the logrank test, provided the logrank test was performed after successive equal numbers of events.

Sellke and Siegmund (1983) presented asymptotic arguments which imply that group sequential boundaries have appropriate size under the null hypothesis, when the logrank test is performed at intervals defined by equal number of events.

Slud (1984) has shown that sequentially computed logrank tests have uncorrelated increments for various follow-up patterns.

Tsiatis (1981, 1982) demonstrated that logrank increments are asymptotically normal and independent, under the null hypothesis.

DeMets and Gail (1985) showed that boundaries of Pocock, O'Brien and Fleming and Haybittle are still valid, even if the analyses are performed at equal intervals of time instead of events. They are robust also when different test statistics, whose distribution in large samples is approximately normal, are used (Geller and Pocock, 1987).

DeMets and Ware (1980) proposed one-sided group sequential tests. They considered two methods, modifications of repeated significance testing procedures and a third derived from the SPRT of Wald.

In a following paper, DeMets and Ware (1982) produced tests with more stringent requirements for stopping early, based on the two-sided test of O'Brien and Fleming. A common feature of these designs is the lack of simmetry, that give rise to additional parameters in the boundaries and apparently to a certain grade of arbitrariness in the choice of the tests.

One-sided tests can be also derived through the approach of repeated confidence intervals (Jennison and Turnbull, 1989), described in section 1.2.2.5.

Gould (1983) noticed that for not life-threatening diseases, early interruptions due to extremely negative results are appropriate, but if interim results suggest an advantage for the experimental treatment the study should be continued till the planned end, in order to provide adequate informations about secondary endpoints, and/or safety, and/or sub-groups of patients. Similarly, if safety is the primary endpoint, the study should be interrupted only in case of negative results.

Jennison (1987) for constant size groups and a fixed number of analyses derived tests that minimize a target function in order to detect the optimal region of continuation.

Emerson and Fleming (1989) developed a family with one parameter symmetrical designs, whose boundaries are almost totally efficient when compared with the optimal test of Jennison.

1.2.2.3 Alpha-spending functions

Group sequential approaches require number and timing of interim analyses to be specified in advance and this may represent a difficulty in their application. For example, if the study duration is longer than what planned due to a lower accrual rate, the number of annual interim analyses should consequently be changed.

Lan and DeMets (1983), Lan et *et al.* (1989a, 1989b) proposed a more flexible implementation of the group sequential boundaries through an "alpha-spending function", which permits to overcome these restrictions. Previous work of Slud and Wei (1982) was of similar nature.

The spending function allocates the amount of type I error which can be spent at each analysis as a function of the proportion τ of the total information available.

τ, called the information fraction, may be estimated as the fraction of the enrolled patients (or of the observed events) at a given time divided by the total number required on the basis of the sample size calculation. If τ specifies the position of each interim analysis along the trial, the alpha-spending function is a monotonically increasing function on the information fraction $\alpha(\tau)$, $\tau \in [0,1]$, such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$, with α the amount of type I error desired at the final analysis. $\alpha(\tau)$ is defined as the significance level which results if the study is stopped on either boundary when information fraction is τ.

The group sequential boundaries of Pocock and O'Brien-Fleming can be, respectively, approximated and reformulated in terms of the "spending function" as:

$$\alpha_1(\tau) = \quad 2 - 2\phi(\tfrac{1}{2}\eta/\sqrt{\tau}) \qquad \text{OBF} \tag{1.10}$$

$$\alpha_2(\tau) = \quad \alpha \ln[1 + (e+1)\tau] \qquad \text{Pocock} \tag{1.11}$$

where Φ is the cumulative standard normal distribution function and *eta* is the $100(1-\eta)^{th}$ upper percentile of the same distribution.

Many other spending functions have been constructed (Hwang and Shiy, 1990; Kim and Demets, 1987a).

At the i^{th} inspection, Z_i and V_i are computed and compared with symmetric stopping limits l_i and u_i ($l_i = -u_i$), chosen so that the probability of stopping at or before the current inspection under $\theta = 0$ is $\alpha(\tau)$. Then l_i and u_i are progressively chosen to satisfy:

$$P[Z_j \notin (l_j, u_j \quad \text{for some} \quad j = 1, \ldots, i)] = \alpha(\tau), \qquad i = 1, 2, \ldots \tag{1.12}$$

The number and timing of analyses have not to be pre-specified, but only the maximum amount of information to collect has to be anticipated in order to define an alpha-spending function procedure.

The calculation of CIs following the alpha-spending function approach is described in Kim and Demets (1987b).

1.2.2.4 Stochastic curtailment

Stochastic curtailment (Halperin *et al.*, 1982; Lan *et al.*, 1982; Lan and Wittes, 1988), by taking into account the information available at a given interim analysis, tries to predict the final results that would be obtained would the trial continue until the planned end.

It is also referred as conditional power (CP) since it considers the test that would be performed using both data already available and data that would be collected if the trial were prolonged, and judges the properties of this test in the conditional distribution given what was observed so far.

Stochastic curtailed testing was proposed and used as a decisional tool for stopping a trial before its planned termination, when the treatments appear to be convincingly different or if they appear convincingly not different.

The CP is defined, at a given information fraction τ as the probability $p_\tau(\theta)$ that a statistical test will reject the null hypothesis H_0 at the end of the study. The null hypothesis can be accepted at an information fraction τ, assuming that a certain value of θ is smaller than some prespecified value.

Since future data are unknown, several scenarios can be supposed, such as positive trend, negative trend or no trend at all. In fact, we can calculate the probability of different outcomes conditional on observing certain interim results. Thus, several values of θ can be thought of, and consequently the CP values can be quite different depending upon the values chosen (Pepe and Anderson, 1992):

- the parameter value θ_R as specified in the study design under H_1;
- the parameter value based on the data observed so far;
- the parameter value based on a limit of the CI for the parameter estimate.

The decision to stop the study and accept the null hypothesis is based on a CP, calculated under a parameter value θ, falling below some prespecified value of probability π. Pepe and Anderson (1992) recommended values of $\pi \leq 0.3$. Betensky (1997) proposed values ranging from 0.1 (conservative) to 0.3 (nonconservative). Ware *et al.* (1985) used a value equivalent to 0.33.

Finally the optimal information fraction k must be chosen. For Pepe and Anderson (1992) values between 0.25 ond 0.5 have intuitive appeal.

Criticism to CP is that this approach can be very conservative and, furthermore, it does not give any information about θ in terms of point and interval estimation, but only about the likely conclusion of the reference test.

1.2.2.5 Repeated confidence intervals

The method of repeated confidence intervals (RCIs) creates a sequency of 100(1-γ) percent CIs, one at each performed analysis, having the property that they all contain θ with probability $1-\gamma$.

Denoting the i^{th} interval by $(\theta_{Li}, \theta_{Ui})$, the definitions is:

$$P[(\theta_{Li}, \theta_{Ui}) \ni \theta \quad \text{for all i=1,2,\ldots}] = 1-\gamma. \tag{1.13}$$

Each individual interval, including the final one, satisfies

$$P[(\theta_{Li}, \theta_{Ui}) \ni \theta] \geq 1-\gamma. \tag{1.14}$$

and therefore it is more conservative than a "classic" CI, from whom therefore it must be distinguished, with error probabilities around 0.045 rather than 0.05. This conservatorism is the price to be paid for the great flexibility.

RCIs have been described by Jennison and Turnbull (1984, 1989, 1990). Durrleman and Simon (1990) considered their application to non-inferiority trials.

When RCI approach is used as a sequential design, the trial is stopped as soon as the current RCI excludes the value corresponding to treatment equality. If the trial is stopped before the pre-planned end, the intervals will be conservative, since the left-hand side of (1.13) will exceed $1-\gamma$. In the same way, if the trial continues beyond the planned end, no more valid intervals can be defined.

1.2.3 Bayesian approach

The Bayesian philosophy of statistical inference differs from that underlying frequentist approach, the main difference lying in the way they deal with probability. The difference is quite radical, and, although the conclusions reached may be qualitatively similar, the way of expressing and interpretating those conclusions are different.

For frequentists, the probability of an event is the limit of the relative frequency with which it occurs in series of suitably relevant observations in which it could occur. Bayesians, in contrast, interpretate probability as a personal degree of belief of a relevant observer concerning whether the event will or not will occur on a particular observation.

In frequentist analysis the unknown parameters in a statistical model are fixed but unknown quantities, and it is not possible to make probability about them. In Bayesian approach the parameters are random variables, having probability distributions. In Bayesian approach, there may be as many different probabilities of an event as are observers, whereas for frequentists each event has a unique probability. Frequentists talks about their probability as being "objective", in contrast with Bayesian probability termed "subjective", and since subjectivity is thought to connote arbitrariness and bias, they considered frequentist approach more suitable for scientific research.

On the other side, Bayesians assert that a subjective view of probability does not mean that probability is arbitrary and that their approach, using more available information, can produce stronger results than frequentist method and is more appropriate for problems of decision making.

Bayesian probabilities are direct, since a Bayesian analysis of hypothesis results precisely in the probability that it is true, whereas the p-value for frequentist is an indirect measure, being the probability if we repeat the analysis many times to falsely reject the null hypothesis, even if it is usually interpreted to mean the more understandable Bayesian statement.

Bayesian intervals are called "credible intervals" to make it clear that they are different from frequentist confidence intervals. A 95% CI for a certain parameter says that if we repeated the experiment under the same conditions many times, and we calculated an interval each time, then 95% of those intervals would contain the true, but unknown value of the parameter. The degree of probability of a credible interval is really the chance of the parameter lying in the particular interval.

1.2.3.1 Bayesian framework

Bayesians express their prior knowledge concerning a parameter of interest in a prior distribution function. Subsequently they observe data as a result of an experiment. The product of the prior distribution function and the information about this parameter contained in the data and expressed in the likelihood, leads to the posterior distribution function through the so called Bayes' theorem.

The posterior distribution can thus be viewed as an update of the prior information or as the prior belief modified by data. So while frequentist method uses only the likelihood, Bayesian uses both the likelihood and the prior information, the posterior estimates being a compromise between prior and data estimates with a higher precision than either information sources separately.

In mathematical terms, let $P(\theta)$ be the pre-study opinion or prior probability about the treatment effect size and $P(data|\theta)$ be the likelihood of obtaining the observed data, given the effect size, then

$$P(\theta|data) \propto P(data|\theta)P(\theta) \tag{1.15}$$

is the posterior probability about the effect size, given the observed results, upon which any inference is derived.

In fact, when a Bayesian analysis reports a credible interval for a parameter, this is a posterior interval, derived from the parameter's posterior distribution, based not only on the data but also on whatever other information or knowledge the investigator possesses.

When the prior information is very weak, relative to data information, the prior distribution gets so little weight in Bayes' theorem that the posterior distribution is effectively just the likelihood. In this situation Bayesian methods lead to similar inferences to conventional frequentist methods.

1.2.3.2 Prior distribution

The prior information allows Bayesian approach to access more information and thus to produce stronger inferences, even though it is the main butt of criticisms of frequentists due to its subjectivity. This drawback can be overcome, by choosing a prior information based on a widespread evidence and examining a sufficient number of observations, so that differences in prior positions can be shaded.

We need to specify the prior distribution with sufficient reliability and accuracy. It is important to note that the choice of the prior may not be necessarily unique, but a range of prior distributions could be presented, thus reflecting different perspectives. This can include the subjective prior opinion of the trial investigators and/or other experts, as well as the results of previous similar studies.

Possible priors include (Spiegelhalter *et al.*, 1993; Parmar *et al.*, 1994):

- *uninformative* prior, representing a lack of prior opinion or information as to the likely treatment difference. It is the more unrealistic prior, corresponding more or less to the frequentist approach of significance testing. Nevertheless such a prior is the least subjective and can be used as a reference against which to measure the impact of the choice of other priors;

- *clinical* prior, representing the opinion of experts. It may change during the course of the trial, since the clinician view can be influenced for instance by new external evidence;

- *sceptical* prior, representing the opinion of someone who, unenthusiastic about the treatment under study, thinks that there is only a small probability that the alternative hypothesis is likely to be true. It represents some scepticism about the treatment effect and its conservatorism leads to a behaviour comparable to that of group sequential designs (Freedman and Spiegelhalter, 1989). It is useful in order to counteract over-enthusiastic opinion due to extremely positive results that could be observed by chance during a clinical trial;

- *enthusiastic* prior, representing the opinion of individuals who are persuaded that experimental treatment effect is greater than control.

The argument in favor of representing no prior information is that this avoids any criticism about subjectivity. There have been numerous attempts to find a formula for representing prior ignorance, but without any consensus.

The idea of using "sceptical" prior is that if a sceptic can be persuaded by the data, then anyone with a less sceptical prior position would also be persuaded, denoting that the data are strong enough to reach a firm conclusion. If, on the other hand, the data are not strong enough to yield a high posterior probability for that hypothesis, then we should not yet claim any definite inference about it. This prior can be useful to counterbalance early positive results, which could be lead to premature termination of the study.

On the other side, the enthusiatic prior distribution is useful when initial results suggest a detrimental effect of the experimental treatment, again to avoid a premature termination of the trial.

As already explained, the usual approach to specifying a prior distribution for some parameters consists of first specifying a few features of the distribution, such as a prior expectation and some measure of prior uncertainty (e.g. the prior variance), then choosing a suitable distribution to fit these features.

It is sensible to make choice of distributions on grounds of simplicity and convenience. Mathematically, in some simple statistical problems there exist classes of priors known as conjugate priors that are particularly convenient.

In fact, while the likelihood function is often determined by the nature of data, as a rule, prior distribution can be of any form. In practice the analysis is simpler if the prior distribution is chosen so that the posterior is a member of the same distributional family. Such a family is called "conjugated" for that particular likelihood function.

As an example, when both the prior and the likelihood are normal, also the posterior has a normal distribution with mean lying between the mean of the prior and the observed effect. The posterior distribution variance is smaller than the prior one, since further data have been incorporated (Abrams *et al.*, 1994).

For time to event data, Tsiatis (1982) has shown that the quantity $4Z/n$, where n is the total observed number of events in the two groups, has an asymptotic normal distribution with mean θ and variance $4/n$. Therefore, supposing that the prior distribution for the $\ln(HR)$ is normal with mean μ_0 and variance σ_0^2, and the likelihood from the data has mean μ_L and variance σ_L^2, then the posterior distribution is also normally distributed with mean $\frac{(\mu_0\sigma_L^2+\mu_L\sigma_0^2)}{(\sigma_0^2+\sigma_L^2)}$ and variance $\frac{(\sigma_L^2\sigma_0^2)}{(\sigma_0^2+\sigma_L^2)}$.

1.2.3.3 Trial monitoring

When cumulative data from a trial are analyzed sequentially following the Bayesian approach, the posterior distribution describes the currently available information about the parameter of interest. This information can be used to decide whether to stop the study because enough evidence is already gathered or whether additional evidence is needed.

Two important features of Bayesian analysis are that all available information can be used in deciding whether to stop a trial. First, the decision of an early termination may depend not only on data from the trial but also on external information, which may have become available after the beginning of the study. In this latter case the prior may be changed in order to take into consideration this new information. Second, it is more flexible, since interim analyses can be introduced without affecting the final conclusions; they do not need to be planned in advance and there is no penalty for the repeated analyses, due to the lack of dependence on study design.

Decision theory provides another good example of the flexibility of Bayesian inference. In this theory we have a set of possible decisions and an utility function that specifies how good it would be to make a particular decision, if the parameters turned out to have particular values. For instance, for a hypothesis test we could define an utility function that states it would be good (high utility) to accept the hypothesis if it turned out to be true, or to reject it if it turned out to be false, but otherwise the utility would be low.

If we knew the parameters, it would be easy to reach a decision, since we would just choose the decision with the largest utility for those values of the parameters. However, the parameters are generally unknown. Decision theory says we should choose the decision with the highest (posterior) expected utility. This expectation is the value of the utility, averaged with the posterior distribution of the parameters.

There are two Bayesian approaches to stopping decision. The first is the decision theoretic approach (Anscombe, 1963; Colton, 1963), in which a fixed total number of patients is assumed and costs and utilities to various decisions and outcomes are assigned: the consequences of continuing and of stopping a trial are weighted using the current distribution of θ. In practice, since it is difficult to quantify the number of the patients and the costs, this approach has been criticized (Peto, 1985) and rarely applicated. The second Bayesian approach is enterely based on the posterior distribution of θ. It is often useful to present the results of an interim analysis under several alternative prior distributions, so that the impact of the data may be reviewed according to different levels of scepticism.

Chapter 2

Aims

Several important questions have been raised about decision of stopping a trial early, and on what basis to reach such a decision. Differences between treatments may be larger than expected or unanticipated adverse effects may occur, and either of these may justify early termination of a trial. Early stopping may sometimes be suggested because continuing a trial would not provide sufficiently useful information to warrant continuation. Existing evidence from outside the trial, such as meta-analyses of data from comparable trials, other existing evidence external to the trial and the nature of the condition and its alternative treatments may be also taken into consideration when deciding to stop or continue a trial. Data from the trial may later constitute the base for management of future patients, therefore the evidence should also be sufficiently convincing to the wider clinical and patient communities to determine future practice.

In chronic life-threatening diseases, like cancer, evidence of early therapeutic - benefits may be even more compelling than in other diseases and the detection of the best trade-off between obtaining results as earlier as possible and getting good estimation of the magnitude of treatment effect is particularly important.

A range of formal statistical approaches can be used as a basis for judging at what point such differences are so extreme as to be sufficiently unlikely to reflect the play of chance. These analyses help to control for errors in decision making and estimation; however, although interim analyses in clinical trials are widely known, information on explicit adoption of some form of planned monitoring, even in long-term trials, is still scarce and basically driven by the published reports of studies, which rarely include details on both the strategies for data monitoring and interim analysis plan. Further research is therefore needed for comparing the performance of different statistical approaches in real life situations.

It seems therefore of interest to investigate the forms of monitoring used in cancer clinical trials and in particular to gather information on the role of interim analyses in the data monitoring of a clinical trial.

More specifically, the project addresses the following issues:

- what are the operative characteristics of different interim analysis approaches;
- how often interim analyses are used and which types of statistical analysis are more frequently adopted;
- how information derived from statistical analyses is taken into consideration in the decisional process.

Regarding the first aim, research will focus on comparison of the more commonly used statistical approaches to interim analysis on real data derived from cancer clinical studies, in order to increase knowledge on their relative efficiency, in terms of anticipation of the final effect and accuracy of the effect size estimation. Potentially influencing factors, such as severity of disease, magnitude of treatment effect and number of analyses will be taken into consideration, too.

As for the second: to investigate how often interim analyses are used and which types of analysis are more frequently adopted, the project will focus mainly on the prevalence of the different types of interim analysis and of data monitoring in clinical trials. Source of investigation will be literature data and the protocols of cancer clinical trials included in Italian registry of clinical trials.

The final aim, focusing on the actual impact of results of interim analyses on the decision of modifying the planned study conduct, deals with two separate tasks: first, to provide information on how often an early termination is caused by results of an interim analysis. For this proposal, causes of early interruption will be investigated based on published papers of preliminary or early clinical trial results. Secondly, to obtain information from the literature on the modality of implementation of interim analyses in the data monitoring, investigating which are the more reported forms and rules for such a process.

Chapter 3

Methods

3.1 Comparison of different statistical approaches

3.1.1 Monitoring approaches

Among all those proposed for study monitoring, the following methods have been considered:

1. **Frequentist**

   - **Naïve**, in which no adjustment for multiple testing has been used;

   - **alpha-spending function** (Lan *et al.*, 1983; DeMets *et al.*, 1994) with OBF boundaries. The spending function controls the rate at which the total α error is spent as a continuous function of information fraction available at the moment of the analysis and thus determines the corresponding boundary. The nominal significance levels at each analysis depend on the prespecified overall α error, and on the number of analyses. At the final analysis, the nominal significance level is close to that of the conventional fixed sample design (see Section 1.2.2.3);

- **Triangular test**, introduced by Whitehead (1983, 1992, 1994), in which the boundaries are a function of the information fraction and the trial stops when the sample path crosses a boundary. The boundaries depend on prespecified α and β errors, on the expected value of θ_R, and on the expected rate of events in the control group. After each analysis, in order to maintain the total α error at the desired level, boundaries are adjusted thus assuming a "christmas tree shape" (see Section 1.2.2.1);
- **Restricted procedure** (Armitage, 1957, 1975; Whitehead, 1992), another boundary test, which if the slope value is set equal to zero, as in our case, produces boundaries similar to those suggested by O'Brien and Fleming. Like the triangular test, also in the restricted procedure boundaries depend on α, β and θ_R (see Section 1.2.2.1);

2. **Bayesian**, choosing three different prior distributions (Fayers *et al.*, 1997)

   - **Uninformative prior**, that contains no information about prior opinion on treatment effect:

$$Uninformative\ prior \sim N\left(\ln(1), \infty\right).$$

   Even if a normal distribution cannot have an infinite variance and therefore this is an improper uniform distribution, it will anyway used for mathematical considerations;

   - **Sceptical prior**, with mean equal to 0 and a precision such that the prior probability of a true effect as large or larger than what stated by the alternative hypothesis is small, say 5%. The sceptical prior is equivalent to having performed a trial with sample size equal to N_p subjects all of whom have died, and in which no difference has been observed between arms:

$$Sceptical\ prior \sim N\left(\ln(1), \frac{4}{N_p}\right);$$

- **Enthusiastic prior**, corresponding to consider as best guess the alternative hypothesis and having the same precision as the sceptical prior:

$$Entusiastic\ prior \sim N\left(\ln(HR_1), \frac{4}{N_p}\right)$$

where $\ln(HR_1)$ is the natural logarithm of the HR under the alternative hypothesis.

In the Bayesian approach, stopping rules are based on the posterior probabilities. A reasonable criterion is that the posterior probability of one treatment being better, using a sceptical prior opinion, is at least 95%. Alternatively, if a non-zero target is sought, a reasonable criterion might be to accept a posterior probability of at least 90% (Fayers *et al.*, 1997) (see also Section 1.2.3).

3.1.2 Reanalyzed clinical trials

Three published phase III trials, coordinated in Italy by the Istituto di Ricerche Farmacologiche "Mario Negri" of Milan, were considered.

All these studies were aimed at demonstrating superiority, in terms of overall survival, of the experimental treatment versus control. They were chosen to represent different clinical settings for long-term prognosis, recruitment duration and expected effect size, as well as different conclusions that a clinical trial may reach.

ICON 3 Trial (The International Collaborative Ovarian Neoplasm (ICON) Group, 2002) showed no statistically significant difference between arms, while ICON 4/AGO-OVAR 2.2 study (The ICON and AGO Collaborators, 2003) and the GIVIO/SITAC 01 Study (Zaniboni *et al.*, 1998) demonstrated the superiority of experimental treatment.

For each trial, the initial hypothesis on the two-sided α error rate, the power, the expected benefit of the experimental treatment over control and the expected rate of events in the control group were retrieved from the protocols. Using this information, for each trial the monitoring process was rebuilt according to the previously described methods.

Our intention was to compare conclusions drawn from each monitoring method with those actually obtained.

3.1.2.1 ICON 3 Trial

This study was aimed at comparing the safety and efficacy of paclitaxel plus carboplatin with a control of either cyclophosphamide, doxorubicin and cisplatin (CAP) or carboplatin alone in women with advanced ovarian cancer.

"...An independent data-monitoring and ethics committee (DSMC), comprising two clinicians and one statistician who had no involvement in the trial, was established to review the progress and confidential unmasked results, and any other relevant external evidence, about once a year during the accrual period. The committee did not follow any predetermined statistical stopping rules... We expected that 2-year survival would be about 50% in the control groups of the trial. With a 2:1 ratio in favour of control, an initial accrual target of 1000 patients in the control groups was set to allow the reliable detection of an absolute difference in 2-year survival of 10% (from 50% to 60%) with 85% power at the 5% significance level. This absolute difference

translates into a hazard ratio of 0.74. With 1000 patients entered (and about 460 deaths), the 95% CI for a difference of 10% would be estimated as 4-15%. At their first meeting on July 1, 1996, the DSMC recommended that the trial size be increased to allow smaller differences to be detected reliably. This recommendation was endorsed by collaborators at meetings in July, 1996, and a new target of 2000 patients was agreed. This target accrual was sufficient to allow the reliable detection of an absolute difference of 7% in 2-year survival (from 50% to 57%) with 85% power at the 5% significance level, corresponding to a hazard ratio of 0.81. With 2000 patients entered (and about 955 deaths), the 95% CI for an absolute difference of 7% would be estimated as 3-11%. The DSMC reconsidered the evidence at the end of May, 1997, the week after the results of OV 10 and GOG-132 became available, when 1254 patients had been randomised into ICON 3. The committee recommended that the trial continue recruitment to 2000 patients, and this suggestion was endorsed by collaborators at meetings early in June, 1997... ".

Between February, 1995 and October, 1998 2074 women were therefore enrolled. The final analysis was performed at a median follow-up time of 51 months after the observation of 1286 events. The estimate of the treatment effect resulted in a hazard ratio of 0.98 (95% CI 0.87-1.10) with a p-value of 0.74, using the log-rank test.

The authors concluded that *"... up to 5 years from treatment, single-agent carboplatin, CAP, and paclitaxel plus carboplatin are all safe and show similar effectiveness as first-line treatments for women requiring chemotherapy for ovarian cancer. Of these three treatments, carboplatin might be regarded as the preferred treatment because of its better toxicity profile,... "*.

3.1.2.2 ICON 4/AGO-OVAR 2.2 Trial

Run as two parallel trials, this study investigated whether paclitaxel should be given in addition to platinum-based chemotherapy in patients with ovarian cancer relapsing 6 or more months after the end of the previous line of platinum-based chemotherapy, and therefore judged to have platinum-sensitive disease and who would otherwise be treated with more conventional platinum-containing regimens.

"... For ICON 4, an independent data monitoring and ethics committee was e-stablished, of two clinicians and one statistician who had no involvement in the trial. Progress and unmasked results of the trial and any other relevant external evidence were reviewed roughly once yearly during the accrual period. The committee followed no predetermined statistical stopping rules. A data monitoring and ethics committee was not established for the AGO trial....

... When the original sample size calculations were made, few data were available on the outcome of relapsed ovarian cancer. We expected that 2-year survival would be about 5% in the control group and the absolute difference in 2-year survival of no more than 5-10% (5-10%) with 95% power at the 5% significance level, corresponding to a hazard ratio of 0.77. At the final analysis on Feb 8, 2001, the data monitoring and ethics committee noted that the survival in the control group was much higher than in the original power calculations; two-year survival in the conventional treatment group was around 50%. We therefore revised the power calculations. We calculated that 800 patients would be sufficient to detect reliably an absolute difference of 11% in 2-year survival (50-61%) with about 90% power at the 5% significance level, corresponding to a hazard ratio of 0.71. Hence the target accrual remained unchanged....

... From January 1996 to March 2002, 802 patients were enrolled in the study. ... By March, 2003, with a median follow-up of 42 months, 530 patients had died. Survival curves showed a hazard ratio of 0.82 (95% CI 0.69-0.97, p=0.02)... ".

The authors concluded *"Our findings suggest a beneficial effect for paclitaxel in combination with platinum chemotherapy on survival and progression-free survival among patients with platinum-sensitive relapsed ovarian cancer"*.

3.1.2.3 GIVIO/SITAC 01 Trial

This study was designed to assess whether 5-fluorouracil and high dose of folinic acid (HD-FUFA) would increase the overall survival of patients with resectable Dukes B and C colon carcinoma. Early results were published as a part of an international multicenter pooled analysis (IMPACT Investigators, 1995).

The original plan for the study was to detect a 30% relative mortality reduction at 80% power with a conventional two-tailed test of 5%. No DSMC was established for the study.

Overall, 888 patients were randomised, while 869 were considered eligible and therefore included in the analysis. The median follow-up time for the HD-FUFA and control arms were 65 and 63 months, respectively. HD-FUFA significantly reduced the rate of mortality of 25% (95% CI 5-41%, p=0.02): there were 120 deaths in the treatment arm and 159 in the control arm.

Authors concluded that *"this study confirmed that adjuvant therapy clearly reduced mortality among patients with colon carcinoma... "*.

The main characteristics of the three studies are summarized in Table 3.1.

Table 3.1: **Main characteristics of the reanalysed clinical trials**

Study	Duration (yrs) of study	Duration (yrs) of enrollment	Patients	Events	Observed HR (95% CI)	p
ICON 3	7.7	3.7	2074	1286	0.98 (0.87-1.10)	0.74
ICON 4	6.8	6.2	802	530	0.82 (0.69-0.97)	0.02
SITAC 01	7.4	3.0	869	280	0.75 (0.59-0.95)	0.02

3.1.3 Interim analysis planning

For each study, interim analyses were performed at equal intervals of events, precisely at 25, 50, 75 and 100% of the total number of events observed. In reality, to conduct analyses at constant time intervals turns out to be more practicable, because it allows to schedule in advance the DSMC meetings. However, it is anticipated that results are not affected by the approach chosen to select the cut points of the analysis (Fleming and DeMets, 1993).

3.1.4 Statistical analysis

Individual patient data retrieved included:

- Patient code;
- Randomisation date;
- Treatment code;
- Event code;
- Date of event, if present, otherwise date of last follow-up.

Thus retrospectively the progress of each trial was rebuilt and the different monitoring procedures were applied. The date the trial would have been stopped was then determined according to each method. All the sequential re-analyses are idealized, in that all events had occurred by x years are included in the analysis at x years, while in the practice, this is seldom possible, since a reporting lag is always detected.

Differences in survival between treatments were assessed using the log-rank test. HRs and their 95% CIs were calculated using Cox's proportional hazards regression model (Cox, 1972) at the date of stopping decision provided by each method.

When group sequential test procedures were used, standard confidence intervals were adjusted taking into account the sequential nature of analyses. For triangular test and restricted procedure, if not otherwise specified, the 95% CIs for the estimate of the HR were computed using the Woodroofe approach, that allows for the fact that on termination of a sequential trial the distribution of z may not be the standard normal (Woodroofe, 1992).

Point and credibility interval estimates are easily calculable using the Bayesian methods inherent in the non informative, sceptical and enthusiatic schemes.

The monitoring and the sequential analyses using the triangular test and the restricted procedure were performed with the Planning and Evaluation of Sequential Trials (PEST 4.0) statistical software (Medical and Pharmaceutical Statistics Research Unit, 2000), kindly provided by the authors.

The Lan and DeMets method was performed both with the program called LANDEM (Reboussin *et al.* 1996, 2000), publicly available on Internet, and with East software (Cytel Software Corporation, *East* Cambridge: Cytel Software Corporation, 2003).

Otherwise, statistical analyses were computed using SAS (Statistical Analysis System, SAS Institute Inc., Cary, NC, US, Version 8.20) software.

3.1.5 Simulations

We evaluated the properties of the frequentist methods by computer simulations, too.

A SAS macro routine, reported in Appendix A, was developed for computing distribution of ln(HR) at specific time points. The routine allowed for the choice of:

- number of interim analyses;
- underlying hazard ratio;
- probability of event at a particolar time point;
- accrual duration time (years);
- total study duration (years).

Assumptions were made on the uniformity of accrual over time, on exponential distribution of survival times and on normality of distribution of ln(HR) at each cut point.

For each of the three studies (ICON 3, ICON 4, GIVIO/SITAC 01) a total of 10000 runs (denoted *trials*) were generated in order to compare the naïve approach, alpha–spending function with OBF boundaries, the triangular test and the restricted procedure, having the fixed sample size design as benchmark.

We used the actual values of the underlying hazard ratio, of the probability of the event, of the total duration of the study, as well as of the accrual period.

We also considered the consequences of performing the analysis four times (at 25, 50, 75 and 100% of the total number of observed events) and eight times (at 12.5, 25, 37.5, 50, 62.5, 75, 87.5 and 100%).

3.2 Early reports in scientific literature

In order to investigate the characteristics of early publications in the scientific literature, we performed an electronic search, focusing on the description of the type of statistical approaches used.

Source of investigation of literature data was PubMed, available via Entrez retrieval system, and developed by the National Center of Biotechnology Information (NCBI) and by the National Library of Medicine (NLM), located at the National Institute of Health in Bethesda (United States).

PubMed was designed to provide access to citations from biomedical literature, as well as to full-text articles at journal Web sites and other related Web resources. It also provides access to bibliographic information that includes MEDLINE, OLDMEDLINE, and citations that precede the date when the journal was selected for MEDLINE indexing, and some additional life science journals, submitting full text to PubMedCentral and receiving a qualitative review by NLM.

Research strategy focused on cancer randomised clinical trials (RCTs), published from January 2000 to February 2005, and presenting either in the title or in the abstract at least one of the the following terms: "interim", "early", "preliminary", "pilot".

Formally, the research was conducted in the following way:

Field: **Title/Abstract**, Limits: **Publication Date from 2000, Randomized Controlled Trial, Cancer**

(preliminary[Title/Abstract] OR early[Title/Abstract] OR interim[Title/Abstract] OR pilot[Title/Abstract] AND Randomized Controlled Trial[ptyp] AND cancer[sb] AND ("2002"[PDAT]: "3000"[PDAT)].

In order to select studies for further assessments, two independent reviewers scanned the title, abstract section and keywords of every retrieved record. Full articles were taken into account, if the information given suggested that:

- the study was a randomized clinical trial (RCT) in oncological field;
- the primary endpoint was a time to event variable;
- the publication concerned the results of an analysis different from that planned as final.

If there was any doubt regarding these criteria from the information given in the title and abstract, the full article was retrieved for clarification. If different opinions existed, they were resolved by discussion.

A template data extraction form was developed and tested in a pilot phase. Data extraction have been performed independently by two evaluators. Differences in data extraction was resolved by consensus with a third reviewer, referring back to the original article. This form included the following items:

- Journal;
- Journal impact factor;
- Year of publication;
- Type of disease;
- Stage of disease (early/advanced);
- Experimental treatment;
- Control treatment;
- Reason for publication (Interim analysis/ Preliminary results not related to primary efficacy endpoint);

- Presence of DSMC;
- Presence of planned stopping rules;
- Type of stopping rules;
- Analysis performed before accrual termination (Yes/No);
- Results (Study continued/Study stopped for futility/Study stopped for efficacy/Study stopped for other reasons);
- Planned sample size.

Results were described using absolute and relative frequencies, and contingency tables.

A multiple linear regression model was used for estimating the association between importance of journal chosen for publication, represented by the impact factor, the decision of stopping the trial based on results of interim analysis, and presence of DSMC.

A logistic regression model was also performed in order to analyse the association among decision following early analysis results and potential determinants, such as presence of DSMC, presence of interim analysis plan and study sample size.

Analysis was performed using SAS (Statistical Analysis System, SAS Institute Inc., Cary, NC, US, Version 8.20) software.

3.3 Use of interim analyses in randomized oncological trials

We also investigated the kind of interim analyses planned in clinical oncological trials currently submitted for approval to Italian Ethics Committees (ECs).

Reason for this research stems from the observation that analysis of published trials may reflect the criteria adopted by trials planned even several years before and as a consequence data derived even from the most recent published literature may not be completely appropriate to represent the current trends of interim analysis and data monitoring use.

Moreover in published literature the quality of details relative to the description of statistical methods is often scarce, and it is possible that, despite the accuracy of literature search strategy, the information regarding the approaches adopted for monitoring clinical trials is not completely captured.

Since 2002 an electronic registry of all clinical trials submitted to ECs is active in Italy. The registry was initially aimed at mapping clinical researches conducted in Italy and at facilitating the exchange of information among ECs, their coordination, and a shared decision on authorizing the launch of a particular clinical trial. It may also offer a great opportunity for conducting epidemiological public health researches on clinical trial strategies, by investigating the type and relevance of current clinical research questions, as well as the quality of the design and other methodological criteria adopted by researches currently ongoing in Italy.

Our project is actually one of the first examples of research which takes advantage of the registry, because in order to investigate how often interim analyses are used and which types of analysis are more frequent in clinical trials conducted in oncology, we had the opportunity of evaluating the protocols in the *"Osservatorio Nazionale sulla Sperimentazione Clinica dei Farmaci", OsSC* (National Monitoring Centre for

Clinical Trials) of Italian Ministry of Health[1].

We assessed protocols available in the *"Osservatorio"* database, relative to oncological studies submitted to ECs from January 1^{st}, 2000 to May 18^{th}, 2005, restricting the evaluation to protocols of randomised studies with a time to event endpoint, such as overall survival (OS) or progression free survival (PFS).

Similarly to the previous project on early reports in scientific literature, a template data extraction form was developed and tested in a pilot phase. Data extraction have been performed independently by two evaluators. Differences in data extraction were resolved by consensus with a third reviewer, referring back to the original protocol. The form included the following items:

- Identification number;

[1]The *"Osservatorio"* is an instrument developed to improve coordination and surveillance of clinical trials on drugs conducted in Italy.

In fact, to be initiated, every trial has to obtain the release of the motivated opinion of the EC that has to take into account, while evaluating the study protocol, its scientific relevance and technical aspects, such as informed consent, assurance and adequacy of medical facilities. The first step in such a process, is the central registration with the OsSC. Afterwards, the results of the assessment made by each local EC, together with the details of each study - type, phase, therapeutic categories, etc., have also to be reported within the OsSC database. Then, this information becomes accessible to all ECs, so that the assessment process becomes as transparent as possible. Access, however, is restricted to ECs and data are not available to the general public yet.

The informative support of the *"Osservatorio"* consists of three on-line registers which form the database of clinical trials. The Ministry of Health, the local ECs, the sponsors, the Regions and the autonomous Provinces can access the information on these registers, according to their organizational needs. The registers are the following:

- a Register of the local ethics committees;
- a Register of private clinical sites;
- a Register of clinical trials.

- Experimental phase (I/II/III/IV);
- Year of EC opinion release (year for study protocols not yet released was conventionally established as *2005*);
- Type of sponsor (Profit/No profit);
- Involved countries (Italy/Europe/Worlwide);
- Study objective (Superiority/Non inferiority/Mixed);
- Setting [Early (adjuvant or neoadjuvant)/Advanced (locally advanced or metastatic or mixed) for solid tumors; pediatric/adult for haematological tumors];
- Primary endpoint (OS/PFS/Event Free Survival/Recurrent Free Survival);
- Disease localization;
- Number of arms;
- Experimental treatment;
- Expected number of events at the end of the study;
- Expected number of patients at the end of the study;
- Planned number of centers;
- Study total duration (months);
- Accrual duration (months);
- Follow-up duration (months);
- Presence of interim analyses;
- Number of interim analyses, if planned;

- Type of interim analyses, if planned;

- Objective of interim analyses (Efficacy/Safety), if planned;

- Presence and type of timing of interim analyses, if planned;

- Presence of DSMC;

- Composition of DSMC, if present;

- Tasks of DSMC, if present.

Results were described using adequate descriptive statistics, such as absolute and relative frequencies, and contingency tables.

A logistic regression model was also performed in order to assess the association among presence of interim analyses and/or DSMC and potential determinants, such as type of sponsor, involved countries and year of submission.

Analysis was performed using SAS (Statistical Analysis System, SAS Institute Inc., Cary, NC, US, Version 8.20) software.

Chapter 4

Results

4.1 Comparison of different statistical approaches

The results of the analyses performed applying the different statistical approaches on real data are reported later on.

Data derived by simulations are also presented with the purpose of quantifying the probability of early stopping and the estimates of treatment effect, when calculated in case of early interruption.

The most important assumption that should be taken into account for interpreting these results is that the estimates found at the end of the study are to be considered as the "true" treatment effect and therefore taken as reference.

Analyses were performed assuming interim looks at regular proportion of events: in order to assess the effect of an increase in the number of looks, the simulations were performed twice: firstly adopting four analyses, then eight analyses.

For each study, results obtained according to the previously described approaches are presented later on. In the relevant tables, the red colour has been used for frequentist methods to emphasize an analysis in which the relevant stopping boundary has been crossed. For Bayesian approach it denotes the results of an analysis in which either the posterior probability of one treatment being better using a sceptical prior is at least 95 per cent or, if a non-zero target improvement is sought, a posterior probability of at least 90 per cent, as well as a 95% credibility interval that does not include the value of one.

In the tables the term "z-value" refers to the standardized normal statistic, while in the figures $Z = -Z^*$ is reported, where Z^* is the logrank statistic, according to Whitehead (1983) (see Section 1.2.1).

All reported p-values are one-sided.

4.1.1 ICON 3 Trial

Tables 4.1, 4.2, 4.3 and 4.4 report the results of naïve method (analysis without adjustments), alpha-spending function with OBF boundaries, triangular test, and restricted procedure, respectively. Plot of results of these analyses are also reported in Figures 4.1, 4.2, 4.3.

Table 4.5 shows estimates (HRs and 95% CIs) calculated at study closure, considered as the first crossed boundary or the last scheduled analysis, according to the results obtained by each considered method.

Tables 4.6, 4.7 and 4.8 report the results obtained adopting Bayesian approach.

Of note, the naïve method, which for all analyses considers as stopping boundary the value of the standardized normal statistic equal to 1.96, corresponding to an one-sided p-value equal to 0.025, would have conducted to early stopping at the second analysis (HR=0.84; 95% CI: 0.71-0.99; p=0.02, one sided), with a relative overestimate of 13%.

The triangular test, which allows for an asymmetric power design, requiring a high effect magnitude to detect superiority of experimental treatment, while more relaxed boundaries against refusal of the null hypothesis, would have conducted to conclude for futility at the fourth (final) analysis. The observed p-value was 0.10, the unbiased estimate of the HR was 0.95 (95% CI: 0.85-1.08).

OBF approach and restricted procedure, which both require very extreme results in order to stop, particularly in the early analyses, show superimposable results: with these approaches the study would have not been interrupted.

Although it is rather difficult to directly compare Bayesian with frequentist approaches, Bayesian approach showed that under the sceptical prior, the point estimate at the second look is 0.88 and the relative overestimate is reduced to 8%. The probability of an effect size larger than zero is 96% at the second look, while the probability of an effect size larger than 2.5% is reduced to 76%, but the probability to obtain at least a 5% reduction in mortality is 36%, suggesting that the benefit, if any, is very small. At the end of the study, the probability of some effect is still quite high under the sceptical prior (71%), but the chance to have a clinically relevant effect (say at least 5% reduction in mortality) is 1% under the sceptical and 6% under the "enthusiastic" prior, respectively.

Table 4.9 reports the results of the simulation with four looks.

As also summarized in Table 4.31, using the naïve method, the overall chance of early stopping (calculated as the percent of the number of *trials* with HR point estimate outside the boundaries, divided by the total number of *trials*) was 17.2%: 14.9% of *trials* crossed the lower boundary, indicating an effect in favour of experimental arm, while 2.3% crossed the upper boundary. The larger proportion of *trials* outside the boundaries was observed in the first and second looks (58% of *trials* crossing the lower boundary and 77% of *trials* crossing the upper boundary occurred in these first two analyses).

The median HR of *trials* with results crossing the lower boundary ranged from 0.77 to 0.89, and on the average, the overestimate was 13.6%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 1.12 to 1.28, with an underestimate of effect equal to 27%.

Simulation adopting alpha-spending function with OBF boundaries and restricted procedure produced similar results.

For alpha-spending function the percentage of early stopping was 9.5% (0.6% at upper boundary and 8.9% at the lower boundary) and the chances of early stopping were higher at the last two analyses (93.4% of early interruptions for crossing the lower boundary and 96.5% of interruptions for crossing the upper boundary). The median HR of *trials* with results crossing the lower boundary ranged from 0.78 to 0.88, and on the average, the overestimate was 10.9%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 1.13 to 1.30, with an underestimate of effect equal to 19%.

A similar pattern was produced by the restricted procedure: the percentage of early stopping was 10.6% (0.7% at upper boundary and 9.9% at the lower boundary) and the chances of early stopping were also higher at the third and fourth analyses (89.7% of early interruptions for crossing the lower boundary and 94.5% of interruptions for crossing the upper boundary). The median HR of *trials* with results crossing the lower boundary ranged from 0.79 to 0.88, and on the average, the overestimate was 10.7%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 1.13 to 1.29, with an underestimate of effect equal to 19%.

The adoption of triangular test would have conducted, instead, to 95.3% of early interruptions (87.8% at upper boundary and only 7.5% at the lower boundary). The higher percentage of interruptions (64.8% of *trials* crossing the lower boundary and 74.6% of *trials* crossing the upper boundary) occurred at the second and third analyses. Also in this case, there was a convergence of the median HR towards the "real" effect: the median HR of *trials* with results crossing the lower boundary ranged from 0.70 to 0.87, and on the average, the overestimate was 13.6%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 0.93 to 1.16, with an average underestimate of effect equal to 3.7%.

Table 4.10 reports the results of the simulation using eight looks.

With the naïve method, the overall chance of early stopping was 24.2%: 20.0% of the *trials* crossed the lower boundary. The larger proportion of *trials* outside the boundary was observed in the first 4 looks (64.3% of *trials* crossing the lower boundary and 84.0% of *trials* crossing the upper boundary, occurred in these four earlier analyses). The median HR of *trials* with results crossing the lower boundary ranged from 0.70 to 0.89, and on the average, the overestimate was 16.1%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 1.12 to 1.43, with an underestimate of effect equal to 35.7%.

For the alpha-spending function the percentage of early stopping was 9.6% (0.4% at upper boundary and 9.2% at the lower boundary, respectively) and the chances of early stopping were higher after the fourth analysis (82.6% of early interruptions for crossing the lower boundary, and 92.7% of interruptions for crossing the upper boundary). The median HR of *trials* with results crossing the lower boundary ranged from 0.57 to 0.88, and on the average, the overestimate was 11.4%. For *trials* stopped because results crossed the upper boundary, the median HR ranged from 1.13 to 1.40, with an underestimate of effect equal to 21.7%.

For the restricted procedure, the percentage of early stopping was 9.1% (0.5% at upper boundary and 8.6% at the lower boundary) and again the chances of early stopping were higher after the fourth analysis (93.3% of early interruption for crossing the lower boundary, and 74.5% of interruptions for crossing the upper boundary). The median HR of *trials* with results crossing the lower boundary ranged from 0.71 to 0.88, and on the average, the overestimate was 11.4%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 1.13 to 1.62, with an underestimate of effect equal to 21.9%.

With the triangular test 94.0% of *trials* interrupted early (87.0% at upper boundary, but only 7.0% at the lower boundary). Again, the higher percentage of interruptions (71.1% of *trials* crossing the lower boundary and 75.9% of *trials* crossing the upper boundary) occurred between the third and sixth analysis. The median HR of *trials* with results crossing the lower boundary ranged from 0.53 to 0.87, and on the average, the overestimate was 14.3%. For *trials* stopped because results crossed the upper boundary, the median HR ranged from 0.92 to 1.56, with an underestimate of effect equal to 4.4%.

4.1.1.1 Frequentist approaches

Table 4.1: **Analysis without adjustment for multiple tests**

				Obtained		**Stopping boundary**	
N	**Date**	**Pts**	**Events**	**z-value**	**p-value**	**z-value**	**p-value**
1	28/09/1997	1577	323	+0.56790	0.28505	±1.96000	0.02500
2	15/10/1998	2069	643	+2.04657	0.02033	±1.96000	0.02500
3	17/12/1999	2074	965	+1.75946	0.03926	±1.96000	0.02500
4	05/09/2003	2074	1286	+0.60437	0.27279	±1.96000	0.02500

Table 4.2: **Alpha-spending function with O'Brien and Fleming boundaries**

	Obtained		Stopping boundary	
N	z-value	p-value	z-value	p-value
1	+0.56790	0.28505	±4.33263	0.00001
2	+2.04657	0.02033	±2.96311	0.00152
3	+1.75946	0.03926	±2.35902	0.00916
4	+0.60437	0.27279	±2.01406	0.02200

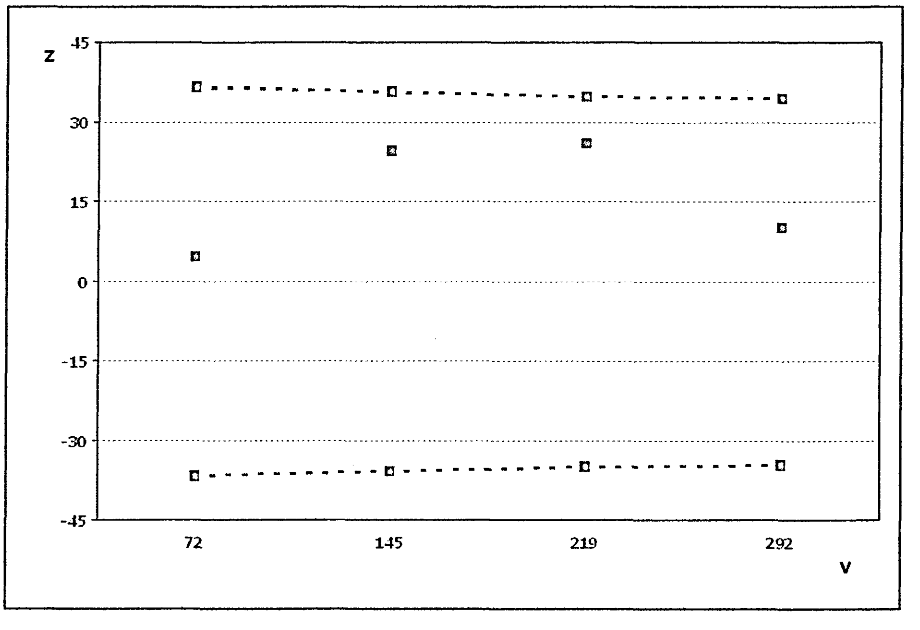

Figure 4.1: **Alpha-spending function with O'Brien and Fleming boundaries**

Table 4.3: **Triangular test**

N	Obtained		Stopping boundary			
			Upper		Lower	
	z-value	p-value	z-value	p-value	z-value	p-value
1	+0.56874	0.28477	+2.95100	0.00158	−0.98111	0.16327
2	+2.04979	0.02019	+2.42652	0.00762	+0.37193	0.35497
3	+1.76041	0.03917	+2.26268	0.01183	+1.17727	0.11954
4	+0.60784	0.27165	+2.20930	0.01358	+1.76454	0.03882

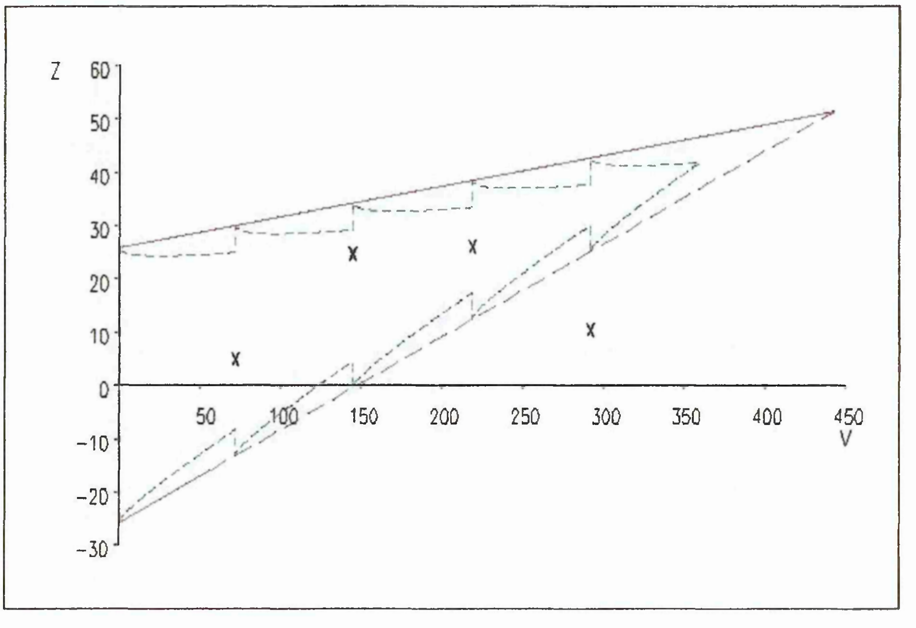

Figure 4.2: **Triangular test**

Table 4.4: **Restricted procedure**

N	Obtained		Stopping boundary	
	z-value	p-value	z-value	p-value
1	+0.56874	0.28477	±4.00620	0.00003
2	+2.04979	0.02019	±2.81635	0.00243
3	+1.76041	0.03917	±2.28897	0.01104
4	+0.60784	0.27165	±1.98305	0.02368

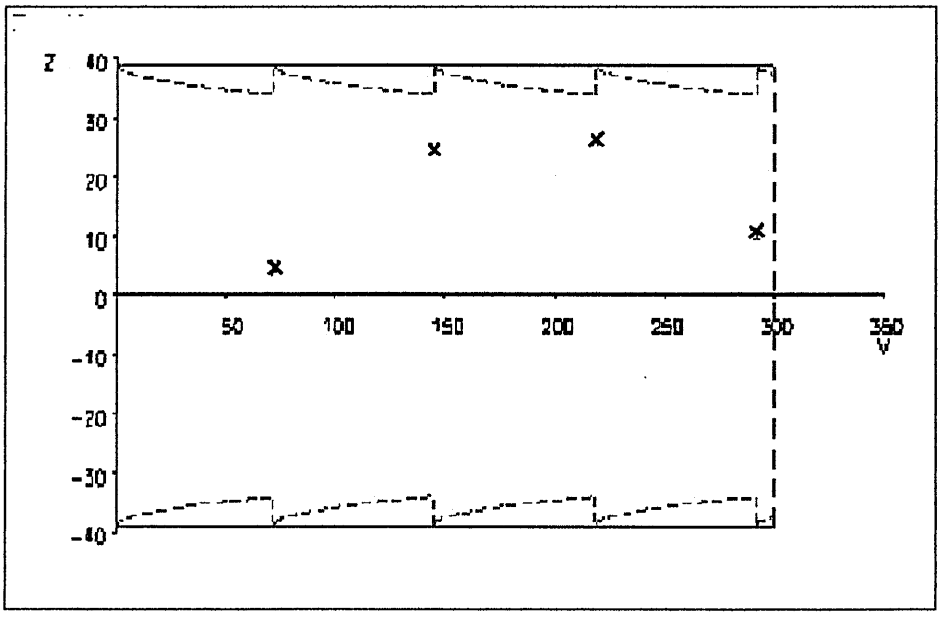

Figure 4.3: **Restricted procedure**

Table 4.5: **Hazard ratios (95% CIs) at study closure**

Method	**Analysis number**	**HR (95% CI)**
SSD	-	0.965 (0.860-1.083)
Naïve	2	0.839 (0.709-0.993)
OBF	4	0.965 (0.895-1.081)
Triangular test	4	0.946 (0.847-1.081)
Restricted procedure	4	0.965 (0.860-1.082)

For the restricted procedure the median unbiased estimate of the HR and its 95% CI have been calculated

4.1.1.2 Bayesian approach

At two-year the overall survival for the experimental group was assumed to be $surv_e = 0.57$, whereas for the control group $surv_c = 0.50$. Using the notation of sections 1.2.3 and 3.1.1, it follows that:

$$\ln(HR_1) = \ln[\ln(0.57)/\ln(0.50)] = -0.210$$

corresponding to a HR of 0.81.

Therefore $\sigma_{scep} = |\ln(HR_1)/1.6445| = 0.1281$ e $N_p = 4.5/0.1281^2 = 274$ subjects. In order to calculate N_p, it has been taken into account the unbalancement in the randomization ratio (1 experimental : 2 control), using the formula $N_p = \left(\frac{(r+1)^2/r}{\sigma^2_{scep}}\right)$, where r is the randomization ratio.

The following prior, likelihood and posterior distributions have been calculated:

- Uninformative prior *expressed as* $N[\ln(1), \infty]$;

- Sceptical prior $\sim N[\ln(1), 0.016]$;

- Enthusiastic prior $\sim N[\ln(0.81), 0.016]$.

Table 4.6: **Likelihood and posterior distributions**

N	Likelihood	Uninformative	Sceptical	Enthusiastic
1	N[ln(0.934), .014]	N[ln(0.934), .014]	N[ln(0.964), .008]	N[ln(0.875), .008]
2	N[ln(0.839), .007]	N[ln(0.839), .007]	N[ln(0.884), .005]	N[ln(0.830), .005]
3	N[ln(0.886), .005]	N[ln(0.886), .005]	N[ln(0.910), .004]	N[ln(0.869), .004]
4	N[ln(0.965), .003]	N[ln(0.965), .003]	N[ln(0.971), .003]	N[ln(0.936), .003]

Table 4.7: **Probabilities of improvement greater than:**

Target improvement	**Hazard Ratio**	**N**	**Uninformative**	**Sceptical**	**Enthusiastic**
0.00	1.00	1	0.719	0.665	0.938
		2	0.982	0.961	0.996
		3	0.962	0.941	0.990
		4	0.727	0.708	0.892
0.025	0.93	1	0.484	0.339	0.758
		2	0.890	0.763	0.947
		3	0.759	0.638	0.870
		4	0.264	0.208	0.451
0.05	0.86	1	0.250	0.101	0.435
		2	0.629	0.361	0.707
		3	0.347	0.187	0.453
		4	0.029	0.014	0.064
0.07	0.81	1	0.116	0.023	0.191
		2	0.342	0.109	0.369
		3	0.098	0.028	0.127
		4	0.002	0.000	0.004

Table 4.8: **Hazard ratios and their 95% credibility intervals**

N	Uninformative	Sceptical	Enthusiastic
1	0.934 [0.741-1.177]	0.964 [0.813-1.143]	0.875 [0.738-1.037]
2	0.839 [0.712-0.988]	0.884 [0.771-1.014]	0.830 [0.724-0.952]
3	0.886 [0.775-1.013]	0.910 [0.809-1.024]	0.869 [0.772-0.977]
4	0.965 [0.859-1.084]	0.971 [0.874-1.079]	0.936 [0.842-1.040]

4.1.1.3 Simulations

The following values have been used:

- underlying HR: 0.965;

- daily event rate in the control arm: 0.00094887;

- accrual duration time: 3.69884 years;

- total study duration: 7.66872 years.

Table 4.9: **Results of simulation with four analyses**

Methods	Boundary	HR	Analysis 1	Analysis 2	Analysis 3	Analysis 4
Naïve	Upper	Median	1.28	1.19	1.14	1.12
		10-90pct	1.25-1.35	1.17-1.23	1.14-1.17	1.12-1.14
		N	122	50	29	21
	Lower	Median	0.77	0.84	0.87	0.89
		10-90pct	0.72-0.80	0.80-0.85	0.85-0.88	0.87-0.89
		N	498	377	344	279
	Not crossing	N				8280
OBF	Upper	Median	-	1.30	1.18	1.13
		10-90pct	-	1.29-1.31	1.17-1.21	1.12-1.15
		N	-	2	15	40
	Lower	Median	-	0.78	0.84	0.88
		10-90pct	-	0.74-0.79	0.82-0.86	0.86-0.89
		N	-	59	304	528
	Not crossing	N				9052
Triangular test	Upper	Median	1.16	1.01	0.96	0.93
		10-90pct	1.12-1.26	0.98-1.08	0.93-1.00	0.91-0.96
		N	998	3852	2699	1226
	Lower	Median	0.70	0.81	0.85	0.87
		10-90pct	0.66-0.72	0.77-0.82	0.83-0.86	0.86-0.88
		N	44	245	243	221
	Not crossing	N				472
Restricted procedure	Upper	Median	-	1.29	1.18	1.13
		10-90pct	-	1.26-1.33	1.16-1.23	1.12-1.15
		N	-	4	17	52
	Lower	Median	-	0.79	0.85	0.88
		10-90pct	-	0.76-0.80	0.82-0.86	0.86-0.89
		N	-	102	302	584
	Not crossing	N				8939

Table 4.10: **Results of simulation with eight analyses**

Methods	Boundary	HR	Analysis 1	2	3	4	5	6	7	8
Naïve	Upper	Median	1.43	1.28	1.22	1.19	1.16	1.14	1.13	1.12
		10-90pct	1.38-1.54	1.25-1.35	1.20-1.27	1.17-1.22	1.15-1.18	1.14-1.16	1.13-1.15	1.12-1.16
		N	184	91	46	36	29	22	11	6
	Lower	Median	0.70	0.78	0.82	0.84	0.86	0.87	0.88	0.89
		10-90pct	0.64-0.73	0.74-0.80	0.79-0.83	0.82-0.85	0.84-0.87	0.85-0.88	0.87-0.89	0.87-0.90
		N	422	358	259	246	215	162	175	162
	Not crossing	N								7576
OBF	Upper	Median	-	-	1.40	1.32	1.23	1.17	1.15	1.13
		10-90pct	-	-	1.38-1.42	1.32-1.32	1.21-1.28	1.17-1.18	1.14-1.16	1.13-1.15
		N	-	-	2	1	5	9	12	12
	Lower	Median	-	0.57	0.72	0.78	0.81	0.84	0.87	0.88
		10-90pct	-	0.57-0.57	0.70-0.73	0.74-0.79	0.79-0.83	0.83-0.86	0.85-0.87	0.87-0.89
		N	-	1	5	36	119	186	247	329
	Not crossing	N								9036
Triangular test	Upper	Median	1.56	1.18	1.07	1.01	0.97	0.95	0.93	0.92
		10-90pct	1.53-1.65	1.14-1.28	1.04-1.12	0.99-1.05	0.96-1.01	0.94-0.98	0.92-0.96	0.91-0.94
		N	25	688	1773	1933	1652	1239	852	534
	Lower	Median	0.53	0.70	0.76	0.80	0.83	0.85	0.86	0.87
		10-90pct	0.53-0.53	0.66-0.70	0.72-0.77	0.78-0.82	0.81-0.84	0.83-0.86	0.85-0.87	0.86-0.88
		N	1	25	88	141	143	129	148	110
	Not crossing	N								519
Restricted procedure	Upper	Median	-	1.62	1.41	1.27	1.23	1.18	1.15	1.13
		10-90pct	-	1.62-1.62	1.41-1.41	1.26-1.28	1.20-1.27	1.17-1.20	1.14-1.17	1.12-1.16
		N	-	1	1	2	9	12	15	21
	Lower	Median	-	-	0.71	0.78	0.82	0.85	0.87	0.88
		10-90pct	-	-	0.68-0.72	0.75-0.79	0.79-0.83	0.83-0.86	0.85-0.87	0.86-0.89
		N	-	-	8	49	127	179	220	274
	Not crossing	N								9082

4.1.2 ICON 4/AGO-OVAR 2.2 Trial

Tables 4.11, 4.12, 4.13 and 4.14 report the results of naïve method, alpha-spending function with OBF boundaries, triangular test, and the restricted procedure, respectively. Plot of results of these analyses are also reported in Figures 4.4, 4.5, 4.6.

Table 4.15 summarises the estimates (HRs and 95% CIs) obtained at study closure, considered as the first crossed boundary or the last scheduled analysis, according to the results of each considered method.

Tables 4.16, 4.17 and 4.18 report the results obtained adopting Bayesian approach.

The naïve method would have conducted to early stopping at the second analysis (HR=0.74; 95% CI: 0.58-0.94; $p<0.01$, one sided), with an absolute overestimate of 9%.

The other methods would have conducted to anticipated closure at the third analysis, with an overestimate of 6.3% for alpha spending function approach, and of 4.5% for the other approaches. Bayesian approach showed that under the sceptical prior the overestimate at second look is 4%. The probability of an effect size larger than null is 98% at the second look, the probability of an effect size larger than 2.5% is 92%, while the probability of obtaining at least 5% reduction in mortality is 76%, suggesting that the benefit may be relevant. At the end of the study, the probability of some effect is high even under the sceptical prior (97%), and the chance to have a clinically relevant effect (say at least 5% reduction in mortality) is 54% under the sceptical and 77% under the enthusiastic priors.

Table 4.19 reports the results of the simulation with four looks.

Using the naïve method, the overall chance of early stopping was 63.2%: 62.9% of the *trials* crossed the lower boundary, indicating an effect in favour of experimental arm. Again the larger proportion of *trials* outside the boundaries was observed in the first and second looks (57% of *trials* crossing the lower boundary and 96% of *trials* crossing the upper boundary). The median HR of *trials* with results crossing the lower boundary ranged from 0.66 to 0.82, and on the average, the overestimate was 10.4%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 1.19 to 1.45, with an underestimate of effect equal to 70.3%.

For alpha-spending function the percentage of early stopping was 55.6% (all at the lower boundary) and the chances of early stopping at the third or fourth analyses were 87.4%. The median HR of *trials* with results crossing the lower boundary ranged from 0.53 to 0.87, and on the average, the overestimate was 7.6%.

As for the restricted procedure, the percentage of early stopping was 54.5% (all at the lower boundary) and the chances of early stopping were also higher after the second analysis (84.3%). The median HR of *trials* with results crossing the lower boundary ranged from 0.49 to 0.81, and on the average, the overestimate was 7.8%.

The triangular test produced 86.6% of early interruptions (38.3% at upper boundary and only 48.3% at the lower boundary). The higher percentage of interruptions (66.8% of *trials* crossing the lower boundary and 57.0% of *trials* crossing the upper boundary) occurred at the second and third analyses. The median HR of *trials* with results crossing the lower boundary ranged from 0.57 to 0.80, and on the average, the overestimate was 10.1%. For *trials* stopped because results crossed the upper boundary, the median HR ranged from 0.88 to 1.24, with an underestimate of effect equal to 14.1%.

Table 4.20 reports the results of the simulation using 8 looks.

With the naïve method, the overall chance of early stopping was 63.3%: of those 99.0% at the lower boundary. Again, the larger proportion of *trials* outside the boundary was observed in the first 4 looks (61.8% of *trials* crossing the lower boundary and 98.5% of *trials* crossing the upper boundary). The median HR of *trials* with results crossing the lower boundary ranged from 0.56 to 0.83, and on the average, the overestimate was 13.1%. For *trials* stopped because results crossed the upper boundary, the median HR ranged from 1.20 to 1.72, with an underestimate of effect equal to 97.7%.

For alpha-spending function the percentage of early stopping was 53.0% (all at the lower boundary) and the chances of early stopping were higher at the final analyses (88.0% of early interruption for crossing the lower boundary occurred after the fourth analysis). The median HR of *trials* with results crossing the lower boundary ranged from 0.44 to 0.82, and on the average, the overestimate was 8.4%.

For restricted procedure the percentage of early stopping was 52.5% (all at the lower boundary) and the chances of early stopping were higher after the fourth analysis (87.0% of early interruption for crossing the lower boundary). The median HR of *trials* with results crossing the lower boundary decreased from 0.46 to 0.82, and on the average, the overestimate was 8.6%.

With the triangular test 86.5% of *trials* interrupted early (47.1% at the lower - boundary). The higher percentage of interruptions (65.2% of *trials* crossing the lower boundary and 65.4% of *trials* crossing the upper boundary) occurred between the third and sixth analysis. The median HR of *trials* with results crossing the lower boundary ranged from 0.34 to 0.81, and the overestimate was 11.0%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 0.88 to 2.04, with an underestimate of effect equal to 15.7%.

4.1.2.1 Frequentist approaches

Table 4.11: **Analysis without adjustment for multiple tests**

				Obtained		Stopping boundary	
N	Date	Pts	Events	z-value	p-value	z-value	p-value
1	18/05/1999	409	133	+1.84183	0.03272	±1.96000	0.02500
2	09/10/2000	607	264	+2.43969	0.00735	±1.96000	0.02500
3	19/11/2001	772	398	+2.60082	0.00467	±1.96000	0.02500
4	26/03/2003	802	530	+2.10033	0.01784	±1.96000	0.02500

Table 4.12: **Alpha-spending function with O'Brien and Fleming boundaries**

N	Obtained		Stopping boundary	
	z-value	p-value	z-value	p-value
1	+1.84183	0.03272	±4.33263	0.00001
2	+2.43969	0.00735	±2.96311	0.00152
3	+2.60082	0.00467	±2.35902	0.00916
4	+2.10033	0.01784	±2.01406	0.02200

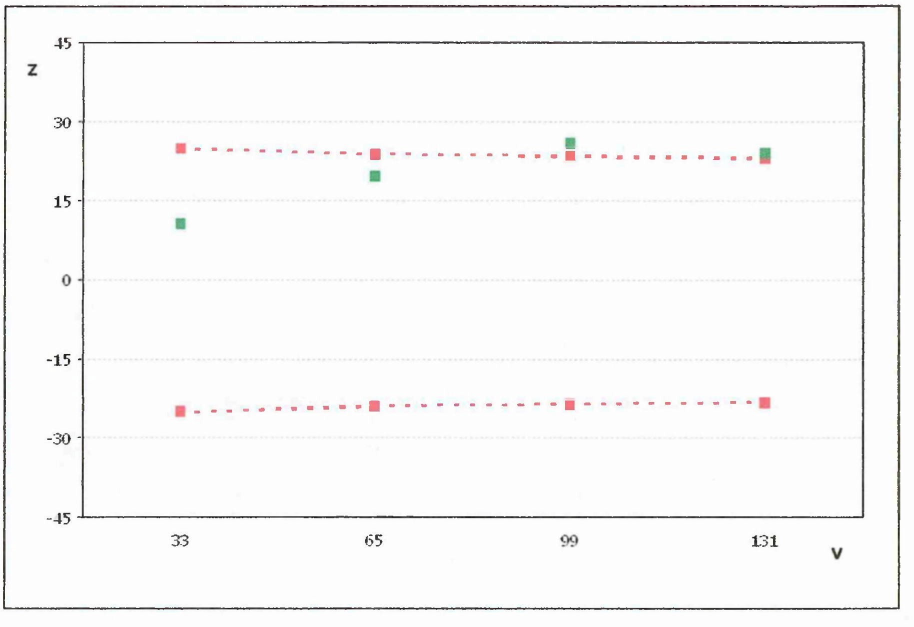

Figure 4.4: **Alpha-spending function with O'Brien and Fleming boundaries**

Table 4.13: **Triangular test**

			Stopping boundary			
	Obtained		Upper		Lower	
N	z-value	p-value	z-value	p-value	z-value	p-value
1	+1.84873	0.03225	+2.96346	0.00152	−1.00314	0.15790
2	+2.45129	0.00712	+2.45536	0.00704	+0.29749	0.38305
3	+2.60389	0.00461	+2.27815	0.01136	+1.10438	0.13471
4	+2.09855	0.01793	+2.22153	0.01316	+1.67725	0.04675

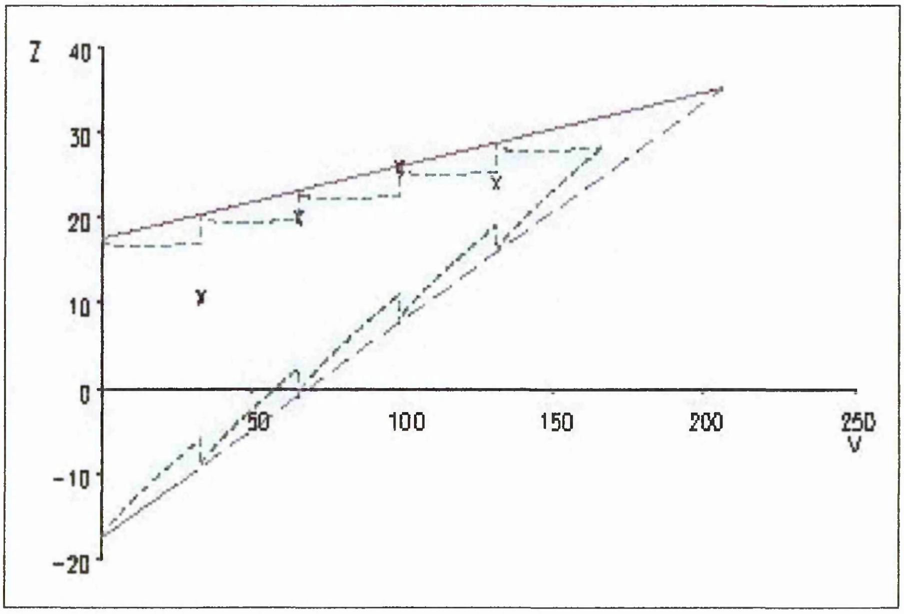

Figure 4.5: **Triangular test**

Table 4.14: **Restricted procedure**

N	Obtained		Stopping boundary	
	z-value	**p-value**	**z-value**	**p-value**
1	+1.84873	0.03225	±4.00908	0.00003
2	+2.45129	0.00712	±2.86073	0.00211
3	+2.60389	0.00461	±2.32252	0.01010
4	+2.09855	0.01793	±2.01899	0.02174

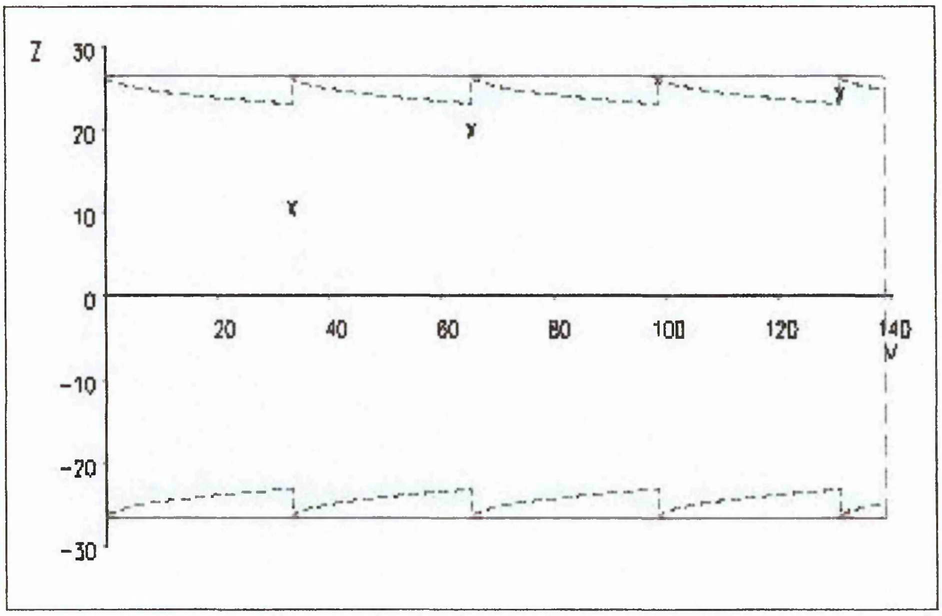

Figure 4.6: **Restricted procedure**

Table 4.15: **Hazard ratios (95% CIs) at study closure**

Methods	**Analysis number**	**HR (95% CI)**
SSD	-	0.833 (0.702-0.988)
Naïve	2	0.739 (0.579-0.942)
OBF	3	0.770 (0.634-0.942)
Triangular test	3	0.788 (0.628-0.968)
Restricted procedure	3	0.788 (0.637-0.952)

4.1.2.2 Bayesian approach

At two-year the overall survival for the experimental group was assumed to be $surv_e = 0.61$, whereas for the control group $surv_c = 0.50$. It follows that:

$$\ln(HR_1) = \ln[\ln(0.61)/\ln(0.50)] = -0.342$$

corresponding to a HR of 0.71.

Therefore $\sigma_{scep} = |\ln(HR_1)/1.6445| = 0.208$ e $N_p = 4/0.208^2 = 92$ subjects. The following prior, likelihood and posterior distributions have been calculated:

- Uninformative prior *expressed as* $N[\ln(1), \infty]$;
- Sceptical prior $\sim N[\ln(1), 0.043]$;
- Enthusiastic prior $\sim N[\ln(0.71), 0.043]$.

Table 4.16: **Likelihood and posterior distributions**

N	Likelihood	Uninformative	Sceptical	Enthusiastic
1	N[ln(0.723), .030]	N[ln(0.723), .030]	N[ln(0.826), .018]	N[ln(0.718), .018]
2	N[ln(0.739), .015]	N[ln(0.739), .015]	N[ln(0.799), .011]	N[ln(0.731), .011]
3	N[ln(0.769), .010]	N[ln(0.769), .010]	N[ln(0.808), .008]	N[ln(0.758), .008]
4	N[ln(0.833), .008]	N[ln(0.833), .008]	N[ln(0.856), .006]	N[ln(0.814), .006]

Table 4.17: **Probabilities of improvement greater than:**

Target improvement	Hazard Ratio	N	Uninformative	Sceptical	Enthusiastic
0.00	1.00	1	0.969	0.925	0.994
		2	0.993	0.983	0.998
		3	0.996	0.991	0.999
		4	0.982	0.974	0.995
0.025	0.93	1	0.926	0.813	0.974
		2	0.969	0.923	0.988
		3	0.971	0.940	0.988
		4	0.897	0.849	0.952
0.05	0.86	1	0.845	0.628	0.916
		2	0.895	0.764	0.940
		3	0.874	0.765	0.925
		4	0.656	0.538	0.767
0.11	0.71	1	0.468	0.136	0.481
		2	0.386	0.141	0.406
		3	0.226	0.083	0.252
		4	0.037	0.011	0.050

Table 4.18: **Hazard ratios and their 95% credibility intervals**

N	Uninformative	Sceptical	Enthusiastic
1	**0.723 [0.515-1.016]**	**0.826 [0.636-1.072]**	0.718 [0.553-0.932]
2	0.739 [0.581-0.941]	0.799 [0.649-0.984]	0.731 [0.594-0.900]
3	0.769 [0.632-0.936]	0.808 [0.677-0.964]	0.758 [0.635-0.904]
4	0.833 [0.703-0.988]	**0.856 [0.731-1.002]**	0.814 [0.695-0.952]

4.1.2.3 Simulations

The following values have been used:

- underlying HR: 0.833;
- daily event rate in the control arm: 0.000949;
- accrual duration time: 6.19576 years;
- total study duration: 6.85558 years.

Table 4.19: **Results of simulation with four analyses**

Methods	**Boundary**	**HR**	**Analysis**			
			1	**2**	**3**	**4**
Naïve	Upper	Median	1.45	1.31	-	1.19
		10-90pct	1.42-1.53	1.28-1.34	-	1.19-1.19
		N	21	4	-	1
	Lower	Median	0.66	0.75	0.79	0.82
		10-90pct	0.58-0.70	0.69-0.78	0.75-0.82	0.79-0.84
		N	1875	1692	1546	1177
	Not crossing	N				3684
OBF	Lower	Median	0.46	0.66	0.76	0.81
		10-90pct	0.40-0.46	0.61-0.69	0.71-0.78	0.77-0.83
		N	4	697	2349	2515
	Not crossing	N				4435
Triangular test	Upper	Median	1.24	1.00	0.92	0.88
		10-90pct	1.20-1.40	0.97-1.08	0.90-0.98	0.87-0.92
		N	222	1142	1333	1136
	Lower	Median	0.57	0.71	0.77	0.80
		10-90pct	0.52-0.59	0.65-0.73	0.73-0.79	0.77-0.82
		N	263	1494	1731	1337
	Not crossing	N				1342
Restricted procedure	Lower	Median	0.49	0.68	0.76	0.81
		10-90pct	0.46-0.49	0.63-0.70	0.71-0.79	0.77-0.83
		N	19	838	2302	2294
	Not crossing	N				4547

Table 4.20: **Results of simulation with eight analyses**

Methods	Boundary	HR	Analysis 1	2	3	4	5	6	7	8
Naïve	Upper	Median	1.72	1.44	1.35	1.27	-	-	1.20	-
		10-90pct	1.63-1.90	1.42-1.47	1.35-1.45	1.27-1.27	-	-	1.20-1.20	-
		N	52	10	3	1	-	-	1	-
	Lower	Median	0.56	0.67	0.73	0.76	0.79	0.80	0.82	0.83
		10-90pct	0.46-0.61	0.61-0.70	0.68-0.75	0.72-0.78	0.75-0.80	0.78-0.82	0.79-0.83	0.81-0.84
		N	1166	1102	984	865	785	651	616	492
	Not crossing	N								3272
OBF	Lower	Median	-	0.44	0.59	0.67	0.73	0.77	0.79	0.82
		10-90pct	-	0.39-0.47	0.54-0.61	0.62-0.69	0.69-0.74	0.73-0.78	0.77-0.81	0.79-0.83
		N	-	4	149	483	1058	1197	1296	1109
	Not crossing	N								4704
Triangular test	Upper	Median	2.04	1.28	1.09	1.00	0.95	0.92	0.90	0.88
		10-90pct	1.96-2.41	1.23-1.43	1.05-1.17	0.97-1.06	0.93-0.99	0.90-0.95	0.88-0.93	0.87-0.90
		N	7	144	435	660	744	727	679	539
	Lower	Median	0.34	0.55	0.65	0.71	0.75	0.77	0.79	0.81
		10-90pct	0.27-0.36	0.51-0.58	0.60-0.67	0.67-0.73	0.71-0.76	0.74-0.79	0.77-0.81	0.78-0.82
		N	7	199	531	771	959	819	824	602
	Not crossing	N								1353
Restricted procedure	Lower	Median	-	0.46	0.59	0.67	0.73	0.77	0.79	0.82
		10-90pct	-	0.45-0.48	0.55-0.61	0.63-0.69	0.70-0.75	0.73-0.78	0.77-0.81	0.79-0.83
		N	-	7	165	510	1081	1145	1218	1122
	Not crossing	N								4752

4.1.3 GIVIO-SITAC 01 Trial

Tables 4.21, 4.22, 4.23 and 4.24 report the results of naïve method, alpha-spending function with OBF boundaries, triangular test, and the restricted procedure, respectively. Plot of results of these analyses are also reported in Figures 4.7, 4.8, 4.9.

Table 4.25 summarises the estimates (HRs and 95% CIs) obtained at study closure, considered as the first crossed boundary or the last scheduled analysis, according to the results of each considered method.

All these methods would have allowed to reach a statistically significant result at the fourth analysis.

Tables 4.26, 4.27 and 4.28 report the results obtained adopting Bayesian approach.

Bayesian approach showed that under the sceptical prior, the probability of an effect size larger than zero is 98% at the fourth look, the probability of an effect size larger than 2.5% is 87%, while the probability to obtain at least 5% reduction in mortality is 53%, supporting the evidence of a moderate effect.

Table 4.29 reports the results of the simulation with four looks.

Using the naïve method, the overall chance of early stopping was 74.0%: 73.9% crossed the lower boundary. The larger proportion of *trials* outside the boundaries was observed in the first and second look (61% of *trials* crossing the lower boundary and 100% of *trials* crossing the upper boundary occurred in these two analyses). The median HR of *trials* with results crossing the lower boundary ranged from 0.56 to 0.76, and on the average, the overestimate was 11.2%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 1.44 to 1.71, with an underestimate of effect equal to 125.3%.

For alpha-spending function the percentage of early stopping was 66.5% (all occurring at the lower boundary) and the chances of early stopping at the third or fourth analyses were 82.9%. The median HR of *trials* with results crossing the lower boundary ranged from 0.33 to 0.75, and on the average, the overestimate was 7.9%.

For the restricted procedure, the percentage of early stopping was 66% (all at the lower boundary) and the chances of early stopping were also higher at the third and fourth analysis: 79.7%. The median HR of *trials* with results crossing the lower boundary ranged from 0.36 to 0.75, and on the average, the overestimate was 8.4%.

With the triangular test 87.6% of early interruptions would have been observed (27.3% at upper boundary and 60.3% at the lower boundary). The higher percentage of interruptions (63.5% of *trials* crossing the lower boundary and 67.4% of *trials* crossing the upper boundary) occurred at the second and third analyses. The median HR of *trials* with results crossing the lower boundary ranged from 0.45 to 0.74, and on the average, the overestimate was 11.6%. For *trials* stopped because results crossed the upper boundary, the median HR ranged from 0.84 to 1.35, with an underestimate of effect equal to 23.2%.

Table 4.30 reports the results of the simulation using eight looks.

With the naïve method, the overall chance of early stopping is 75.9%: 75.6% crossed the lower boundary. The larger proportion of *trials* outside the boundaries was observed in the first 4 looks (65.4% of *trials* crossing the lower boundary and 100% of *trials* crossing the upper boundary). The median HR of *trials* with results crossing the lower boundary ranged from 0.45 to 0.78, and on the average, the overestimate was 14.7%. For *trials* stopped because the upper boundary was crossed, the median HR ranged from 1.64 to 2.02, with an underestimate of effect equal to 164.4%.

For alpha-spending function the percentage of early stopping was 65.7% (all at the lower boundary) and the chances of early stopping were higher after the fourth analysis (82.5% of early interruptions). The median HR of *trials* with results crossing the lower boundary ranged from 0.34 to 0.76, and on the average, the overestimate was 9.6%.

For the restricted procedure, the percentage of early stopping was 64.8% (only one trial crossed the upper boundary) and the chances of early stopping were also higher at the final analyses (81.8% of early interruption for crossing the lower boundary occurred after the fourth analysis). The median HR of *trials* with results crossing the lower boundary ranged from 0.35 to 0.76, and on the average, the overestimate was 9.5%. The trial who stopped at the upper boundary had a median HR equal to 1.65, with an underestimate of 120.9%.

With the triangular test 89.1% of *trials* interrupted early (60.9% at upper boundary and 28.2% at the lower boundary). The higher percentage of interruptions (68.4% of *trials* crossing the lower boundary and 69.9% of *trials* crossing the upper boundary) occurred between the fourth and seventh analyses. The median HR of *trials* with results crossing the lower boundary ranged from 0.20 to 0.75, and on the average, the overestimate was 12.7%. For *trials* stopped because results crossed the upper boundary, the median HR ranged from 0.83 to 2.86, with an underestimate of effect equal to 46.1%.

4.1.3.1 Frequentist approaches

Table 4.21: **Analysis without adjustment for multiple tests**

				Obtained		**Stopping boundary**	
N	**Date**	**Pts**	**Events**	**z-value**	**p-value**	**z-value**	**p-value**
1	15/11/1991	840	70	+1.59104	0.05582	±1.96000	0.02500
2	31/12/1992	869	140	+1.80292	0.03572	±1.96000	0.02500
3	24/01/1994	869	210	+1.75359	0.03976	±1.96000	0.02500
4	04/09/1996	869	279	+2.40664	0.00804	±1.96000	0.02500

Table 4.22: **Alpha-spending function with O'Brien and Fleming boundaries**

N	Obtained		Stopping boundary	
	z-value	**p-value**	**z-value**	**p-value**
1	+1.59104	0.05582	±4.33263	0.00001
2	+1.80292	0.03571	±2.96311	0.00152
3	+1.75359	0.03976	±2.35902	0.00916
4	+2.40664	0.03976	±2.01406	0.02200

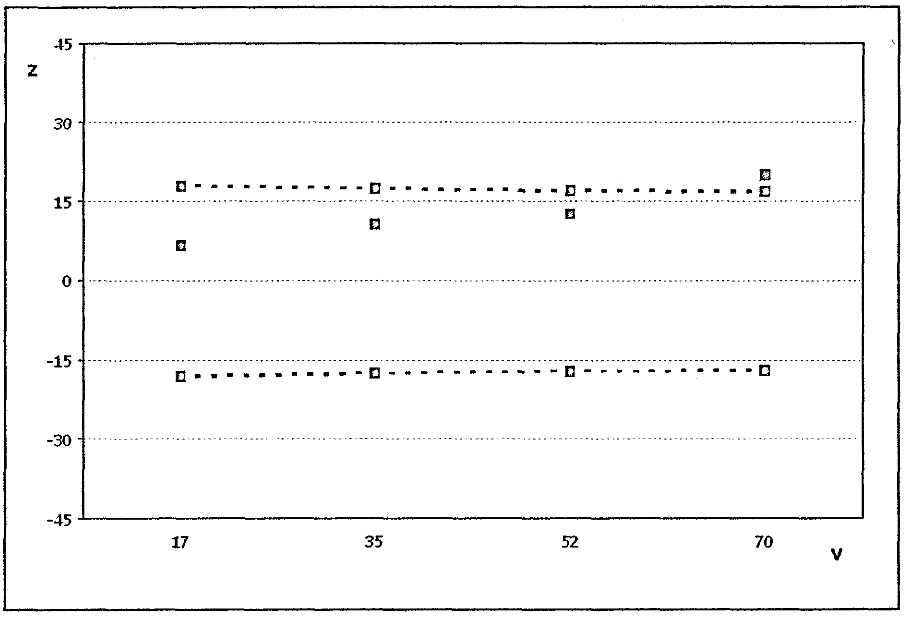

Figure 4.7: **Alpha-spending function with O'Brien and Fleming boundaries**

Table 4.23: **Triangular test**

N	Obtained		Stopping boundary			
			Upper		Lower	
	z-value	p-value	z-value	p-value	z-value	p-value
1	+1.60044	0.05475	+2.95701	0.00155	−0.99176	0.16066
2	+1.80926	0.03521	+2.43572	0.00743	+0.34745	0.36413
3	+1.75761	0.03941	+2.27292	0.01152	+1.13655	0.12786
4	+2.41566	0.00785	+2.21703	0.01331	+1.71313	0.04334

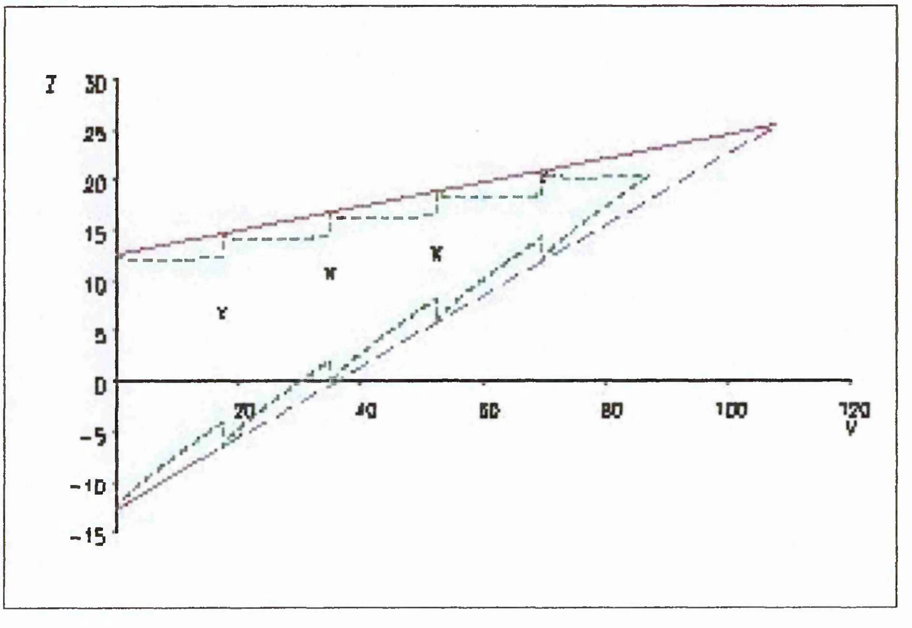

Figure 4.8: **Triangular test**

Table 4.24: **Restricted procedure**

N	Obtained		Stopping boundary	
	z-value	p-value	z-value	p-value
1	+1.60044	0.05475	±4.03311	0.00003
2	+1.80924	0.03521	±2.84670	0.00221
3	+1.75511	0.03962	±2.32402	0.01006
4	+2.18473	0.01446	±2.01827	0.02178

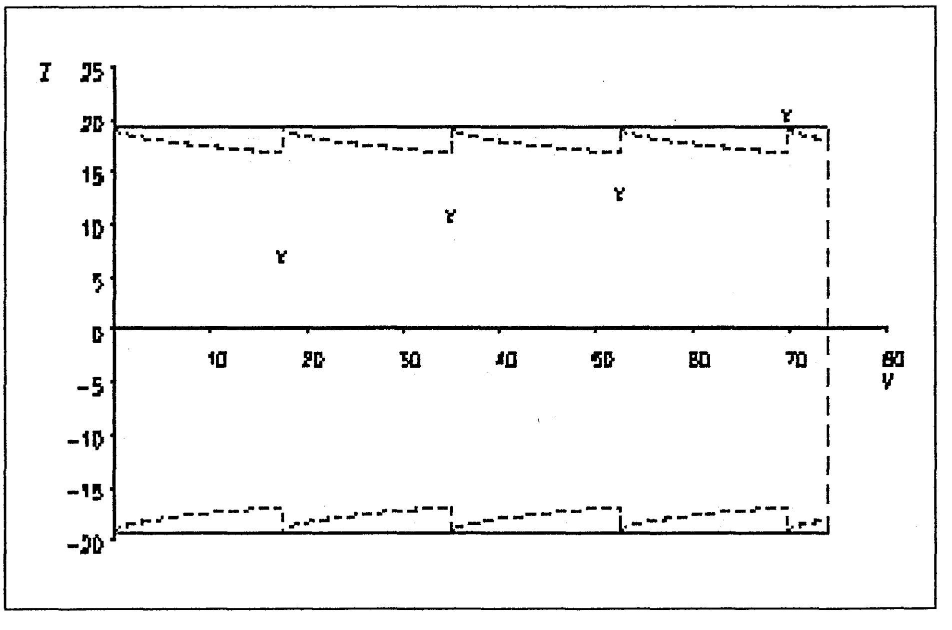

Figure 4.9: **Restricted procedure**

Table 4.25: **Hazard ratios (95% CIs) at study closure**

Methods	**Analysis number**	**HR (95% CI)**
SSD	-	0.747 (0.590-0.947)
Naïve	4	0.747 (0.590-0.947)
OBF	4	0.747 (0.596-0.969)
Triangular test	4	0.760 (0.575-0.990)
Restricted procedure	4	0.768 (0.593-0.968)

4.1.3.2 Bayesian approach

After five years the overall survival for the experimental group was assumed to be $surv_e = 0.78$, whereas for the control group $surv_c = 0.70$. It follows that:

$$\ln(HR_1) = \ln[\ln(0.70)/\ln(0.78)] = -0.355$$

corresponding to a HR of 0.70.

Therefore $\sigma_{scep} = |\ln(HR_1)/1.6445| = 0.216$ e $N_p = 4/(0.208)^2 = 86$ subjects. The following prior, likelihood and posterior distributions have been calculated:

- Uninformative prior *expressed as* $N[\ln(1), \infty]$;
- Sceptical prior $\sim N[\ln(1), 0.047]$;
- Enthusiastic prior $\sim N[\ln(0.70), 0.047]$.

Table 4.26: **Likelihood and posterior distributions**

N	Likelihood	Uninformative	Sceptical	Enthusiastic
1	N[ln(0.679), .057]	N[ln(0.679), .057]	N[ln(.840), .026]	N[ln(.691), .026]
2	N[ln(0.735), .029]	N[ln(0.735), .029]	N[ln(.826), .018]	N[ln(.722), .018]
3	N[ln(0.784), .019]	N[ln(0.784), .019]	N[ln(.841), .014]	N[ln(.759), .014]
4	N[ln(0.747), .014]	N[ln(0.747), .014]	N[ln(.800), .011]	N[ln(.736), .011]

Table 4.27: **Probabilities of improvement greater than:**

Target improvement	Hazard Ratio	N	Uninformative	Sceptical	Enthusiastic
0.00	1.00	1	0.947	0.861	0.989
		2	0.966	0.924	0.993
		3	0.961	0.931	0.991
		4	0.993	0.984	0.998
0.025	0.90	1	0.882	0.670	0.952
		2	0.887	0.744	0.952
		3	0.844	0.724	0.931
		4	0.942	0.874	0.974
0.05	0.81	1	0.764	0.399	0.833
		2	0.709	0.428	0.798
		3	0.581	0.358	0.699
		4	0.740	0.532	0.809
0.08	0.70	1	0.543	0.121	0.520
		2	0.375	0.100	0.394
		3	0.196	0.052	0.230
		4	0.279	0.093	0.300

Table 4.28: **Hazard ratios and their 95% credibility intervals**

N	Uninformative	Sceptical	Enthusiastic
1	0.679 [0.425-1.085]	0.840 [0.614-1.150]	0.691 [0.505-0.946]
2	0.735 [0.528-1.024]	0.826 [0.636-1.072]	0.722 [0.556-0.937]
3	0.784 [0.598-1.028]	0.841 [0.670-1.057]	0.759 [0.604-0.953]
4	0.747 [0.591-0.944]	0.800 [0.652-0.982]	0.736 [0.600-0.903]

4.1.3.3 Simulations

The following values where used:

- underlying HR: 0.747;
- daily event rate in the control arm: 0.0001953;
- accrual duration time: 2.99247 years;
- total study duration: 7.43874 years.

Table 4.29: **Results of simulation with four analyses**

Methods	Boundary	HR	Analysis 1	2	3	4
Naïve	Upper	Median	1.71	1.44	-	-
		10-90pct	1.61-1.96	1.44-1.44	-	-
		N	9	1	-	-
	Lower	Median	0.56	0.67	0.73	0.76
		10-90pct	0.46-0.61	0.59-0.71	0.67-0.76	0.72-0.79
		N	2309	2229	1702	1152
	Not crossing	N				2598
OBF	Lower	Median	0.33	0.57	0.68	0.75
		10-90pct	0.29-0.35	0.50-0.60	0.62-0.71	0.70-0.78
		N	23	1116	2911	2605
	Not crossing	N				3345
Triangular test	Upper	Median	1.35	0.99	0.89	0.84
		10-90pct	1.28-1.54	0.95-1.09	0.86-0.95	0.82-0.89
		N	117	723	1008	878
	Lower	Median	0.45	0.62	0.69	0.74
		10-90pct	0.39-0.49	0.55-0.65	0.64-0.72	0.69-0.76
		N	486	1962	2104	1483
	Not crossing	N				1239
Restricted procedure	Lower	Median	0.36	0.58	0.68	0.75
		10-90pct	0.32-0.38	0.51-0.61	0.62-0.72	0.70-0.78
		N	33	1309	2834	2420
	Not crossing	N				3404

Table 4.30: **Results of simulation with eight analyses**

Methods	Boundary	HR	Analysis 1	2	3	4	5	6	7	8
Naïve	Upper	Median	2.02	1.64	-	-	-	-	-	-
		10-90pct	1.95-2.31	1.61-1.77	-	-	-	-	-	-
		N	23	9	-	-	-	-	-	-
	Lower	Median	0.45	0.58	0.64	0.69	0.72	0.74	0.76	0.78
		10-90pct	0.34-0.50	0.49-0.62	0.58-0.68	0.63-0.71	0.67-0.74	0.70-0.76	0.72-0.77	0.75-0.79
		N	1406	1331	1153	1056	860	731	600	423
	Not crossing	N								2408
OBF	Lower	Median	-	0.34	0.48	0.57	0.64	0.69	0.73	0.76
		10-90pct	-	0.31-0.35	0.43-0.50	0.52-0.60	0.59-0.66	0.65-0.71	0.69-0.75	0.73-0.78
		N	-	15	268	868	1323	1545	1431	1123
	Not crossing	N								3427
Triangular test	Upper	Median	2.86	1.40	1.13	1.00	0.93	0.89	0.86	0.83
		10-90pct	2.83-2.89	1.32-1.58	1.08-1.24	0.97-1.07	0.91-0.99	0.87-0.93	0.84-0.90	0.82-0.86
		N	2	85	264	424	508	576	460	497
	Lower	Median	0.20	0.44	0.55	0.62	0.66	0.70	0.72	0.75
		10-90pct	0.16-0.23	0.38-0.47	0.49-0.58	0.57-0.64	0.62-0.68	0.66-0.72	0.69-0.74	0.72-0.76
		N	8	353	798	1100	1077	1075	914	764
	Not crossing	N								1095
Restricted procedure	Upper	Median	-	-	-	1.65	-	-	-	-
		10-90pct	-	-	-	1.65-1.65	-	-	-	-
		N	-	-	-	1	-	-	-	-
	Lower	Median	-	0.35	0.48	0.58	0.64	0.69	0.73	0.76
		10-90pct	-	0.29-0.36	0.43-0.51	0.53-0.60	0.60-0.66	0.65-0.71	0.69-0.75	0.73-0.77
		N	-	24	264	891	1268	1492	1454	1089
	Not crossing	N								3517

4.1.4 Summary of results

Table 4.31 reports the summary of the main results of the simulations: the more used statistical approaches reduce the risk of "incorrect" early stopping, compared with the adoption of no stopping rule.

Figures 4.10, 4.11, 4.12, show the estimated HRs and the relative unadjusted 95% CIs as a function of the percent of the total number of events observed during the course of the three studies considered: as expected the fluctuation is very high at the beginning of the study, but estimates are more stable after about half of information fraction is achieved.

It is therefore important to stress the fact that alpha-spending function and restricted procedures allow early termination at the end of the study, favouring a reduction in the magnitude of estimation bias.

On the contrary naïve method is more prone to lead to a termination at early stages, with a higher estimation bias.

Triangular test performs well regarding overestimation, but the more relaxed criteria for stopping for futility, increase the study probability of an early termination concluding for no effect difference. Even in "positive" *trials*, the chance of stopping for futility are 39% in ICON 4 and 28.2% in GIVIO-SITAC 01 study.

It is also important to note that the chance of early stopping due to an overestimate is directly related to the "true" magnitude of effect; the inverse is true in case of stopping for futility.

The number of analyses has a moderate impact on estimation, when a specific approach has been adopted, but it is important when no criteria for controlling for multiple analyses are used.

Table 4.31: **Simulation result summary**

Study	**Approach**	**Four analyses** Rate of early stops	Early stop greatest rate	Bias size related to stop	**Eight analyses** Rate of early stops	Early stop greatest rate	Bias size related to stop
ICON 3	Naïve	17.2%	looks 1–2		24.2%	looks 1–2–3–4	
HR 0.965	- overestimate	14.9%	58.0%	13.6%	20.0%	64.3%	16.1%
	- underestimate	2.3%	77.0%	27.0%	4.2%	84.0%	35.7%
	OBF	9.5%	looks 3–4		9.6%	looks 5–6–7–8	
	- overestimate	8.9%	93.4%	10.9%	9.2%	82.6%	11.4%
	- underestimate	0.6%	96.5%	19.0%	0.4%	92.7%	21.7%
	Triangular	95.3%	looks 2–3		94.0%	looks 3–4–5–6	
	- overestimate	7.5%	64.8%	13.6%	7.0%	71.1%	14.3%
	- underestimate	87.8%	74.6%	3.7%	87.0%	75.9%	4.4%
	Restricted	10.6%	looks 3–4		9.1%	looks 5–6–7–8	
	- overestimate	9.9%	89.7%	10.7%	8.6%	93.3%	11.4%
	- underestimate	0.7%	94.5%	19.0%	0.5%	74.5%	21.9%
ICON 4	Naïve	63.2%	looks 1–2		63.3%	looks 1–2–3–4	
HR 0.833	- overestimate	62.9%	57.0%	10.4%	62.7%	61.8%	13.1%
	- underestimate	0.3%	96.0%	70.3%	0.6%	95.2%	97.7%
	OBF	55.6%	looks 3–4		53.0%	looks 5–6–7–8	
	- overestimate	55.6%	87.4%	7.6%	53.0%	88.0%	8.4%
	- underestimate	-	-	-	-	-	-
	Triangular	85.6%	looks 2–3		86.5%	looks 3–4–5–6	
	- overestimate	48.3%	66.8%	10.1%	47.1%	65.2%	11.0%
	- underestimate	38.3%	57.0%	14.1%	39.4%	65.4%	15.7%
	Restricted	54.5%	looks 3–4		52.5%	looks 5–6–7–8	
	- overestimate	54.5%	84.3%	7.3%	52.5%	87.0%	8.6%
	- underestimate	-	-	-	-	-	-

cont.

cont., Table 4.31

Study	Approach	Four analyses: Rate of early stops	Four analyses: Early stop greatest rate	Four analyses: Bias size related to stop	Eight analyses: Rate of early stops	Eight analyses: Early stop greatest rate	Eight analyses: Bias size related to stop
SITAC 01	Naïve	74.0%	looks 1–2		75.9%	looks 1–2–3–4	
HR 0.747	- overestimate	73.9%	61.0%	11.2%	75.6%	65.4%	14.7%
	- underestimate	0.1%	100.0%	125.3%	0.3%	100.0%	164.4%
	OBF	66.5%	looks 3–4		65.7%	looks 5–6–7–8	
	- overestimate	66.5%	82.9%	7.9%	65.7%	82.5%	9.6%
	- underestimate	-	-	-	-	-	-
	Triangular	87.6%	looks 2–3		89.1%	looks 4–5–6–7	
	- overestimate	60.3%	63.5%	11.6%	60.9%	68.4%	12.7%
	- underestimate	27.3%	67.4%	23.2%	28.2%	69.9%	46.1%
	Restricted	66.0%	looks 3–4		64.8%	looks 5–6–7–8	
	- overestimate	66.0%	79.7%	8.4%	64.8%	81.8%	9.5%
	- underestimate	-	-	-	-	-	-

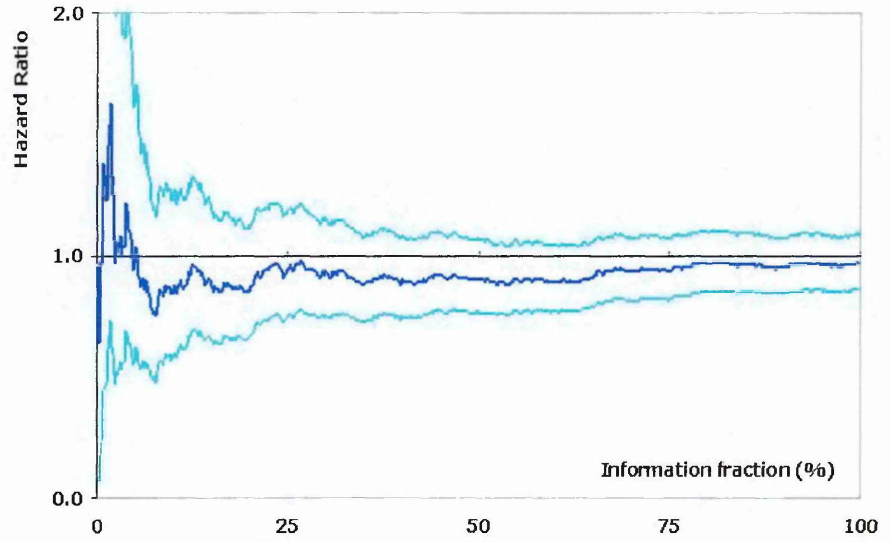

Figure 4.10: **ICON 3 Trial - Trend over time of the hazard ratio estimate (dark blue) and its lower and upper 95% CI bounds (blue)**

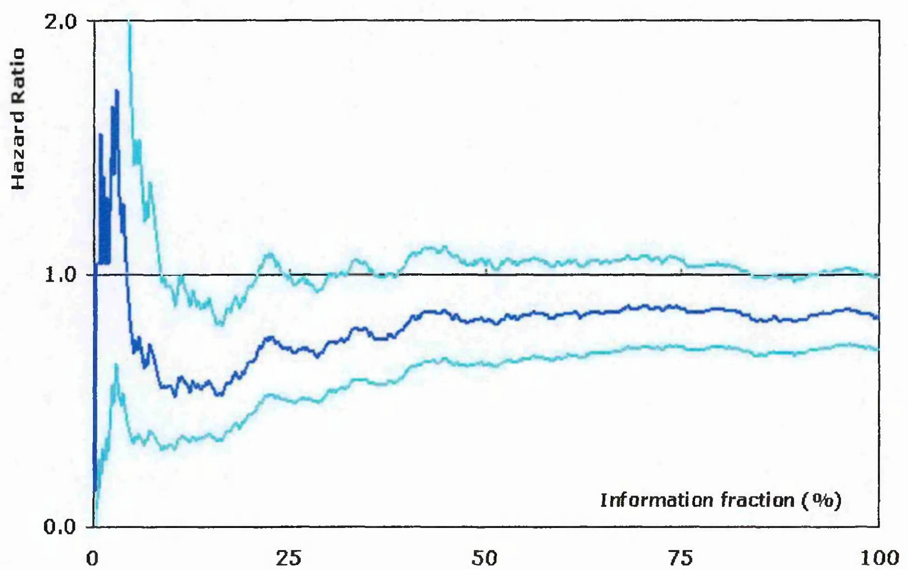

Figure 4.11: **ICON 4/AGO-OVAR 2.2 Trial - Trend over time of the hazard ratio estimate (dark blue) and its lower and upper 95% CI bounds (blue)**

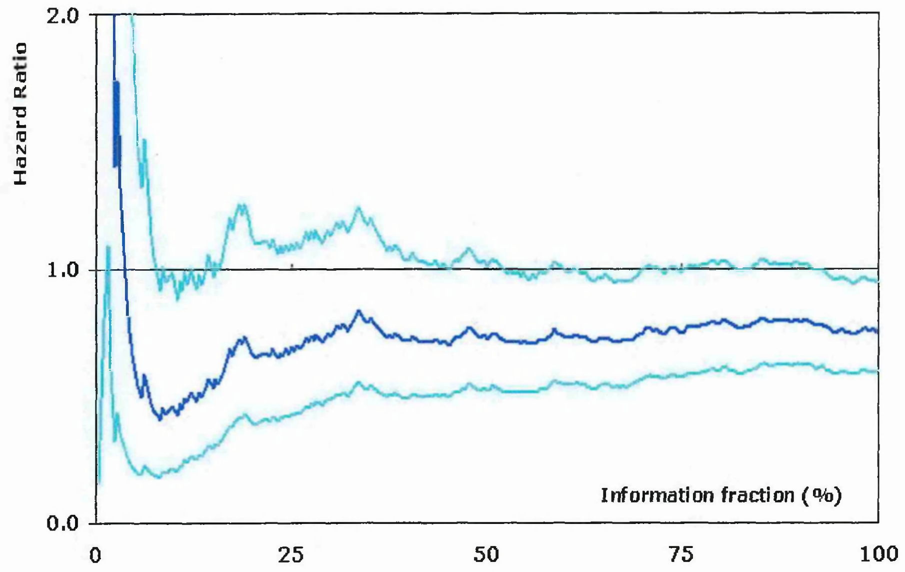

Figure 4.12: **GIVIO/SITAC 01 Trial - Trend over time of the hazard ratio estimate (dark blue) and its lower and upper 95% CI bounds (blue)**

Tables 4.32, 4.33 and 4.34 reports the values of the normal standardized statistic corresponding to the stopping rules, with 4 equally spaced analyses and overall type I error 0.05, for the three considered studies. The graphical plot of same data are presented in Figures 4.13, 4.14 and 4.15, in which for triangular test, only the levels corresponding to the upper boundaries are showed.

Although there is no formal need to present stopping boundaries within the Bayesian approach, the criterion such as "stop if posterior probability that the treatment is beneficial is greater than 97.5%" has been also investigated and results under the sceptical prior are reported as well.

Achieved results are in line with what expected by the different approaches adopted: for alpha-spending function and restricted procedure the boundaries are very similar, while for triangular test and the Bayesian approach the boundaries are more relaxed at the beginning, but a little bit stringent at the end of the study.

As far as Bayesian approach is concerned, it assures a protection similar to the triangular test, and although simulations have not been performed, it seems reasonable to consider that the performance of Bayesian method is similar to what achieved using the triangular test.

Table 4.32: **ICON 3 Trial - Values of the normal standardized statistic corresponding to stopping rules in case of four analyses**

N	Naïve	OBF	Triangular test	Restricted procedure	Sceptical prior
1	±1.9600	±4.33263	+2.95100	±4.00620	±2.66461
2	±1.9600	±2.96311	+2.42652	±2.81635	±2.34060
3	±1.9600	±2.35902	+2.26268	±2.28897	±2.22085
4	±1.9600	±2.01406	+2.20930	±1.98305	±2.15869

For triangular test only the levels corresponding to the upper boundaries are reported

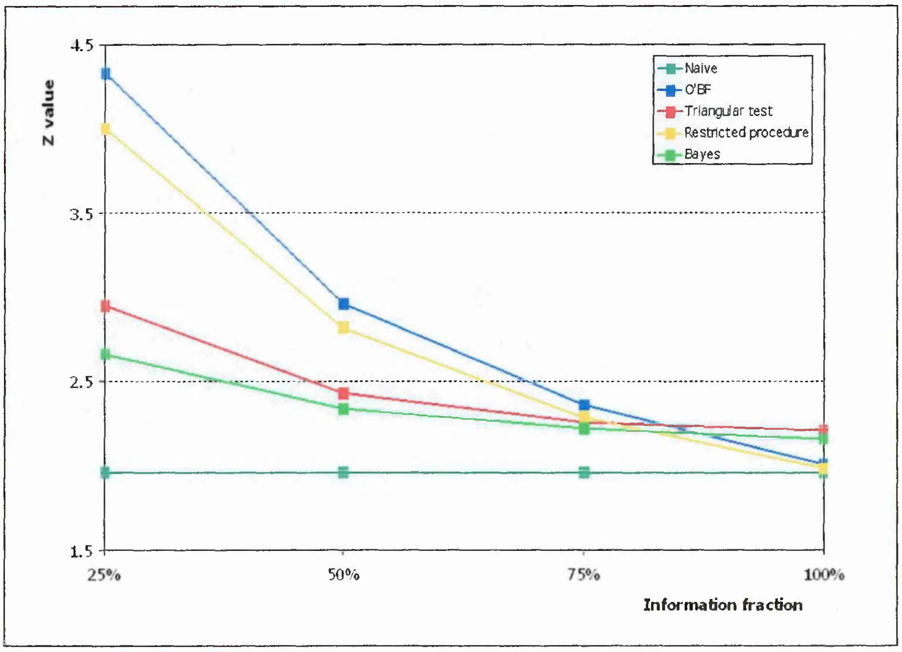

Figure 4.13: **ICON 3 Trial - Values of the normal standardized statistic corresponding to stopping rules in case of four analyses**

Table 4.33: **ICON 4/AGO-OVAR 2.2 - Values of the normal standardized statistic corresponding to stopping rules in case of four analyses**

N	Naïve	OBF	Triangular test	Restricted procedure	Sceptical prior
1	±1.9600	±4.33263	+2.96346	±4.00908	±2.54926
2	±1.9600	±2.96311	+2.45536	±2.86073	±2.27599
3	±1.9600	±2.35902	+2.27815	±2.32252	±2.17473
4	±1.9600	±2.01406	+2.22153	±1.01899	±2.12327

For triangular test only the levels corresponding to the upper boundaries are reported

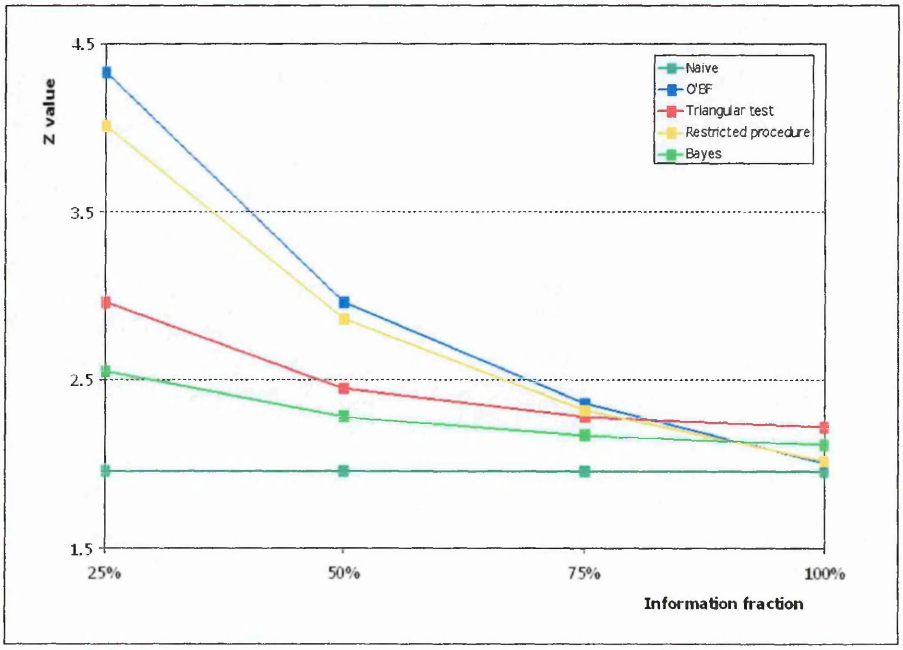

Figure 4.14: **ICON 4/AGO-OVAR 2.2 Trial - Values of the normal standardized statistic corresponding to stopping rules in case of four analyses**

Table 4.34: **GIVIO/SITAC 01 Trial - Values of the normal standardized statistic corresponding to stopping rules in case of four analyses**

N	Naïve	OBF	Triangular test	Restricted procedure	Sceptical prior
1	±1.9600	±4.33263	+2.95701	±4.03311	±2.92591
2	±1.9600	±2.96311	+2.43572	±2.84670	±2.49022
3	±1.9600	±2.35902	+2.27292	±2.32402	±2.32694
4	±1.9600	±2.01406	+2.07325	±1.01827	±2.24178

For triangular test only the levels corresponding to the upper boundaries are reported

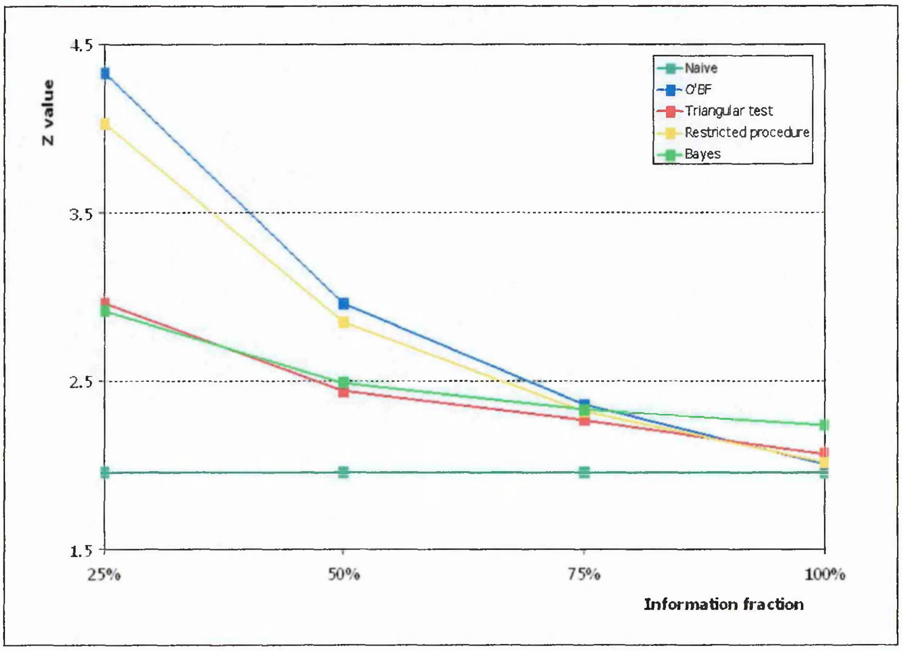

Figure 4.15: **GIVIO/SITAC 01 Trial - Values of the normal standardized statistic corresponding to stopping rules in case of four analyses**

4.2 Early reports in scientific literature

As showed in Figure 4.16, according to the research strategy, a subgroup of 1495 out of 12347 (12%) cancer phase III RCT papers was identified and manually checked in order to pick up the relevant publications, i.e. early reports of therapy trials in oncology.

A vast majority of papers was not relevant for the research, since the choice of using generic terms, such as "early" or "preliminary" increased sensitivity, but at the same time reduced specificity. In fact, "early" is seldom related to the preliminary nature of the publication; rather, it is mainly used for referring to stage of disease or to results obtained in similar previous researches. Likewise, the term "preliminary" is often referred to phase I-II clinical trials.

After the initial manual screening, 178 papers were considered in details: 83 were found to be eligible for the analysis.

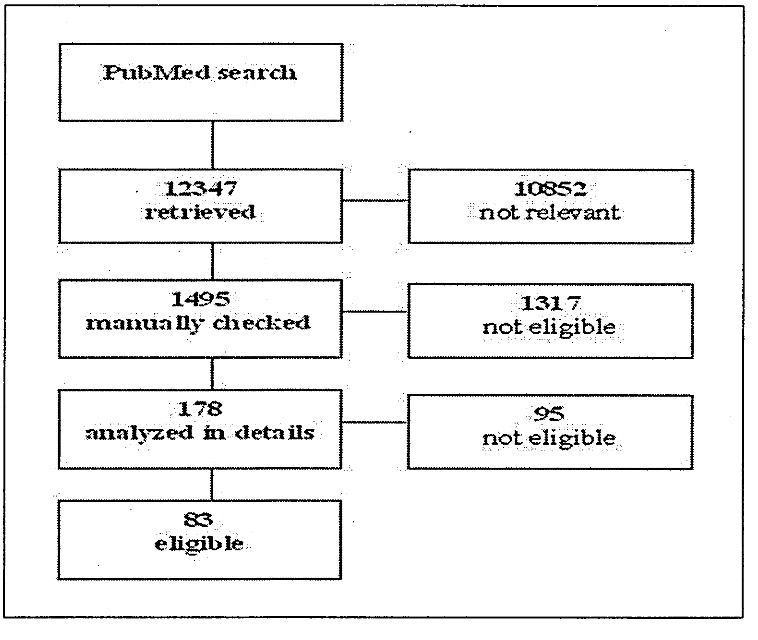

Figure 4.16: **Search flow diagram**

Table 4.35 reports the reasons for publication by year. Sixtyfour out of the 83 RCTs (77%) were interim analyses related to efficacy, while 19 were articles aimed at describing either patient population characteristics or results not related to efficacy endpoints.

We will mainly focus on interim analysis reports on efficacy.

Table 4.35: **Reasons for publication by year**

	Publication reason				**All**	
	Interim analysis		**Early results no efficacy**			
	N	%	N	%	N	%
Year						
2000	14	87.5	2	12.5	16	19.3
2001	8	66.7	4	33.3	12	14.5
2002	17	81.0	4	19.0	21	25.3
2003	17	77.3	5	22.7	22	26.5
2004	6	60.0	4	40.0	10	12.0
2005	2	100.0	.	.	2	2.4
All	**64**	**77.1**	**19**	**22.9**	**83**	**100**

Studies were performed particularly on breast, lung and haematological neoplasms, and the most investigated treatment was chemotherapy (43 out of 64 trials, 67%). These data are shown in Tables 4.36, 4.37 and 4.38.

Table 4.36: **Disease localisation**

Disease localisation	**N**	**%**
Breast	12	18.8
Lung	12	18.8
Haematological	9	14.1
Colorectal	5	7.8
Gastric	4	6.3
Prostate	4	6.3
Genital	4	6.3
Urological	3	4.7
CNS	3	4.7
Head & neck	2	3.1
Pancreatic	2	3.1
Esophageal	2	3.1
Melanoma	1	1.6
GIST	1	1.6
All	**64**	**100**

Table 4.37: **Investigated treatment**

Investigated treatment	N	%
Chemotherapy	43	67.2
Ormonotherapy	7	10.9
Radiotherapy	7	10.9
Surgery	4	6.3
Radiotherapy+chemotherapy	3	4.7
All	**64**	**100**

Table 4.38: **Disease localisation and investigated treatment**

Disease localisation	**Investigated treatment**	N	%
Breast		**12**	**18.8**
	Chemotherapy	7	58.8
	Ormonotherapy	4	33.3
	Surgery	1	8.3
Lung		**12**	**18.8**
	Chemotherapy	12	100.0
Haematological		**9**	**14.1**
	Chemotherapy	8	88.9
	Surgery	1	11.1
Colorectal		**5**	**7.8**
	Chemotherapy	4	80.0
	Radiotherapy+chemotherapy	1	20.0

cont.

cont., Table 4.38

Disease localisation	Investigated treatment	N	%
Gastric		**4**	**6.3**
	Chemotherapy	3	75.0
	Surgery	1	25.0
Prostate		**4**	**6.3**
	Ormonotherapy	2	50.0
	Radiotherapy	2	50.0
Genital		**4**	**6.3**
	Chemotherapy	3	75.0
	Surgery	1	25.0
Urological		**3**	**4.7**
	Chemotherapy	2	66.7
	Ormonotherapy	1	33.3
CNS		**3**	**4.7**
	Radiotherapy	2	66.7
	Radiotherapy+ormonotherapy	1	33.3
Head & neck		**2**	**3.1**
	Radiotherapy	2	100.0
Pancreatic		**2**	**3.1**
	Chemotherapy	2	100.0
Esophageal		**2**	**3.1**
	Radiotherapy	1	50.0
	Radiotherapy+chemotherapy	1	50.0

cont.

cont., Table 4.38

Disease localisation	Investigated treatment	N	%
Melanoma		**1**	**1.6**
	Chemotherapy	1	100.0
GIST		**1**	**1.6**
	Chemotherapy	1	100.0

The results of the trials were published mainly on specialized journals, but even when published on general journals, the impact factor was always high, as shown in Table 4.39.

Table 4.39: **Type of journal and impact factor**

Type of journal	Impact Factor	N	%
INTERIM ANALYSIS			
Specialized	≤ 2	19	29.7
	≤ 4	14	21.9
	> 4	25	39.1
General	> 4	6	9.4
All		**64**	**100**
EARLY RESULTS			
Specialized	≤ 2	8	42.1
	≤ 4	5	26.3
	> 4	6	31.6
All		**19**	**100**

Choice of publication on journals with higher impact factor is associated with the presence of a DSMC and the decision of stopping the trial early, based on the results obtained by the interim analysis (Tables 4.40 and 4.41).

Table 4.40: **Type of journal and impact factor (n=64)**

	Impact factor			
	N	Median	Mean	SD
TYPE OF JOURNAL				
Specialized	58	3.605	4.846	4.014
General	6	34.833	32.080	6.743
TRIAL STOPPED?				
No	27	2.381	3.548	3.454
Yes	37	6.511	10.209	10.770
DSMC				
No	51	3.605	5.555	7.064
Yes	13	10.864	14.635	12.346
STOPPING RULES				
No	29	2.381	3.563	3.749
Yes	35	6.511	10.577	10.853
SAMPLE SIZE				
≤100	5	1.159	3.095	4.566
≤250	24	2.459	6.004	9.456
≤500	22	7.799	7.436	4.568
>500	13	3.772	11.568	13.468

Table 4.41: **Association between higher impact factor with decision of stopping and DSMC presence (n=64)**

Variable	**DF**	**Parameter estimate**	**Standard error**	**t-value**	**p-value**
Intercept	1	3.29	1.567	2.096	0.0402
Trial stopped? (yes/no)	1	4.63	2.199	2.105	0.0395
DSMC (yes/no)	1	7.08	2.699	2.622	0.0110

In 29 (45%) of the trials no formal approach of stopping rules was described.

Bayesian approach was never used, since in the remaining 35 papers a frequentist method was adopted, even if in 20 cases (57.1%) without the specification of the chosen type.

The presence of a DSMC was reported in 13 out of 64 (20%) studies, all in publications of trials using some form of specified approach of interim analysis. DSMC is more reported in trials of medium-large dimensions, i.e. with a number of expected patients greater than 250.

All these data are reported in Tables 4.42 and 4.43.

Table 4.42: **Presence of stopping rules and of a DSMC**

Presence	Type	N	%
STOPPING RULES			
No		**29**	**45.3**
Yes		**35**	**54.7**
	Frequentist, unspecified	20	58.8
	Frequentist, α-spending OBF	6	17.6
	Frequentist, OBF	4	11.8
	Frequentist, α-spending OBF+CP	1	2.9
	Frequentist, Peto+CP	1	2.9
	Frequentist, restricted procedure	1	2.9
	Frequentist, Pocock	1	2.9
	Frequentist, CP	1	2.9
DSMC			
No		**51**	**79.7**
	Presence of stopping rules	29	56.9
	Absence of stopping rules	22	43.1
Yes		**13**	**20.3**
	Presence of stopping rules	13	100
All		**64**	**100**

Table 4.43: **Presence of stopping rules and expected sample size (n=64)**

Sample size	Stopping rules				DSMC			
	No		Yes		No		Yes	
	N	%	N	%	N	%	N	%
≤100	4	13.8	1	2.9	4	7.8	1	7.7
≤250	15	51.7	9	25.7	22	43.1	2	15.4
≤500	7	24.1	15	42.9	17	33.3	5	38.5
>500	3	10.3	10	28.6	8	15.7	5	38.5

The effect of interim analysis results on continuation of the trial are showed in Table 4.44.

In four papers, the trials were stopped for reasons other than efficacy or futility, i.e. for an accrual lower than expected or lack of resources. In 27 out of 64 papers the study was continued. Stopping for efficacy reasons was reported in 24 papers, while in 9 cases the study was reported to have been stopped for futility.

Interestingly, in those papers not mentioning any form of policy for interim analyses, the studies continued in 20 (69%) out of 29, while in papers reporting some form of policy, only 7 (20%) trials continued and the remaining 35 were stopped early. The principal reason for stopping was efficacy, accounting for 24 studies (37.5%); the ratio between reason for stopping (effficacy/futility) was 2.7:1, not associated to reporting of policy presence.

Table 4.44: **Decision taken based on interim analysis results (n=64)**

Stopping rules	**Continued**		**Stopped efficacy**		**Stopped futility**		**Stopped other**	
	N	%	N	%	N	%	N	%
No	20	69.0	5	17.2	2	6.9	2	6.9
Yes	7	20.0	19	54.3	7	20.0	2	5.7
All	**27**	**42.2**	**24**	**37.5**	**9**	**14.1**	**4**	**6.2**

In 36 (56%) studies the early report was drawn while the study accrual was still ongoing, in 19 out of 36 the study was interrupted for efficacy and in 6 out of 36 for futility. Instead, only in 9 out of 28 (32%) studies in which analysis was performed after the end of the accrual, the trial was stopped (Table 4.45).

The relationship among the decision of interrupting the study (and therefore of producing an early publication) and the presence of a policy of stopping rules, the presence of a DSMC and the trial size was assessed by means a logistic regression model, whose results are reported in Table 4.46. They suggest that, when a trial is stopped, the probability of publishing the results of an interim analysis is increased in presence of specific stopping rules.

Table 4.45: **Accrual status, presence of stopping rules and results of interim analysis (n=64)**

		Results of interim analysis							
		Continued		**Stopped efficacy**		**Stopped futility**		**Stopped other**	
		N	%	N	%	N	%	N	%
Accrual status	**Stopping rules**								
Ongoing	No	3	33.3	3	33.3	1	11.1	2	22.2
	Yes	5	18.5	16	59.3	5	18.5	1	3.7
Closed	No	17	85.0	2	10.0	1	5.0	.	.
	Yes	2	25.0	3	37.5	2	25.0	1	12.5
All		**27**	**42.2**	**24**	**37.5**	**9**	**14.1**	**4**	**6.2**

Table 4.46: **Relationship among of probability of early publication and study characteristics**

Variable	Point estimate	95% Wald confidence intervals	
DSMC (yes:no)	7.170	0.504	101.923
Stopping rules (yes:no)	6.566	1.832	23.533
Sample size	0.888	0.702	1.124

4.3 Use of interim analyses in randomized oncological trials

According to the research strategy, a subgroup of 836 cancer protocols was identified and manually checked in order to pick up the relevant trials.

Figure 4.17 reports the flow of the selection of relevant protocols: 143 (17.1%) out of 836 protocols were eligible and evaluable for analysis. 128 (15.3%) investigated outcomes different from time to event, such as pain control, 406 (48.6%) were single arm trials, 64 (7.7%) considered a time to event endpoint only as secondary outcome measure. Ninetyfive (11.4%) studies at the moment of the analysis were not yet fully included in the registry, since their protocol was not available.

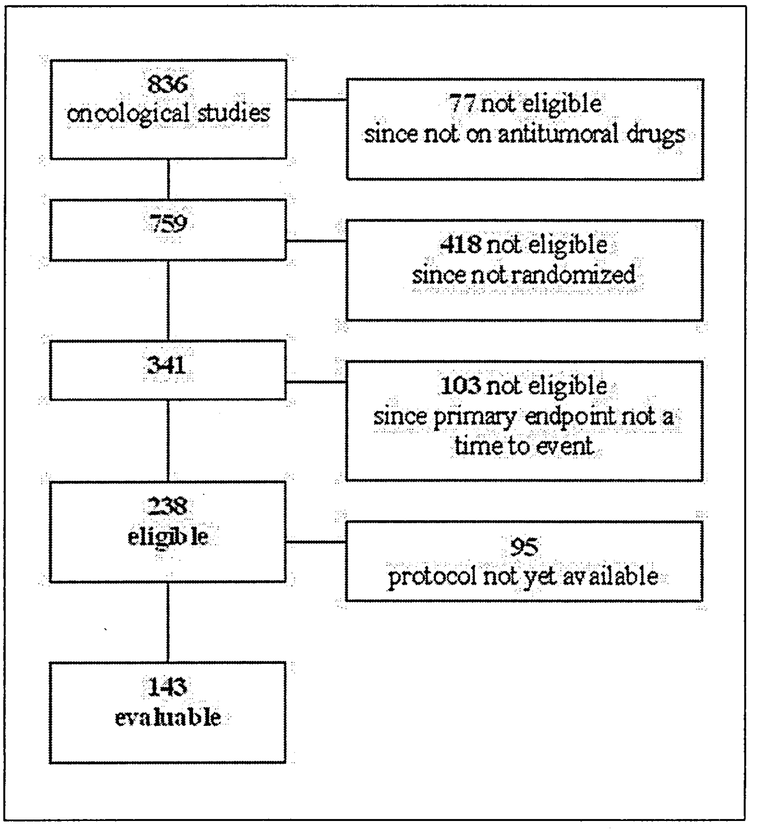

Figure 4.17: **Search flow diagram**

Comparison of the 143 fully eligible and evaluable protocols with the 95 eligible but not yet evaluable trials (Table 4.47) shows that not evaluable studies are less recent, since mostly evaluated before 2002 (68.4% compared to 23.1% in evaluable trials, χ^2=57.9, 5 df, p< 0.0001), launched generally only in Italy (46.0% *vs.* 25.2%, χ^2=10.4, 2 df, p=0.0054), more often of phase II (22.1% *vs.* 7.0%, χ^2=10.1, 2 df, p=0.012) and preferentially conducted by no profit organisations (χ^2=14.04, 1 df, p=0.0002).

Table 4.47: **Comparison between evaluable and not evaluable trials**

	Not evaluable		Evaluable	
	N	%	N	%
YEAR				
2000	36	37.9	13	9.1
2001	29	30.5	20	14.0
2002	13	13.7	19	13.3
2003	10	10.5	31	21.7
2004	4	4.2	38	26.6
2005	3	3.2	22	15.4
SPONSOR				
No profit	56	59.0	46	32.2
Profit	39	41.0	97	67.8
PHASE				
II	21	22.1	10	7.0
III	72	75.8	132	92.3
IV	2	2.1	1	0.7
INVOLVED COUNTRIES				
Italy	35	46.0	36	25.2
Europe	1	1.3	9	6.3
Worlwide	40	52.6	98	68.5
not specified	19			
All	**95**		**143**	

Tables from 4.48 to 4.53 report the characteristics concerning disease localisation, the investigated treatment and the clinical setting of the 143 evaluable studies.

The large majority of the studies are to be conducted on solid tumors (85.3%), and the more frequently investigated diseases are lung (37, 25.9%) and breast cancers (30, 21.0%).

Chemotherapy is the preferred investigated strategy, accounting for 103 (72%) of the protocols, while ormonotherapy with or without chemotherapy is tested in 14 studies, 11 out of the 30 (36.7%) breast cancer protocols and all of the 3 protocols on advanced prostate cancer.

Table 4.48: **Disease localisation (n=143)**

Disease localisation	N	%
Solid		
Lung	37	25.9
Breast	30	21.0
Colorectal	10	7.0
Head & Neck	7	4.9
Urogenital	7	4.9
Gynaecological	6	4.2
Melanoma	5	3.5
Pancreas	4	2.8
Liver	3	2.1
Prostate	3	2.1
Gastric	3	2.1
Sarcoma	2	1.4
Adrenocortical	1	0.7
Biliary	1	0.7
Brain	1	0.7
GIST	1	0.7
Anal	1	0.7
Haematological	21	14.7

Table 4.49: **Investigated treatment**

Investigated treatment	N	%
Chemotherapy	103	72.0
Chemotherapy+radiotherapy	10	7.0
Ormonotherapy	7	4.9
Chemotherapy+ormonotherapy	7	4.9
Immunotherapy	6	4.2
Chemotherapy+immunotherapy	5	3.5
Other	3	2.1
Chemotherapy+other	2	1.4
All	**143**	**100**

Table 4.50: **Clinical setting**

Tumor	**Clinical setting**		**N**	**%**
Solid			**122**	**85.3**
	Early		23	18.9
		Adjuvant	19	82.6
		Neoadjuvant	4	17.4
	Advanced		99	81.2
		Locally advanced	55	55.5
		Metastatic	26	26.3
		Both	18	18.2
Haematological			**21**	**14.7**
	Adult		19	90.5
	Pediatric		2	9.5
All			**143**	**100**

Table 4.51: **Clinical setting and tumor localisation** ***(only for solid tumors)***

Disease localisation	Setting	N	%
Lung		**37**	
	Early	3	8.1
	Advanced	34	91.9
Breast		**30**	
	Early	8	26.7
	Advanced	22	73.3
Colorectal		**10**	
	Early	4	40.0
	Advanced	6	60.0
Urogenital		**7**	
	Early	2	28.6
	Advanced	5	71.4
Head & neck		**7**	
	Advanced	7	100.0
Gynaecological		**6**	
	Advanced	6	100.0
Melanoma		**5**	
	Early	1	20.0
	Advanced	4	80.0
Pancreas		**4**	
	Advanced	4	100.0
Liver		**3**	
	Early	1	33.3
	Advanced	2	66.7

cont.

cont., Table 4.51

Disease localisation	Setting	N	%
Prostate		**3**	
	Advanced	3	100.0
Gastric		**3**	
	Early	3	100.0
Sarcoma		**2**	
	Early	1	50.0
	Advanced	1	50.0
Anal		**1**	
	Advanced	1	100.0
Adrenocortical		**1**	
	Advanced	1	100.0
Biliary		**1**	
	Advanced	1	100.0
Brain		**1**	
	Advanced	1	100.0
GIST		**1**	
	Advanced	1	100.0
All		**122**	**100**

Table 4.52: **Treatment and clinical setting**

Tumor	Setting	Treatment	N	%
Solid			**122**	**85.3**
	Early		23	18.8
		Chemotherapy	14	60.9
		Ormonotherapy	4	17.4
		Immunotherapy	3	13.0
		Chemo+radiotherapy	1	4.4
		Chemo+ormonotherapy	1	4.4
	Advanced		99	81.2
		Chemotherapy	72	72.3
		Chemo+radiotherapy	8	8.1
		Chemo+ormonotherapy	6	6.1
		Chemo+immunotherapy	5	5.0
		Ormonotherapy	3	3.0
		Immunotherapy	3	3.0
		Other	2	2.0
Haematological			**21**	**14.7**
	Adult		19	90.5
		Chemotherapy	15	79.0
		Chemo+other	2	10.5
		Chemo+radiotherapy	1	5.3
		Other	4	5.3
	Pediatric		2	9.5
		Chemotherapy	2	100.0

Table 4.53: **Disease localisation and investigated treatment** ***(only for solid tumors)***

Disease localisation	Treatment	N	%
Lung		**37**	
	Chemotherapy	32	86.5
	Chemo+radiotherapy	1	2.7
	Immunotherapy	2	5.4
	Chemo+immunotherapy	1	2.7
	Other	1	2.7
Breast		**30**	
	Chemotherapy	17	56.7
	Chemo+radiotherapy	1	3.3
	Chemo+ormonotherapy	6	20.0
	Ormonotherapy	5	16.7
	Other	1	3.3
Colorectal		**10**	
	Chemotherapy	8	80.0
	Chemo+radiotherapy	1	10.0
	Immunotherapy	6	10.0
Urogenital		**7**	
	Chemotherapy	5	71.4
	Chemo+immunotherapy	2	28.6
Head & neck		**7**	
	Chemotherapy	2	28.6
	Chemo+radiotherapy	2	71.4

cont.

cont., Table 4.53

Disease localisation	Treatment	N	%
Gynaecological		**6**	
	Chemotherapy	6	100.0
Melanoma		**5**	
	Chemotherapy	2	40.0
	Immunotherapy	1	20.0
	Chemo+immunotherapy	2	40.0
Pancreas		**4**	
	Chemotherapy	4	100.0
Liver		**3**	
	Chemotherapy	2	33.3
	Immunotherapy	2	66.7
Prostate		**3**	
	Chemo+ormonotherapy	1	33.3
	Ormonotherapy	2	66.7
Gastric		**3**	
	Chemotherapy	3	100.0
Sarcoma		**2**	
	Chemotherapy	2	100.0
Anal		**1**	
	Chemo+radiotherapy	1	100.0
Adrenocortical		**1**	
	Chemotherapy	1	100.0

cont.

cont., Table 4.53

Disease localisation	Treatment	N	%
Biliary		**1**	
	Chemotherapy	1	100.0
Brain		**1**	
	Chemotherapy	1	100.0
GIST		**1**	
	Chemotherapy	1	100.0
All		**122**	**100**

Table 4.54 describes the planned number of patients to randomize, of the events to observe toghether with the expected proportion of events at the end of the study, a good index of patient prognosis, calculated as the ratio of these two latter variables.

Table 4.55 shows the planned duration of the study, as well as those of the accrual and follow-up.

Table 4.54: **Planned number of patients, events and expected proportion of events at final analysis**

	N	Median	Mean	SD	Min	Max
N. of patients	143	490.0	712.6	846.96	36	5800
N. of events	103	390.0	439.4	263.92	6	1280
Events/ patients	103	0.70	0.65	0.196	0.13	0.95

Table 4.55: **Duration of the study, accrual and follow-up (months)**

	N	Median	Mean	SD	Min	Max
Study duration	110	42.5	49.8	26.99	15.0	150.0
Accrual	116	24.0	26.9	13.32	9.0	84.0
Follow-up	109	18.0	23.9	19.11	2.0	120.0

Table 4.56 reports a description of interim analyses, if planned, and the presence of a DSMC.

Interim analyses were planned in 92 (64.3%) of the protocols, while DSMC was reported in 83 (58.0%) out of the 143 evaluable protocols.

Table 4.56: **Presence of stopping rules and of a DSMC**

		N	%
Interim analysis			
No		51	35.7
Yes		92	64.3
DSMC			
No		60	42.0
Yes		83	58.0
Interim analysis	**DSMC**		
No	No	38	26.6
	Yes	13	9.1
Yes	No	22	15.4
	Yes	70	49.0
All		**143**	**100**

The median number of interim analyses was 1, ranging from 0 to 5 for efficacy, and from 0 to 15 for safety assessment, as reported in Table 4.57.

Table 4.57: **Number of interim analyses for efficacy or safety**

	N	Median	Mean	SD	Min	Max
For efficacy	92	1.0	1.3	0.88	0	5
For safety	90	0.0	0.7	1.98	0	15

Table 4.58 shows the main characteristics of the interim efficacy analyses, while Table 4.59 those of interim safety analyses.

Of note, among the 80 protocols reporting the objective of efficacy analysis, in 6 (7.5%) cases the endpoint is related to activity, and therefore it is different from the main objective of the final analysis.

Interim analyses for efficacy were planned according to the proportion of observed events in 54 out of 80 evaluable protocols, according to the proportion of patients in 20 protocols and based on calendar time in the remaining 6 protocols.

In the 22 trials reporting the planned timing strategy for safety analyses, they were generally based on patients proportion (14), on calendar time (4) or on the number of toxic events (4).

Table 4.58: **Characteristics of the interim efficacy analyses**

	N	%
Objective		
Not stated	12	13.0
Activity endpoint	6	6.5
Efficacy endpoint	72	78.5
Mixed endpoint	2	2.2
Timing		
No	14	15.2
Yes	78	84.8
Type of timing		
Not stated	12	13.0
Events	51	55.4
Patients	20	21.7
Calendar time	6	6.5
Patients and events	3	3.3
All	**92**	**100**

Table 4.59: **Characteristics of the interim safety analyses**

	N	%
Objective		
Not stated	67	72.8
Safety endpoint	25	27.2
Timing		
No	75	81.5
Yes	17	18.5
Type of timing		
Not stated	70	76.1
Events	4	4.3
Patients	14	15.2
Calendar time	4	4.3
All	**92**	**100**

The most frequent type of statistical approach for the analysis was the frequentist method, implementing the OBF boundaries (70 out of 92, 76.1%); conditional power was used in 3 studies, while Bayesian approach was taken into consideration only in one study, as shown in Table 4.60.

A DSMC was planned in 13 out of 51 protocols (25%) without a planned interim analysis, while it was not considered in 22 out of 72 protocols in which interim analysis was planned.

Overall, no form of monitoring was found in 38 out of 143 protocols (27%).

Table 4.60: **Interim analysis and presence of DSMC**

	DSMC present		DSMC not present		All	
Type	N	%	N	%	N	%
Not planned	13	15.7	38	63.3	51	35.7
Freq, OBF[a]	53	63.9	17	28.3	70	49.0
Freq, Peto[b]	2	2.4	1	1.7	3	2.1
Freq, TT[c]	1	1.2	0	-	1	0.7
Freq, CP[d]	2	2.4	1	1.7	3	2.1
Freq, other	11	13.2	3	5.0	14	9.8
Bayesian	1	1.2	0	-	1	0.7
All	**83**	**58.0**	**60**	**42.0**	**143**	**100**

[a] Frequentist, Alpha-spending function with O'Brien and Fleming boundaries
[b] Frequentist, Peto approach
[c] Frequentist, Triangular test
[d] Frequentist, Conditional power

Table 4.61 shows that in 18 out of 83 protocols, the only commitment of DSMC was safety. Efficacy was considered in 65 protocols, but only in two it was the only commitment of DSMC, while in the remaining 63 the DSMC had the task of monitoring safety and efficacy. In one case, DSMC had also the task of deciding the timing and type of statistical analysis.

Composition and frequence of meetings for DSMC were reported only in 6 and 18 protocols, respectively. Usually the DSMC was composed by 3 or 4 people, always with one statistician, and the frequency of the meetings was generally every 6 months or every year.

Table 4.61: **DSMC tasks**

	N	%
Task		
Efficacy	2	2.4
Safety	18	21.7
Efficacy+safety	62	74.7
Efficacy+safety+statistical analysis	1	1.2
Number of DSMC meetings		
Not defined	77	92.8
3	5	6.0
4	1	1.2
Frequency of DSMC meetings		
Not defined	71	85.5
Every 6 months	8	9.6
Every 12 months	4	4.8
All	**83**	**100**

Table 4.62 shows the association among selected characteristics of the study protocols (type of sponsorship, international collaboration, year of submission) and the presence of both planned interim analyses and DSMC.

Although all these factors but year of submission show an association at univariate analysis, at multivariate logistic analysis (Table 4.63) the most important factor associated with the presence of interim analysis or of DSMC is the international organization of the study, accounting for an odds ratio for the presence of DSMC of 5.6 (95% CI 2.2-14.2) and of 5.1 (95% CI 1.9-13.1) for presence of a planned interim analysis.

Table 4.62: **Association among study protocol characteristics and presence of interim analysis and DSMC (n=143)**

	DSMC				Interim			
	No		Yes		No		Yes	
	N	%	N	%	N	%	N	%
Sponsor								
No profit	29	48.3	17	20.5	22	43.1	24	26.1
Profit	31	51.7	66	79.5	29	56.9	68	73.9
International study								
No	33	55.0	12	14.5	27	52.9	18	19.6
Yes	27	45.0	71	85.5	24	47.1	74	80.4
Year of submission								
2000	5	8.3	8	9.6	5	9.8	8	8.7
2001	11	18.3	9	10.8	6	11.8	14	15.2
2002	11	18.3	8	9.6	8	15.7	11	12.0
2003	13	21.7	18	21.7	12	23.5	19	20.7
2004	11	18.3	27	32.5	12	23.5	26	28.3
2005	9	15.0	13	15.7	8	15.7	14	15.2

Table 4.63: **Relationship between presence of both interim analyses and DSMC and selected protocol characteristics**

Variable	Point estimate	95% Wald CI	
Odds ratio estimates for presence of interim analyses			
Sponsor (profit/no profit)	0.857	0.331	2.214
International collaboration (yes/no)	5.106	1.988	13.114
Year of submission	0.969	0.763	1.231
Odds ratio estimates for presence of DSMC			
Sponsor (profit/no profit)	1.553	0.614	3.930
International collaboration (yes/no)	5.585	2.200	14.179
Year of submission	1.144	0.898	1.457

Chapter 5

Discussion

5.1 Statistical findings and methodological context

The most important findings of our research can be summarised as follows:

- the adoption of a statistical approach for data monitoring, no matter of which type chosen, effectively protects from the risk of an incorrect early stopping;

- if no stopping rule is adopted, the probability of early stopping with a higher estimation bias is noticeably increased;

- performance of restricted procedure and alpha-spending function with OBF boundaries are very similar, while triangular test yields results which resemble those obtained by Bayesian approach. Triangular test performs well regarding overestimation, but the more relaxed criteria for stopping for futility increase the study probability of interruption, concluding for no difference between treatments;

- the chance of early stopping due to an overestimate is directly related to the "true" magnitude of effect; the inverse holds for stopping for futility;

- the number of analysis has a moderate impact on estimation, when some approach is adopted, but it is important when no criteria for controlling for multiple analyses are used.

- stabilization of the estimates appears to happen when a substantial amount of events has occurred. Therefore, it seems appropriate to conduct interim analyses only after about half of the expected events occurred, in order to reduce bias. With respect to this, alpha-spending function and restricted procedure are more protective against stopping at the beginning of the study, favouring a reduction in the magnitude of estimation bias.

These findings confirm previous research in the field (Pocock and Hughes, 1989; Korn *et al.*, 2004) and stress the importance of adopting some form of statistical approach for data monitoring.

Our research, rather than finding differences, emphasizes the qualitative similarity of the various options, in that they all are conservative in protecting against an inappropriate early stopping of a trial.

Nevertheless, if bias can be reduced, it cannot be avoided, since at any particular interim analysis, clinical trials with an observed effect that is by chance greater than the true effect ("random high"), are more likely to exceed the stopping boundaries than trials on a "random low", especially if the true effect is relatively small.

Our research confirms on real-life data most of the components associated to the bias due to multiple looks and well described in the literature (Hughes and Pocock, 1988), which can be summarised as follows:

- bias is higher when no sequential design is adopted; however, even when it is used, a distribution skewed toward overestimation of effect is observed;

- bias increases with the ease with which a trial can stop early;

- the size of trial affects the results: smaller trials usually produce less extreme estimates than larger trials but, if they stop early, result in a more exaggerate estimation of treatment effect at corresponding stage;

- bias is more marked when the true risk ratio can be detected with a reasonable power;

- for moderate treatment effects the bias is increased for designs that allow early termination more readily;

- different frequencies of monitoring may have a moderate impact on biased estimation.

An other relevant issue is related to the impact on bias estimation of different frequencies of monitoring.

In agreement with this last statement, Freidlin *et al.* (1999) argued that in terms of protection level and power there is little reason not to monitor frequently the relative treatment efficacy, and frequent monitoring offers advantages in being able to end some trials earlier.

However, although the additional bias seems to be small, the size of the adjustment is dependent on the unknown true hazard ratio and there are no means of knowing whether the surprisingly large observed effect is true. Therefore it is always important to consider the possibility that chance has played a part in achieving the observed results.

The finding that estimation bias tends to be reduced when the observed number of events is closer to the planned size is particularly important for cancer clinical trials also for non-statistical reasons. With time to event endpoints, a potential problem with stopping a trial earlier is that the early experience with short follow-up may not reflect accurately the complete survival experience.

A new treatment may be very toxic, leading to a few early deaths, but may also have much better long-term results than the standard treatment. Overall, the new treatment may be viewed as better than the standard, but the early look at the data may suggest stopping for lack of efficacy. The opposite is also possible, since early suggestions of treatment efficacy may decline over time. This problem can be lessened by deferring the formal interim monitoring.

One further factor that should be considered is the scientific and ethical relationship that links the decision of stopping a trial early with its implication on ongoing trials, addressing the same clinical question, and on the chance that a confirmatory trial is planned. In either these situations, it is suggested that the issues of interpretation of observed effect (plausibility: unrealistically large results needed for stopping; precision: imprecision due to the small sample size in the early analyses) might be better faced in Bayesian context.

Recognition of requirement for large effects to stop the trial early leads to shrinkage methods to produce plausible estimates.

This results in a Bayesian approach, whereby the plausibility of different treatment effects is quantified beforehand. This quantification is of course subjective, but it is important to note that the classical specification of nominal significance for stopping is arbitrary, too and the choice among the various monitoring rules may also reflect prior opinion.

Bayesian approach has the advantage of emphasizing estimation rather than significance testing, and the choice of prior can reflect the degree of expectation of genuine treatment differences. As an example, in tumors for which no or small previous advances in therapy have been made, the prior should be centred toward zero effect, thus requiring more data before apparent treatment effect would justify stopping.

5.2 Contribute of surveys on protocols and published early reports

Analysis of protocols and early reports suggests that, although the field of methodology of interim analyses of clinical trials is largely covered and different approaches are available, the implementation of these procedures in a monitoring strategy is still scarce.

According to the sources of data investigated, analysis of statistical aspects of RCT protocols in oncology, systematically collected in the National Monitoring Centre for Clinical Trials, reveals that the most recent trend, based on the analysis of the international and national trials with participation of Italian centres, is not yet completely satisfactory.

The most important figures derived indicate that only sixty-four percent of the protocols incorporate statistical interim analysis plans. Despite of the large availability of statistical methods for interim analysis, the almost uniquely approach is the frequentist method, with OBF boundaries.

DSMCs are present in 58% of protocols, but there is lack of information on their composition and on the rules to be implemented.

The only factor clearly associated to the adoption of planned interim analyses and to the presence of DSMC is the multinational participation to the study.

Some positive findings are that, in accordance to the characteristics of survival analysis, usually the timing is related to the number of events and almost half of the trials adopt not more than one planned interim analysis, thus reducing the risk of estimation bias.

When looking at the data derived from early reports, the adoption of a formal process of interim analysis affects only a minority (13%) of published trials and slightly more than half (55%) of early publications based on interim analysis.

Again, the largely preferred approach is the frequentist method, generally with OBF boundaries.

Explicit use of DSMC is reported only in 10% of published reports and in 20% of reports of early publication based on interim analyses, with lack of information regarding its rules and composition. DSMC presence is associated to the size of clinical trial.

Publication of early results on high impact factors journals is associated to the decision of stopping the trial early, based on the results obtained by the interim analysis and to the presence of a DSMC.

These data, together with the finding that the publication of trials interrupted due to the results of interim efficacy analysis are associated with planned stopping rules, suggest that access to the publication of early reports, particularly on highly referenced journals, may depend not only from the observed effect of the treatment but also from the reliability of monitoring process.

The findings related to interim analyses are not so dissimilar from those emerged from the Italian registry of clinical trial protocols, and are complementary to those published in the Health Technology Assessment (HTA) report on issues in data monitoring and interim analysis of trials (Grant *et al.*, 2005).

In that report, it was shown that planned interim analysis was reported in 16% and presence of DSMC in 18% of randomised clinical trials published in selected general medical journal and specialist medical journals in 2000. Focusing on oncology journals, these figures were even worse, since only 8% of trials reported DSMC presence in the same year.

Since our research strategy for selecting relevant papers is different from the approach used in the HTA report study, these results are not easily comparable.

We chose only papers reporting early publications from 2000 to 2005, while the survey of HTA focused on retrieving all the trials published in 1990 and 2000 on selected journals.

The reason of our choice was that we were particularly interested in picking up reports in which some mention on the form of interim analysis would have to be expected, thus overcoming the potential problem of underreporting in HTA analysis. Nevertheless, although the data may suggest some sort of improvement in reporting details of interim analysis, the data are still unsatisfactory, since even in this more favourable situation, 45% of reports did not mention any form of planned stopping rule.

Moreover, our data, particularly those derived from the protocol search, confirm the scarce attitude to provide details on the rules that a DSMC should adopt, on the relationship between study Steering Committee and DSMC, so reinforcing the importance of recommending that explicit guidelines must be prepared for each DSMC prior to the start of the trial, specifying clearly how it will operate (Sydes *et al.*, 2004).

5.3 Conclusions

The most important 'take-home' message of our research is that interim analyses play a fundamental role in the balance between the need of timely information regarding the treatment effect and the control of false positive errors and estimation bias.

Since trials are often analysed before their planned end, it is absolutely necessary to implement statistical stopping rules in the context of the monitoring process, which should be performed by an appropriate DSMC.

The most discussed and popular approaches appear to have good performance.

However, the use of interim analyses is still limited basically to the OBF frequentist approach, while the Bayesian method is not considered, although in the context of monitoring it would be more useful for its characteristics of flexibility in incorporating external evidence.

Interim analysis plan are still scarcely described, even in more recent protocols, denoting a not yet sufficient attention to this issues not only by the researchers, but also by the regulatory boards that should consider the ethical and scientific aspects of the submitted studies.

Also, the evidence of underreporting of statistical methods on journals, even when early reports based on results of interim analyses are published, can be considered as a further signal of the gap between methodological availability of statistical methods and their actual use.

These considerations induce some further thoughts relative to the discrepancy between the perceived importance of data monitoring boards and the scarcity of data regarding the description of their presence in clinical trials, their composition and their role.

It can be argued that the importance of adoption of a monitoring strategy is far more relevant than the choice of a particular type of statistical analysis. As a matter of facts, when monitoring clinical trials, many problems can face for a sort of different issues. For example, some of the most important problems may be related to the different endpoints chosen for interim analysis, whose relevance can also be different when compared to the time of analysis, or to the evidence of effects in selected subgroups of patients, but not in others (Ellenberg et *et al.*, 2003).

Conflicting results on type of endpoint, chosen time (long term vs. early) and in subgroups of patients constitute perhaps the most important conceptual issue of clinical trials interpretation and they are even more problematic in interim analyses.

Conflicts in ethical and philosophical point of views also constitute other group of issues: some investigators argue that, at least for some trials, the objective should be to produce results that are persuasive enough to effect changes in medical practice (Liberati, 1994).

In particular situation, such as in confirmatory pragmatic trials where researchers are more interested in effect estimation than in assessing the efficacy of the experimental treatment, trial interruption may provoke an important loss of precision in the estimate of benefit or detriment associated to the treatment under investigation (Sohuami, 1994).

On the other hand, other investigators may not be keen to accept that a trial might be continued far longer than necessary to persuade most knowledgeable clinical researchers, requiring continue randomisation of participants to an inferior treatment.

Determining the optimal length of follow-up can be difficult in a clinical trial having an early beneficial trend. Ideally, evaluating the duration of treatment benefit while continuing to assess possible side-effects or toxicity over a longer period of time would provide the maximum information for clinical use.

However this solution is not always viable, since for patients with a life-threatening disease such as advanced cancer, evidence of short-term therapeutic benefits may be compelling even it is unknown whether these benefits are sustained over the longer term. It is also important to recognise the effect of early stopping on the complete pattern of knowledge which is expected from a trial: after the trial has been early closed, patients on the control arm may begin to receive the new beneficial treatment, comparisons of the study arms become less meaningful and evaluating long-term side-effects and whether benefit is sustained becomes more difficult. In these cases the choice of implementing interim analyses, and the weight of statistical results on the decision to stop the trial, can vary according to different clinical scenarios: for patients with a chronic disease the long-term effects of the therapy may be of greater importance in evaluating the benefit-to-risk ratio. In this case, a focus on longer-term outcomes may sometimes be justified even in the presence of a strong but short-term beneficial trend. When such diseases are progressive, however, there will inevitably be a conflict between the desire to prevent irreversible disease progression in as many patients as possible, and the desire to understand the long-term effects of the treatment, which should be solved on a "case by case" basis.

A further issue raises when an unexpected toxicity profile begins to emerge. In this case, the level of a safety concern that would lead to a recommendation for modification or termination of the study will necessarily vary with the level of benefit being observed. If the treatment appears to be offering a survival benefit, for example, a strong suggestion of a serious and unexpected safety problem might lead to some changes in the protocol to reduce the problem, while the same magnitude of safety concern might lead to a recommendation to terminate the study if the interim efficacy results were less promising.

For this reason when interim monitoring of comparative data is conducted to assess safety issues, efficacy data should also be reviewed in order to enable an informed assessment of the benefit-to-risk profile. In some trials, no apparent trends of either beneficial or harmful effects emerge as the trial progresses toward its planned conclusion. In such instances, decision should take into account the investment in participants, physicians and resources, as well as the discomfort of the trial on patients.

All these examples illustrate the difficulties of the decisions that sometime have to be taken, when monitoring a clinical trial and the important effects such decisions may have. They make even more clearer how important is the role of DSMC and why interim analyses must be considered an important tool to be used as guideline for decision.

There are good examples in literature regarding the positive dialectic between the results of interim analysis and the decision of DSMC.

Wheatley and Clayton (2003) stated that the preliminary results of the twelfth Medical Research Council acute myeloid leukemia trial (Wheatley *et al.*, 2002) showed no evidence of a survival advantage for five courses of therapy compared to four courses in a randomised comparison involving 1078 patients (HR 1.09, 95% CI 0.87-1.37, p=0.4). However, the data presented to the DSMC at both its reviews in 1998 suggested large benefits for the additional course with HRs of 0.47 and 0.55 (95% CIs 0.29-0.77 and 0.38-0.80, p=0.003 and p=0.002, respectively).

Despite these highly significant findings, the DSMC did not recommend closure of the randomisation. In this example, the choice of fixed stopping rules based on p-value was questionable. According to Whitehead (2004), the best choice for this trial would have been a triangular test with asymmetric boundaries, since in this case experimental arm was more toxic and expensive than control arm. Using this approach, the study would have continued and terminated with a non-significant conclusion. However, the main reason for not closing the randomization was not related to the maybe inappropriate statistical method chosen for interim analysis, rather it was based on the consideration that the treatment effects observed early (53% and 45% reductions in the odds of death) were considered too large to be clinically plausible, despite the p-values associated with them.

Investigations did not identify any clinical explanation, such as different types of patients in the earlier and later parts of the trial, to explain the loss of benefit as the trial progressed. Thus, the most likely current explanation for the large benefit observed early on was the play of chance.

According to the authors, the following considerations have to be made: fixed stopping rules based simply on a rigid predefined p-value should never by employed in a trial to dictate when it should be stopped. Stopping rules should be recommendations, nothing more, and need to be interpreted wisely. Other factors, both internal and external to the trial, should always be taken into account.

In this case, the internal factor that the treatment effect was implausibly large was given greater weight than the observed p-value. Thus, although the analyses were based on the traditional frequentist approach, their interpretation used an informal

Bayesian approach, that took account of the prior beliefs and expectations that any benefit from adding a fifth course would be, at best, moderate.

Even more important, in our view, is the appropriate use of interim analyses, particularly when clinical study design and ultimate study question are in conflict with early evaluation of results. In these situations, the importance of DSMC to make the "right" choice is clearly pivotal.

The Letrozole study (Goss *et al.*, 2003), a trial which had a considerable effect on the treatment of early-stage breast cancer, is very helpful as example to highlight these issues and therefore it will be described in details.

The study was led by the National Cancer Institute of Canada Clinical Trials Group and was a joint effort of the North American Intergroup and the Breast International Group .

The rationale of the study was based on the fact that it is accepted that in hormone-dependent breast cancer, five years of postoperative tamoxifen therapy, but not tamoxifen therapy of longer duration, prolongs disease-free and overall survival. Since the aromatase inhibitor letrozole, by suppressing estrogen production, might improve the outcome after the discontinuation of tamoxifen therapy, this double-blind, placebo-controlled trial was therefore conducted to test the effectiveness of five years of letrozole therapy in postmenopausal women with breast cancer who had completed five years of tamoxifen therapy. The primary end point was disease-free survival. A total of 5187 women were enrolled (median follow-up: 2.4 years).

At the first interim analysis, there were 207 local or metastatic recurrences of breast cancer or new primary cancers in the contralateral breast, 75 in the letrozole group and 132 in the placebo group, with estimated four-year DFS rates of 93 percent and 87 percent, respectively, in the two groups ($p < 0.001$ for the comparison of DFS).

A total of 42 women in the placebo group and 31 women in the letrozole group died (p=0.25 for the comparison of overall survival). Low-grade hot flashes, arthritis, arthralgia, and myalgia were more frequent in the letrozole group, while vaginal bleeding was less frequent. There were new diagnoses of osteoporosis in 5.8 percent of the women in the letrozole group and 4.5 percent of the women in the placebo group (p=0.07); the rates of fracture were similar.

After the first interim analysis, conducted using the Lan-DeMets alpha spending function with OBF boundaries, applied accordingly to what planned in the study protocol, the DSMC recommended termination of the trial and prompt communication of the results to the participants.

The decision, provoked a lot of reactions.

According to Bryant and Wolmark (2003), although the decision of early stopping seemed justifiable on statistical and ethical sides, on the other hand, the decision to close the study after a median follow-up of only 2.4 years, to inform all participants of the findings and the treatment they received, and to offer letrozole to the women who were originally assigned to placebo undeniably diminished the clinical usefulness of the data.

In addition, the relative reduction of 24 percent in the hazard of death from any cause in the letrozole group as compared with placebo group reduction was not statistically significant and it is possible that a survival advantage would never be documented, since ongoing follow-up is confounded by crossover. The findings cannot be useful for supporting recommendation of five years of letrozole treatment, since none of the participants have been followed up to five years, and follow-up for adverse events has been even shorter. It was also not possible to quantify the magnitude of a potential benefit with respect to disease free survival, not only because of the small number of events that have been reported to date, but also because of uncertainty about the interval for which the treatment benefit may persist.

The result had obvious implications also for concurrent trials: ongoing placebo-controlled trials of treatment with aromatase inhibitors after five years of tamoxifen therapy could be modified or terminated in response to the announcement of these study results.

Therefore, there was no opportunity to collect data from a placebo-controlled trial that would help to evaluate the risks of long-term adverse events.

As a consequence of all these issues, Bryant and Wolmark suggested that the DSMC should not generally stop a trial early except for reasons of safety, if doing so would compromise the primary aim of the trial. The same authors also suggested that the protocol document should specify a minimal level of follow-up to be completed before allowing early reporting if the reason for early reporting is efficacy and warned that this example showed that stopping rules are based on simplified models of reality and will never capture all elements of the decision-making process.

Difficulties may also arise when there is lack of proportionality of hazard, especially when the curves separate and then come together (a frequent event in non-curative treatments) and also when there is an early detriment due to toxicity or other cause but a longer-term advantage so that the curves cross over.

The importance of DSMC appears more recognised in the recent research protocols, but it is still insufficiently appreciated.

Promotion of guidelines for the structure and organisation of DSMC would be of great importance for improving both the effectiveness and efficiency of monitoring process.

For this reason, the DAMOCLES working party (Grant, 2005) addressed several different issues, using different methodological approaches: systematic literature reviews of DSMC , small group processes in decision-making; sample surveys of: reports of RCTs, recently completed and still ongoing RCTs and policies of major organisations involved in RCTs; case studies of selected DSMCs; and interviews with experienced DSMC members.

The results of these studies clearly indicated that wide variation exists in the structure and organisation of DSMCs, with little guidance on how they should o-perate. The conclusions they reached were that data monitoring should always be considered, and, differently from what is the actual current trend, reasons should be given not for justifying the presence of a DSMC, but where there is no DSMC or when any member is not independent. They also gave some practical advise for optimising the function of DSMC: for example, they stressed the importance of early DSMC meetings and of the agreement with investigators before the study initiation on roles, responsibilities and planned operations. Independence of the members and declaration of absence of conflict of interests were also recognised as important characteristic of DSMC. Finally, the primary roles of DMCS were indicated: DSMC have to ensure that continuing a trial according to its protocol is ethical, taking account of both individual and collective ethics and, in order to properly operate, the DSMC should know in advance the range of recommendations or decisions open to it.

It was also suggested that final reports should be also commented by DSMC and should include information about the data monitoring process and details on DSMC membership. The findings aided the development of a template for a charter guideline for DSMCs, whose widespread use would promote a systematic and transparent approach, and enable them to operate more effectively and efficiently.

Our research is in complete syntony with these conclusions and clearly indicates that much has still to be done for helping in the decision on the kind of statistical analyses that should be implemented, on the contribute of the results of such analyses on the final decision to be taken, and on the role of DSMC.

In our view, our results are a further contribute to the knowledge on data monitoring approaches and are of help for the identification of the questions to be addressed by further researches for improving organisation and conduction of clinical trials. In this sense, the findings from survey of Italian protocols seem of particular interest: although we are aware that they may be valid particularly in Italian research context and not totally generalizable to other countries, we think that this research represents a good point for debating the issues on how to improve monitoring of clinical trials, underlines the importance of the adoption of national registries and encourages the replication of this kind of research, even in other countries were national registries of clinical trials are available.

Chapter 6

References

Abrams, K., Ashby, D. and Errington, D. (1994). Simple Bayesian analysis in clinical trials: a tutorial. *Controlled Clinical Trials*, **15**: 349–359

Anderson, T.W. (1960). A modification of the sequential probability ratio test to reduce sample size. *Annals of Mathematical Statistics*, **31**: 165–197

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, **10**: 89–100

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association*, **58**: 365–383

Armitage, P. (1954). Sequential tests in prophilactic and therapeutic trials. *Q. J. Med.*, **23**: 255–274

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, **44**: 9–26

Armitage, P. (1958). Numerical studies in the sequential estimation of a binomial parameter. *Biometrika*, **45**: 1–15

Armitage, P. (1963). Sequential medical trials: some comments on F. J. Anscombe's paper. *Journal of the American Statistical Association*, **58**: 384–387

Armitage, P. (1967). Some development in the theory and practice of sequential medical trials. In *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, **4**: 791–804

Armitage, P. (1971). Statistical Methods in Medical Research. Oxford: Blackwell

Armitage, P. (1975). Sequential Medical Trials, 2nd ed. Oxford: Blackwell

Armitage, P. (1989). Discussion of "Interim analysis: the repeated confidence interval approach" by Jennison and Turnbull. *Journal of the Royal Statistical Society*, Series B, **51**: 334–335

Armitage, P. (1991). Interim analysis in clinical trials. *Statistics in Medicine*, **10**: 925–937

Armitage, P., McPherson, C.K., and Rowe, B.C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society*, Series A, **132**: 235–244

Barnard, G.A. (1946). Sequential tests in industrial statistics. *Journal of the Royal Statistical Society*, Suppl., **8**: 1–26

Barnard, G.A. (1949). Statistical inference (with discussion). *Journal of the Royal Statistical Society*, Series B, **11**: 115–149

Barraclough, E.D. and Page, E.S. (1959). Tables for Wald tests for the mean of a normal distribution. *Biometrika*, **46**: 169–173

Berger, J.O. (1985). Statistical decision theory and Bayesian analysis, 2nd ed. Springer, New York

Berry, D.A. (1987). Interim analysis in clinical trials: the role of the likelihood principle. *The American Statistician*, **41**: 117–122

Betensky R.A. (1997). Early stopping to accept H0 based on conditional power: approximations and comparisons. *Biometrics*, **53**: 794–806

Birnbaum, A. (1964). The anomalous concept of statistical evidence: axioms, interpretations, and elementary exposition. *Technical Report IMM*, Courant Institute of Mathematical Science, New York Science

Bryant, J. and Walmark, N. (2003). Letrozole after tamoxifen for breast cancer–what is the price of success? *New England Journal of Medicine* **19**: 1855–1857

Bross, I. (1952). Sequential medical plans. *Biometrics* **8**: 188–205

Bross, I. (1958). Sequential medical trials. *Journal of Chronic Disease* **8**: 349–365

Brown, B.W., Jr. (1983). Comments on the Dupont manuscript. *Controlled Clinical Trials*, **4**: 11–12

Canner, P.L. (1977). Monitoring treatment differences in long-term clinical trials. *Biometrics*, **33**: 603–615

Canner, P.L. (1983). Comment on "Statistical inference from clinical trials: choosing the right p-value". *Controlled Clinical Trials*, **4**: 13–17

Colton, J. (1963). A model for selecting one of two medical treatments. *Journal of the American Statistical Association*, **58**: 388–400

Cox, D.R. (1972). Regression model and life-tables (with discussion). *Journal of the Royal Statistical Society*, Series B, **34**: 187–220

Cornfield, J. (1966a). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician*, **20**: 18–23

Cornfield, J. (1966b). A Bayesian test of some classical hypotheses – with applications to sequential clinical trials. *Journal of the American Statistical Association*, **61**: 577–594

Cutler, S.J., Greenhouse, S.W., Cornfield, J., Schneiderman, M.A. (1966). The rule of hypothesis testing in clinical trials. Biometrics seminar. *Journal of Chronic Disease*, **19**: 857–882

DeMets, D.L. and Gail, M.H. (1985). Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics*, **41**: 1039–1044

DeMets, D.L. and Lan, K.K.G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, **13**: 1341–1352

DeMets, D.L. and Ware, J.H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika*, **67**: 651–660

DeMets, D.L. and Ware, J.H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika*, **69**: 661–663

Dodge, Romig (1929). *Bell System Technical Journal*

Duang-Zheng, X. (1990). Computer analysis of sequential medical trials. New York: Ellis Horwood

Dupont, W.D. (1983). Sequential stopping rules and sequentially adjusted p-values: does one require the other? *Controlled Clinical Trials*, 4: 3–10

Durrleman, S. and Simon, R. (1990). Planning and monitoring of equivalence studies. *Biometrics*, **46**: 329–336

Ellenberg, S.S., Fleming, T.R. and DeMets, D.L. (2003). Data monitoring committees in clinical trials. New York: John Wiley & Sons

Emerson, S.S. and Fleming, T.R. (1989). Symmetric group sequential test designs. *Biometrics*, **45**: 905–923

Fayers, P., Ashby, D. and Parmar, M.K.B. (1997). Tutorial in biostatistics. Bayesian data monitoring in clinical trials. *Statistics in Medicine*, **16**: 1413–1430

Feller, W.K. (1940). Statistical aspects of extra-sensory perception. *Journal of Parapsychology*, **4**: 271–298

Fleming, T.R., Harrington, D.P. and O'Brien, P.C. (1984). Designs for group sequential tests. *Controlled Clinical Trials*, **5**: 348–361

Fleming, T.R. and DeMets, D.L. (1993). Monitoring of clinical trials: issues and recommendations. *Controlled Clinical Trials*, **14**: 183–197

Freedman, L.S. and Spiegelhalter, D. J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials*, **10**: 357–367

Freedman, L.S., Spiegelhalter, D.J. and Parmar, M.K.B. (1994). The what, why and how of Bayesian clinical trials monitoring. *Statistics in Medicine*, **13**: 1371–1383

Freidlin, B., Korn, E.L. and George, S.L. (1999). Data monitoring committees and interim monitoring guidelines. *Controlled Clinical Trials*, **20**: 395–407

Gail, M.H., DeMets D.L., Slud, E.V. (1992). Simulation studies on increments of the two-sample logrank score for survival time data, with application to group sequential boundaries. In: Survival analysis, J. Crowley, R. Johnson (eds.), vol. 2. Hayward, CA: IMS Lecture Note Series.

Geller, N.L. and Pocock, S.J. (1987). Interim analyses in randomised clinical trials: ramifications and guidelines for practitioners. *Biometrics*, **43**: 213–223

Goss, P.E., Ingle, J.N., Martino, S., Robert, N.J., Musse, H.B., Piccart, M.G., Castiglione, M., Dongsheng, T., Shepherd, L.E., Pritchard, K.I., Livingston, R.B., Davidson, N.E., Norton, L., Perez, E.A., Abrams, J.S., Therasse, P., Palmer, M.J., Pater, J.L. (2003). A randomized trial of letrozole in post-menopausal women after five years of tamoxifen therapy for early-stage breast cancer. *New England Journal of Medicine*, **19**: 1793–1802

Gould, A.L. (1983). Abandoning lost causes (early termination of unproductive clinical trials). *Proc. Biopharm. Sec. ASA*, 31–24

Grant, A.M., Altman, D.G., Babiker, A.B., Campbell, M.K., Clemens, F.J., Darbyshire, J.H., Elbourne, D.R., McLeer, S.K., Parmar, M.K.B., Pocock, S.J., Spiegelhalter, D.J., Sydes, M.R., Walker, A.E., Wallace, S.A. and the DAMOCLES study group (2005). Issues in data monitoring and interim analysis of trials. *Health Technology Assessment*, **9**: 1–238

Halperin, M., Lan, K.K.G., Ware, J.H., Johnson, N.J. and DeMets, D.L. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials*, **3**: 311–323

Haybittle, J. L. (1971). Repeated assessments of results in clinical trials of cancer treatment. *British Journal of Radiology*, **44**: 793–797

Hughes M., Pocock, J. (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine*, **7**: 1231–1242

Hwang, I.K. and Shiy, W.J. (1990). Group sequential designs using a family of type I error probability spending function. *Statistics in Medicine*, **9**: 1439–1445

IMPACT Investigators (1995). Efficacy of adjuvant flurouracil and folinic acid in colon cancer. *Lancet*, **345**: 939–944

Jennison, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika*, **74**: 155–165

Jennison, C. and Turnbull, B.W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials*, **5**: 33–45

Jennison, C. and Turnbull, B. W. (1989). Interim analysis: the repeated confidence interval approach. *Journal of the Royal Statistical Society*, Series B, **51**: 305–361

Jennison, C. and Turnbull, B.W. (1990). Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statistical Science*, **5**: 299–317

Jones, D. and Whitehead, J. (1979). Sequential forms of the logrank and modified Wilcoxon tests for censored data. *Biometrika*, **66**: 105–113

Kilpatrick, G.S. and Oldham, P.D. (1954). Calcium chloride and adrenaline as bronchial dilators compared by sequential analysis. *British Medical Journal*, **ii**, 1388–1391

Kim, K., DeMets, D.L. (1987a). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, **74**: 149–154

Kim, K., DeMets, D.L. (1987b). Confidence intervals following group sequential tests in clinical trials. *Biometrics*, **43**: 857–874

Kintchine, A. (1924). Uber einen satz der wahrscheinlichbeitstrechnueg. *Fundamenta Mathematica*, **6**: 9–20

Koepcke, W., Hasford, J., Messerer, D. and Zwingers T. (1982). *Proceedings of the XI International Biometric Conference*, Toulouse, 1982

Korn, E., Freidlin, B. and George, S. (2004). Data monitoring and large apparent treatment effects. *Controlled Clinical Trials*, **1**: 67–68

Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**: 659–663

Lan, K.K.G., DeMets, D.L. and Halperin, M. (1989a). Changing frequency of interim analyses in sequential monitoring. *Biometrics*, **45**: 1017–1020

Lan, K.K.G., DeMets, D.L. and Halperin, M. (1989b). Group sequential procedures: Calendar versus information time. *Statistics in Medicine*, **8**: 1191–1198

Lan, K.K.G., Simon, R., Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics - Sequential Analysis*, **1**: 207–219

Lan, K.K.G. and Wittes, J. (1988). The β-value: A tool for monitoring data. *Biometrics*, **44**: 579–585.

Liberati, A. (1994). The relationship between clinical trials and clinical practice: the risks of underestimating its complexity. *Statistics in Medicine*, **13**: 1484–1491.

Manley, B.F.J. (1970). The choice of a Wald test on the mean of a normal distribution. *Biometrika*, **57**: 91–95.

McPherson, C.K. (1974). Statistics: the problem of examining accumulating data more than once. *New England Journal of Medicine*, **290**: 501–502

McPherson, C.K. (1977). Sequential analysis in clinical trials. Clinical Trials. Editors: J.F. Bithell and R. Coppi. Academic Press, London

McPherson, C.K. and Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society*, Series A, **134**: 15–25

Medical and Pharmaceutical Statistics Research Unit. Planning and Evaluation of Sequential Trials (PEST) 4.0. Reading: Medical and Pharmaceutical Statistics Research Unit, 2000

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**: 549–556.

Pampallona, S. and Tsiatis, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Stat. Plann. Inf.*, **42**: 19–35

Parmar, M.K.B., Spiegelhalter, D.J. and Freedman, L.S. (1994). The CHART trials: Bayesian design and monitoring in practice. *Statistics in Medicine*, **13**: 1297–1312

Pepe, M.S., Anderson, G.L. (1992). Two-stage experimental designs: early stopping with a negative result. *Applied Statistics*, **41**: 181–190

Peto, R. (1985). Discussion of "On the allocation of treatments in sequential medical trials", by A. Bather. *International Statistical Reviews*, **53**: 31–34

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, V., Mantel, N., McPherson, K., Peto, J. and Smith P.G. (1976). Design and analysis of randomised clinical trials requiring prolonged observation on each patient. I. Introduction and design. *British Journal of Cancer*, **34**: 585–612

Piantadosi, S. (1997). Clinical Trials: A Methodologic Perspective, New York, John Wiley & Sons

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**: 191–199

Pocock, S.J. (1982). Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, **38**: 153–162

Pocock, S.J. and Hughes, M.J. (1989). Practical problems in interim analyses with particular regard to estimation. *Controlled Clinical Trials*, **10**: 209S–221S

Reboussin, D.M., DeMets, D.L., Kim, K. and Gordon, L. (1996). Programs for computing group sequential boundaries using the Lan-DeMets Method, Version 2. Madison, Wisconsin: Department of Biostatistics, University of Wisconsin-Madison, Technical Report 95

Reboussin, D.M., DeMets, D.L., Kim, K. and Gordon, L. (2000). Computations for group sequential boundaries using the Lan-DeMets spending function method. *Controlled Clinical Trials*, **21**: 190–207

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, **58**: 527–535

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika*, **70**: 315–326

Slud, E.V. (1984). Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics*, **12**: 551–571

Slud, E.V. and Wei, L.J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association*, **77**: 862–868

Sohuami, R.L. (1994). The clinical importance of early stopping of randomized trials in cancer treatments. *Statistics in Medicine*, **13**: 1293–1295

Spiegelhalter, D.J., Freedman, L.S., and Parmar, M.K.B. (1993). Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine*, **12**: 1501–1511

Sydes, M.R., Altman, D.G., Babiker, A.B., Parmar, M.K.B., Spiegelhalter, D.J. and the DAMOCLES group (2004). Reported use of data monitoring committees in the main published reports of randomized controlled trials: a cross-sectional study. *Clinical Trials*, **1**: 48–59

The International Collaborative Ovarian Neoplasm (ICON) Group (2002). Paclitaxel plus carboplatin versus standard chemotherapy with either single-agent carboplatin or cyclophosphamide, doxorubicin, and cisplatin in women with ovarian cancer: the ICON3 randomised trial. *Lancet*, **360**: 505–515

The ICON and AGO Collaborators (2003). Paclitaxel plus platinum-based chemotherapy against conventional platinum-based chemotherapy in women with relapsed ovarian cancer: the ICON4/AGO-OVAR 2.2 trial. *Lancet*, **361**: 2099–2106

Tsiatis, A. A. (1981). The asymptomatic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, **68**: 311–315

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association*, **77**: 855–861

Wald, A. (1947). Sequential analysis. New York: John Wiley & Sons

Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, **43**: 193–200

Ware, J.H, Muller J.E., Braunwald, E. (1985). The futility index. An approach to the cost-effective termination of randomized clinical trials. *American Journal of Medicine*, **78**: 635–643

Wheatley, K., Burnett, A. K., Gibson, B., Clayton, V. (2002). Optimizing consolidation therapy: four versus five courses, SCT versus chemotherapy - preliminary results MRC AML12. *Haematology Journal*, **3**, S1: 159

Wheatley, K., Clayton, V. (2003). Be skeptical about unexpected large apparent treatment effects: the case of an MRC AML12 randomization. *Controlled Clinical Trials*, **24**: 66–70

Whitehead, J. (1983). The design and analysis of sequential clinical trials. 1st ed., Chichester, Ellis Horwood

Whitehead, J. (1992). The design and analysis of sequential clinical trials. 2nd ed., Chichester, Ellis Horwood

Whitehead, J. (1994). Sequential methods based on the boundaries approach for the clinical comparison of survival times. *Statistics in Medicine*, **13**: 1357–1368

Whitehead, J. (2004). Stopping rules of clinical trials. *Controlled Clinical Trials*, **25**: 69–70

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics*, **39**: 227–236

Woodroofe, M. (1992). Estimation after sequential testing: a simple approach for a truncated sequential probability ratio test. *Biometrika*, **79**: 347–353

Zaniboni, A., Labianca, R., Marsoni, S., Torri, V., Mosconi, P., Grilli, R., Apolone, G., Cifani, S., Tinazzi, A. (1998). GIVIO-SITAC 01: A randomized trial of adjuvant 5-fluorouracil and folinic acid administered to patients with colon carcinoma - long term results and evaluation of the indicators of health-related quality of life. *Cancer*, **82**: 2135–2144

Appendix A

SAS Macro routine

```
*-------------------------------------------------------------------*;
* Program  ... : SIMULA.SAS                                         *;
* Scope ...... : Computing distribution of ln(HR) at specific time  *;
                 points                                             *;
* Version  ... : 1.0                                                *;
* Author  .... : Irene Floriani                                     *;
* Date Created : 16SEP2004                                          *;
* Project .... : Statistical approach to interim analysis:         *;
                 a critical appraisal                               *;
* Macros Used. :                                                    *;
* Usage ...... :                                                    *;
*   __Example__                                                     *;
*      %SIMULA(Nsample=10000,                                       *;
*              N=2074,                                              *;
*              tasso=0.00094887,                                    *;
*              HR=0.965,                                            *;
*              recluta= 3.69884,                                    *;
*              finestudio= 7.66872,                                 *;
*              interim=323 643 965 1286,                            *;
*              zu=1.96*1.96*1.96*1.96,                              *;
*              zl=-1.96*-1.96*-1.96*-1.96,                          *;
*              output=phd.icon3_noadj)                              *;
*                                                                   *;
*-------------------------------------------------------------------*;

\%MACRO SIMULA
(NSAMPLE=                 /*Nr. of sample to be generated (dataset)   */
,RECLUTA=                 /*Accrual Time (Years)                      */
,FINESTUDIO=              /*Maximum nr. of years of follow-up         */
,HR=                      /*Hazard Ratio Treated vs Control           */
,N=                       /*Sample Size                               */
,TASSO=                   /*Daily Incidence Rate in Control Group     */
                          /*(-log[S(t)]/t with t=n.ro days            */
                          /*S(t) expected survival in control group   */
                          /*at day t and exponential risk = costant   */
                          /*time rate                                 */
,INTERIM=                 /*Periods of Interim Analysis               */
```

```
,ZU=                     /*<Description Here>                          */
,ZL=                     /*<Description Here>                          */
,OUTPUT=PARMS            /*Final Output Dataset                        */
                         /*(default=PARMS)                             */
);

%LET MACRO=SIMULA;
%LET VERSION=1.0;

***Start Macro;
%PUT (&MACRO &VERSION) Begin;

***Local Macros;
%DO S=1 %TO &NSAMPLE;

*--------------------------------------------------------------------*;
*                                                                    *;
*       Sample Generation                                            *;
*                                                                    *;
*--------------------------------------------------------------------*;

     data sample&S(keep=time entra esce status treat sample);
         attrib
            entra   length=8    label="Entry Time"
            treat   length=8    label="Arm"
            status  length=8    label="Status"
            time    length=8    label="Time";
         do i=1 to &N;

            /*uniform distribution from 0 to 365 days*/
            entra=int(ranuni(-1)*(365*&RECLUTA))+1;
            /*treatment allocation*/
            treat=ranbin(-1,1,0.5);
            /*Hazard Ratio */
            HR=&HR;
            beta=log(HR);
            /*Daily Rate*/
            lambda=(&tasso);
            /*End of study*/
            c=&FINESTUDIO*365;
            u=ranuni(-2);
            /*Event Time*/
            t=-log(u)/(lambda*(exp(beta*treat)));
            /*Time and Censoring*/
            esce=min(int(entra+t),c);
            if esce<c then status=1;
            else status=0;
            time=esce-entra;
            /*Sample*/
            sample=&S;
            output;
         end;
       run;

       proc sort data=sample&s;
         by entra;
       run;
```

```
*--------------------------------------------------------------------*;
*                                                                    *;
*       Interim analysis                                             *;
*                                                                    *;
*--------------------------------------------------------------------;

       %LET NSCAN=1;
       %LET NEVENT=%SCAN(&INTERIM,&NSCAN);
       %DO %WHILE(&NEVENT NE );
           %LET NSCAN=%EVAL(&NSCAN+1);
           %LET NEVENT=%SCAN(&INTERIM,&NSCAN);
       %END;
       %LET NPLAN=%EVAL(&NSCAN-1);
       %LET NSCAN=1;
       %LET NEVENT=%SCAN(&INTERIM,&NSCAN);
       %DO %WHILE(&NEVENT NE );
           ***Run intetim analysis;
           data sample&S&NSCAN;
                set sample&S;
                by entra;
                retain nevent 0 stop 0;
                nevent=sum(nevent,status);
                if stop eq 0 then output;
                if last.entra and nevent ge &NEVENT then stop=1;
           run;
           proc phreg data=sample&S&NSCAN outest=parms&S&NSCAN noprint;
             model time*status(0)=treat;
           run;
           ***Test if beta>z or <-z;
           %LET _ZU_=%SCAN(&ZU,&NSCAN,*);
           %LET _ZL_=%SCAN(&ZL,&NSCAN,*);
           data parms&S&NSCAN;
             set parms&S&NSCAN;
             coffu=(&_ZU_*2)/sqrt(&NEVENT);
             coffl=(&_ZL_*2)/sqrt(&NEVENT);
             hr=exp(treat);
             hr_u=exp(coffu);
             hr_l=exp(coffl);
             if treat>coffu then stop=1;
             else if treat<coffl then stop=-1;
             else stop=0;
             call symput('stop',abs(stop));
             if stop=-1 then stop=2;
           run;
           %LET NSCAN=%EVAL(&NSCAN+1);
           %LET NEVENT=%SCAN(&INTERIM,&NSCAN);
           %IF &STOP EQ 1 %THEN %LET NEVENT=;
       %END;
       ***Dataset with all interim analysis Beta estimates;
       %LET NINTERIM=%EVAL(&NSCAN-1);
       data p&S;
         set
         %DO NEV=1 %TO &NINTERIM;
             parms&S&NEV(in=in&NEV)
         %END;
         ;
         %DO NEV=1 %TO &NINTERIM;
```

```
            if in&NEV then ninterim=&NEV;
        %END;
        sample=&S;
      run;
   %END;  %***DO I=1 TO NSAMPLE;
   ***Set all estimates;
   data parms;
     set
     %DO I=1 %TO &NSAMPLE;
         P&I
     %END;
     ;
     if stop ne 0 then output;
     if (ninterim=&NPLAN)  then do;
        ninterim=99999;
        stop=99999;
        output;
     end;
   run;
   proc format;
     value stop     2='Estimated at Lower'
                    1='Estimated at Upper'
                99999='Reached final analysis';
     value ninterim 99999='Final';
   run;
   ***Stats of all estimates;
   proc means data=parms n min p5 p10 median p90 p95 max
     nway maxdec=3 noprint;
     class stop ninterim;
     var hr;
     format stop stop. ninterim ninterim.;
     label
       stop = "Boundary"
       ninterim="Nr. Analysis";
     output out=stats n=n min=min max=max
                      p5=p5 p10=p10 p95=p95 p90=p90 median=median;
   run;
   data stats;
     set stats;
     tr=&HR;
     array _stat_ (stat) min p5 p10 median p90 p95 max n;
     _id_="&OUTPUT";
     do stat=1 to 8;
        name=vname(_stat_);
        value=_stat_;
        output;
     end;
   run;
   proc sort data=stats;
     by tr _id_ stop stat name;
   run;
   ***Save Output Results;
   proc transpose data=stats out=&OUTPUT(drop=_name_) prefix=stage;
     id ninterim;
     by tr _id_ stop stat name;
     var value;
   run;
   data &OUTPUT;
     length stage1-stage%LEFT(&NPLAN) stagefinal 8;
     set &OUTPUT;
```

```
    stage='Stage';
  run;

*--------------------------------------------------------------------;
*                                                                   *;
*                         End of Macro                              *;
*                                                                   *;
*-------------------------------------------------------------------*;
  %GOTO TERM;
  %QUIT:
  %TERM:
  ***Clear the environment;
  proc sql noprint;
    drop table stats;
    drop table parms;
    %DO I=1 %TO &NSAMPLE;
        %STR(drop table sample&I;);
        %STR(drop table P&I;);
        %DO T=1 %TO &NINTERIM;
            %str(drop table parms&I&T;);
            %str(drop table sample&I&T;);
        %END;
    %END;
  quit;
  %PUT (&MACRO&VERSION) Finish;
%MEND SIMULA;
```