Interim analyses in long-term clinical trials

omslag: Inge Zorn

Print: Haveka B.V., Alblasserdam, The Netherlands.

Interim analyses in long-term clinical trials

Tussentijdse evaluaties in langdurige therapeutische experimenten

PROEFSCHRIFT

Ter verkrijging van de graad van doctor aan de

Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof. dr. C.J. Rijnvos
en volgens besluit van het college van dekanen.

De openbare verdediging zal plaatsvinden op
woensdag 13 juni 1990 om 15.45 uur

door

Gerrit-Anne van Es

geboren te Polsbroek

Promotiecommissie

Promotor:	Prof. R. van Strik
Overige leden:	Prof. dr. J. Lubsen Prof. dr. J. Pool Dr. Th. Stijnen
Co-promotor:	Dr. J.G.P. Tijssen

Dit proefschrift werd bewerkt binnen het Thoraxcentrum, Academisch Ziekenhuis 'Dijkzigt' Rotterdam.

Het verschijnen van dit proefschrift werd mede mogelijk gemaakt door steun van de Nederlandse Hartstichting.

'Decisions, although important, involve non-statistical issues and should be distinguished from purely statistical issues, which consist of asking what the data show and how certain are the conclusions they will support. Once these are known, decisions and their costs can be considered, but preferably by someone else.'

Jerome Cornfield

aan mijn ouders
aan Inge en Lonneke

CONTENTS

CONTENTS

CONTENTS

chapter 1
INTRODUCTION

In the final stage of the development of a new therapy usually large-scale comparative clinical trials are used: the experimental treatment is compared to placebo treatment. Clinical trials in which the effect of the new treatment can only be assessed over a long period of time, or in which the time period that is needed to recruit the required number of patients is long, are referred to as long-term clinical trials. Repeated inspection of the accumulating data in such long-term clinical trial is both necessary and desirable in order to detect early important benefits or unwanted side effects.

An interim analysis is the assessment of the accumulated data at a certain moment during the course of a clinical trial with respect to efficacy and safety of the compared treatments. Each interim analysis constitutes a decision problem involving prior knowledge, evolving knowledge, statistical considerations, medical judgement, and ethical principles which might lead to adaption of the protocol, to the termination of the trial, or to the extension of the trial. A statistical stopping rule, based on an estimate of the treatment effect and its precision, constitutes a useful guide in this decision proces. Various approaches to design a stopping rule are available.

After completion of a trial the results should be summarized, preferably by estimates of the treatment effects and their precision. According to classical statistical theory the usual estimation methods, which do not take into account the fact that interim analyses are performed, are invalid. Recently, some methods have been developed, which do take into account the interim analyses.

The purpose of this dissertation is to evaluate the usefulness of both stopping rules and estimation methods in long-term clinical trials with interim analyses. The ASPECT trial, a long-term clinical trial to assess the effect of anticoagulant therapy on mortality in patients after myocardial infarction which is currently conducted in the Netherlands, serves as a major example throughout this dissertation.

In this chapter a short historical background of the development of the use of interim analyses in clinical trials is followed by a description of the ASPECT trial.

HISTORICAL BACKGROUND

The first major randomized clinical trial was the British Medical Research Council (1948) trial in which the effectiveness of streptomycin in the treatment of pulmonary tuberculosis was evaluated. It was performed just after World War II and became the basis of what is now generally recognized as the correct way to evaluate new medical treatments: a well designed randomized controlled clinical trial. In the early days of clinical trials there already was the notion that at a certain point during the course of the trial sufficient evidence, in favour or in disfavour of the new treatment, might have been accumulated, such that continuation of the trial would not be ethically justified (Shaw et al., 1970). In the University Group Diabetes Program (1970a; and 1970b) and the Coronary Drug Project (1973; and 1975), clinical trials performed in the 1960s which became the model for many future clinical trials, the accumulating data were already monitored at regular time intervals.

The University Group Diabetes Program (UGDP) was a randomized, double-blind, placebo-controlled clinical trial designed to evaluate the long-term efficacy of 5 hypoglycemic agents in the prevention of vascular complications in patients with adult-onset diabetes. Results obtained from this trial were periodically evaluated for evidence of adverse or beneficial treatment effects. Extensive monitoring reports covering life-endangering conditions and death were prepared biannually and were reviewed by all participating investigators and consultants. These reports served as a basis for decision concerning modification or termination of treatments under study. Eight years after entry of the first patient the tolbutamide treatment was discontinued, mainly due to the excess in cardiovascular mortality of the tolbutamide treated group compared to the other treatment groups. The other treatments were continued.

The Coronary Drug Project was a randomized, double-blind, placebo-controlled clinical trial to evaluate the efficacy of 5 lipid-influencing drugs in the secondary prevention of coronary heart disease. In the first two years interim analyses of the data were performed at regular meetings that were open to all investigators. However, during the course of the CDP the notion evolved that knowledge with the investigators of trends in mortality, morbidity, or incidence of side-effects might result in some investigators pulling out of the trial or 'unblinding' the treatment groups prematurely (Canner, 1983). An independent Data and Safety Monitoring Committee composed of persons knowledgeable in the fields of cardiology, clinical medicine, biostatistics, epidemiology, and biochemistry was established to perform interim analyses every six months. Participating investigators (except for one) were not admitted to this committee and were kept blind of further interim results. Three of the five experimental treatments (the high dose estrogen treatment, the dextrothyroxine treatment and the low dose estrogen treatment) were

terminated early, mainly because the interim results showed a possible harmful effect of these treatments compared to placebo treatment (Coronary Drug Project Research Group, 1981).

Statistical considerations concerning the effects of early termination of the different treatment comparisons were investigated in both UGDP and CDP. A Monte Carlo monitoring procedure and a Bayesian approach were developed during the course of the trial as a means of coping with statistical problems arising from the repeated evaluation of the data. Both methods require a larger difference in the treatment groups being compared before a result is regarded 'statistically significant' than is the case when such a difference is evaluated by a conventional testing procedure.

The Monte Carlo approach is based on computer simulations of possible clinical trial outcomes, and is extensively described by Canner (1977). The Monte Carlo approach was a useful tool, when no adequate statistical theory was available yet.

A Bayesian approach introduces the prior distribution of the treatment effect under investigation, which expresses the state of knowledge before the data are obtained. The posterior distribution can be obtained from the prior distribution and the accumulated data. From this posterior distribution inferences can be made, which might lead to decisions concerning the early termination of the trial. The Bayesian approach accords to the likelihood principle, which states that all observations leading to the same likelihood function should lead to the same conclusion. This means that according to the Bayesian approach the result of the trial does not depend on the reason for stopping the trial. The Bayesian approach, strongly advocated by Cornfield (1966a; and 1966b), does not appear to have widespread use. The need to specify a prior distribution on the parameter of interest probably restrained investigators from the adaption of these methods (DeMets and Lan, 1984).

A different approach, based on the need to control the so-called type I error (i.e. the false positive rate of a statistical test), seemed to be more successful. Armitage, McPherson and Rowe (1969) showed that the repeated testing of data after every observation (pair) inflates the type I error rate. Therefore Armitage (1975) developed the so-called repeated significance methods to control for the type I error rate by lowering the conventional significance level. The statistical theory of repeated significance testing evolved from sequential methods, originating from the work of Wald (1947), and adapted for medical applications by Armitage (1957; and 1975).

In addition the CDP utilized the so-called curtailed testing procedures. Curtailed testing procedures evaluate the likelihood of reversal of currently apparent adverse or beneficial

effects of the treatment at the scheduled termination of the trial. In the CDP one of the treatments was discontinued because, '... even for the most extreme and unlikely situation, no statistically significant beneficial effect would be reached' (Canner, 1983). This procedure reflects the asymmetry of the stopping decision problem: for a beneficial treatment effect more evidence is required than for a detrimental or null treatment effect.

A simple but significant modification of Armitage's repeated significance testing model by Pocock (1977) led to the so-called group sequential approach for obtaining monitoring boundaries. The essence of the modification was to recognize that the data monitoring committees meet at scheduled intervals, at which information of recently enrolled patients is available. Instead of pairing individual patients, the group sequential approach compares groups of patients accrued in the same time interval. As Pocock also utilized a constant adjustment of the significance level for all interim analyses, O'Brien and Fleming (1979) first introduced variable adjustments. The Beta-Blocker Heart Attack Trial (1982) employed group sequential methods. The Beta-Blocker Heart Attack Trial (BHAT), a randomized double-blind placebo-controlled trial to evaluate the efficacy of propranolol in patients with a recent myocardial infarction, O'Brien and Flemings stopping rule was applied, resulting in the termination of the trial 9 months earlier than scheduled. At the sixth interim analysis the propranolol group appeared to have a 26% lower mortality rate than the placebo group, and the stopping boundaries were crossed. Of course many other issues were considered in this decision (DeMets et al., 1984).

In the last decade much work has been done in the theoretical development of the formal statistical stopping procedures. The application of (group) sequential methods to survival studies was conceived by work of Tsiatis (1982). Whitehead and co-workers (1983; and 1986), put much effort in the development of sequential methods Bayesian stopping rules were applied and advocated by Freedman (1983; and 1988) and Spiegelhalter (1988). Formal stopping criteria, evolving from these theoretical developments, have been applied in clinical trials, such as AMIS (1980), CAST (1989), AIMS (1988), ASPECT (1987), and many others.

Still rare are the applications of estimation methods following a clinical trial with interim analyses. In many clinical trials the effect estimates from the data of the trial were reported as if no interim analyses had been performed, despite the fact that it has been known for long that (group) sequential designs lead to biased estimates (Armitage, 1957). The final reports of some trials only do mention that a result should not be judged as 'statistically significant' at the 5% level unless it achieved at least the 1% ($P<0.01$) level of significance (Coronary Drug Project Research Group, 1975; and Aspirin Myocardial Infarction Study Research Group, 1980). Partly due to lack of computational

support, it was not before 1978 that a 'exact' confidence intervals following a sequential test were supported (Siegmund, 1978). Currently several estimation procedures are available for the different types of stopping rules. These methods, involving complex numerical integrations, are still under investigation.

THE ASPECT TRIAL

Cardiological studies indicate that long-term treatment with oral anticoagulants, compared to treatment with placebo, may produce a reduction in mortality after myocardial infarction of 10-20 per cent (Sixty Plus Reinfarction Study Research Group, 1980; Mitchell, 1981; and The EPSIM Research Group, 1982). However, in view of risks associated with the therapy, the role of oral anticoagulants in the secondary prevention of myocardial infarction is still under debate (Resnekow et al., 1986). The Sixty Plus Reinfarction Trial (1980), a randomized double-blind placebo-controlled trial conducted in the Netherlands, showed that old age is not a reason to discontinue oral-anticoagulant therapy. This, however, does not constitute an argument to start the treatment in the first place. Therefore the Federation of Dutch Thrombosis Services decided to conduct a clinical trial to determine whether institution of long-term treatment with oral anticoagulants immediately after acute myocardial infarction leads to a substantial decrease in total mortality. Patient recruitment of this clinical trial, called ASPECT (Anticoagulants in the Secondary Prevention of Effects in Coronary Thrombosis), started in april 1986. The trial was originally planned to end in summer 1989, but due to slow patient entry the trial was prolonged until june 1992.

Design

The primary objective of the ASPECT trial is to determine whether institution of long-term treatment with oral anticoagulants immediately after acute myocardial infarction leads to a substantial decrease in total mortality. The ASPECT trial is designed as a double blind, placebo controlled, randomized clinical trial. Patients who have suffered from acute myocardial infarction are eligible for entry within 2 to 6 weeks after onset, provided they satisfy the inclusion criteria and no exclusion criterion applies. Experimental treatment consists of the anticoagulant drugs acenocoumarin and phenprocoumarin and their matching placebos. The outcome measure to compare the two treatment regimens is all cause mortality over a period of at least one year. A total of 4000 patients was planned to be admitted to the trial over a period of two years. With 2000 patients in both treatment groups, at an event rate of 10% in the placebo treated patients and a reduction of 25% a power of about 85% will be obtained. Follow-up will be continued until one year after admission of the last patient, thus the total duration of the

trial was scheduled at three years. All policy decisions during the course of the trial are made by an independent Policy Board of experts in cardiology, haemostasis, epidemiology, and biostatistics.

The Data Monitoring Committee

According to the protocol (ASPECT Policy Board, 1986) the Data Monitoring Committee (DMC), which is a subcommittee of the Policy Board, was supposed to meet at least twice a year, that is six times in three years, to assess the interim results of the trial. The DMC consists of four permanent members representatives from the following fields: pharmacology, statistics, epidemiology, and internal medicine. An independent statistician (the author of this dissertation) will supply the unblinded data for the DMC. The committee monitors the data for early evidence regarding efficacy and safety. With regard to the efficacy of the treatments concerning overall mortality the DMC acts according to a pre-defined statistical stopping rule (van Es, 1987), specifying when to recommend the Policy Board to terminate the trial. It was appreciated that the stopping rule was not employed as an absolute criterion in deciding whether to continue the trial or not; such a decision requires the weighing of other relevant issues, as mentioned before.

Conduct and progress

Patient recruitment of the ASPECT trial was planned to be completed in 2 years. Interim analyses were planned every six months, and accordingly a stopping rule was defined. However, due to a much lower patient accrual rate than anticipated, the patient entry period was prolonged until the end of 1991. As a consequence interim analyses take place once a year; the stopping rule was adapted to these circumstances. Until now 4 interim analyses have taken place. The unblinded data used for these interim analyses are still confidential.

In april 1986 the first patients were admitted to the ASPECT trial. The progress of the trial at the interim analyses is shown in table 1.1. Due to the slow patient accrual (only 540 patients had entered the trial in the first 13 months) the first interim analysis was not performed until 18 november 1987. The interim results concerned all patients admitted to the trial before 1 may 1987; follow-up of the patients was completed until 30 june 1987. Also based on these results the DMC recommended the Policy Board to continue the trial, although it spoke out its concern about the progress of the trial, with respect to (very) low patient accrual rate.

At the next interim analysis, performed on 6 february 1988, data of 867 patients that

Table 1.1: Progress of the ASPECT trial at the first four interim analyses

interim analysis	maximal follow-up (in months)	number of patients	number of person years	number of deaths
1	15	540	350	16
2	22	867	701	24
3	34	1576	1944	67
4	46	2200	4000	130

entered the trial before 1 december 1987, and with follow-up until 31 december 1987 were available. There was an increase of the patient accrual rate. However, the target of 4000 patients in september 1988 was beyond reach, and it was decided to extend the patient recruitment until the end of 1991; patient follow-up was extended until june 1992.

At the third interim analysis data on 1576 patients were reported to be randomized before 1 december 1988 (with follow-up until 31 december 1988). Apart from the usual considerations, the DMC also regarded the preliminary results that were available on the WARfarin re-Intervention Study (Smith et al., 1989), a clinical trial conducted in Norway with similar objectives as ASPECT. The WARfarin re-Intervention Study (WARIS) showed a beneficial result of anticoagulant treatment relative to placebo, with respect to mortality and re-infarction in a group of post-MI patients. The ASPECT trial was continued.

On request of the DMC the next interim analysis will concern the data of patients that were randomized before 1 december 1989 with follow-up until 31 december 1989. The DMC meeting was in april 1990. The number of patients in this interim analysis was 2200, the number of deaths was 130. For the final analysis, which will be carried out after june 1992, 3500 patients are expected to be admitted to the trial. Based on this total number of patients and based on a constant hazard rate the expected number of deaths is 350.

REFERENCES

AIMS Trial Study Group. Effect of intravenous APSAC on mortality after acute myocardial infarction: preliminary report of a placebo-controlled clinical trial. *Lancet,* 545-549, 1988.

Armitage P, McPherson CK, Rowe BC. Repeated significance testing on accumulating data. *Journal of the Royal Statistical Society A,* 132, 235-244, 1969.

Armitage P. *Sequential medical trials.* New York: Wiley, 1957.

Armitage P. *Sequential medical trials (second edition).* New York: Wiley, 1975.

ASPECT Policy Board. *Anticoagulants in the Secondary Prevention of Events in Coronary Thrombosis: ASPECT study (protocol).* Rotterdam, august 1986.

Aspirin Myocardial Infarction Study Research Group. Randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *Journal of the American Medical Association,* 243, 661-669, 1980.

Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction: I. Mortality results. *Journal of the American Medical Association,* 247, 1707-1714, 1982.

Canner PL. Monitoring treatment differences in long-term clinical trials. *Biometrics,* 33, 603-615, 1977.

Canner PL. The Coronary Drug Project. Monitoring of the data for evidence of adverse or beneficial treatment effects. *Controlled Clinical Trials,* 4, 467-483,1983.

CAST investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppressed after myocardial infarction. *New England Journal of Medicine,* 321, 406-412, 1989.

Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *American Statistician,* 20, 18-23, 1966a.

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

Coronary Drug Project Research Group. The Coronary Drug Project. *Journal of the American Medical Association,* 226, 652-657, 1973.

Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *Journal of the American Medical Association,* 231, 360-381, 1975.

Coronary Drug Project Research Group. Practical aspects of decision making in clinical trials: the Coronary Drug Project as a case study. *Controlled Clinical Trials,* 1, 363-376, 1981.

DeMets DL, Hardy R, Friedman LM, Lan KKG. Statistical aspects of early termination in the beta-blocker heart attack trial. *Controlled Clinical Trials,* 5, 362-372, 1984.

DeMets DL, Lan KKG. An overview of sequential methods and their application in clinical trials. *Communications in Statistics Theory and Methods,* 13, 2315-2338, 1984.

EPSIM Research Group. A controlled comparison of aspirin and oral anticoagulants in prevention of death after myocardial infarction. *New England Journal of Medicine,* 307, 701-708, 1982.

Es GA van, Tijssen JGP, Lubsen J, Strik R van. Early termination of clinical trials with prolonged observation of individual patients: a case study. *Statistics in Medicine,* 6, 927-937, 1987.

Freedman LS. A comparison of Bayesian with frequentist methods of monitoring clinical trials data. *Proceedings of the XIVth International Biometric Conference,* 57-65, 1988.

Freedman LS, and Spiegelhalter DJ. The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician,* 32, 153-160, 1983.

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal,* 2, 769-782, 1948.

Mitchell JRA. Anticoagulants in coronary heart disease: retrospect and prospect. *Lancet,* 257-262, 1981.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics,* 35, 549-556, 1979.

Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika,* 64, 191-199, 1977.

Resnekow L, Chediak J, Hirsh J, Lewis D. Antithrombotic agents in coronary artery disease. *Archives of Internal Medicine,* 146, 469, 1986.

Shaw LW, Chalmers TC. Ethics in cooperative clinical trials. Ann of NY Acad Sci, 169, 487-495, 1970.

Siegmund D. Estimation following sequential tests. *Biometrika,* 65, 341-9,1978.

Sixty Plus Reinfarction Study Research Group. A double blind trial to assess long-term oral anticoagulant therapy in elderly patients after myocardial infarction, *Lancet,* II, 989-994, 1980.

Smith P, Arnesen H. Oral anticoagulants reduce mortality, reinfarction and cerebrovascular events after myocardial infarction - WARIS study. *European Heart Journal (abstact supplement),* 10, 264, 1989.

Spiegelhalter DJ, and Freedman LS. Bayesian approaches to clinical trials. *Bayesian Statistics* 3, 453-477, 1988.

Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association,* 77, 855-861, 1982.

University Group Diabetes Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: I. Design, methods and baseline results. *Diabetes,* 19, 747-785, 1970a.

University Group Diabetes Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: II. Mortality results. *Diabetes,* 19, 787-826, 1970b.

Wald A. *Sequential analysis.* New York: Wiley, 1947.

Whitehead J. *The design and analysis of sequential trials.* Chichester: Ellis Horwood, 1983.

Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika,* 73, 573-581, 1986.

chapter 2
LONG-TERM CLINICAL TRIALS

INTRODUCTION

A clinical trial is a scientific experiment with patients as subjects. Its goal is to learn about the effect of one or more therapeutic interventions for a certain disease. Long-term clinical trials are clinical trials with a relatively long time interval between entry of the first patient until the end of the follow-up. In certain indications clinical trials require long-term follow-up of each patient to evaluate the efficacy of the treatments under comparison (Peto et al., 1976 and 1977). Furthermore, clinical trials may involve long admission periods lasting for several months or years to obtain the required number of patients.

In this chapter some general principles of clinical trials in general, and of long-term clinical trials in particular, are described. From an initial idea about a possible improvement in treatment the objectives of a clinical trial are put forward. The next step is to design the trial, such that the objectives are met. After the conduct, i.e. the execution and data collection, of the trial the results are summarized, i.e. data analysis. The final step is to report the results and to draw conclusions. The specific aspects concerning interim analyses will only be briefly mentioned in this chapter; they will be discussed more extensively in the following chapters.

OBJECTIVES

In clinical practice it is not sufficient just to know that the therapeutic intervention 'works' or 'does not work'; treatment decisions depend on weighing the magnitude of the expected therapeutic benefit against possible adverse effects and against costs and time. Consequently, a clinical trial should be regarded as a means to measure a treatment effect and not as a means to determine whether or not a treatment effect exists. The objective of a clinical trial should accordingly be specified in quantitative terms. This objective is characterized by three elements: 1. the disease entity for which treatment effect assessment is required; 2. the treatments to be compared; and 3. the outcome (Tijssen and Lubsen, 1987a). These three elements can be clearly distinguished in the objective of the ASPECT trial: to determine whether institution of

long-term treatment with oral anticoagulants relative to placebo (2) immediately after myocardial infarction (1) leads to a substantial decrease in total mortality (3). Note that the treatment effect is defined in comparative terms.

Disease entity

In the design phase of a clinical trial the inclusion and exclusion criteria are the defining characteristics of the disease entity to be studied. The inclusion and exclusion criteria by themselves, however, do not define a group of patients. In order to obtain a group of patients to represent this disease entity the investigators must devise a appropriate recruitment scheme. The recruitment scheme ties the selection of patients to a particular place (hospital, city, country) and to a particular time (period of patient accrual). Therefore, after completion of the trial, a precise description of the characteristics of the patients at entry into the trial, usually referred to as the baseline characteristics, and the clinical course in the reference group contain essential additional information concerning the disease entity of interest.

Treatments

In a clinical trial the interest usually centers around one specific treatment. In many instances the treatment of interest is a drug, but also other interventions such as surgical treatments or therapeutic strategies may be studied. The treatment of interest, called the index treatment, may be contrasted to placebo treatment, a standard treatment, or the absence of treatment. The choice of this reference treatment depends on the particular situation and is also guided by considerations of internal validity.

Outcome

A treatment effect in an individual patient is measured in terms of a disease outcome. A disease outcome reflects that aspect of the clinical course that the treating physician intends to influence in that patient. In long-term clinical trials the disease outcome of interest usually is an untoward clinical event which may or may not occur in a certain period of time. The proportion of patients in a certain group of patients who have developed the outcome event then characterizes the occurrence in the whole group. If the disease outcome is measured on a quantitative scale (e.g. blood pressure) the outcome on group level is the mean (or median) of the individual observations.

DESIGN

The overall goal of a clinical trial is to estimate the effect of treatment with as little error as possible. Error in measurement may be classified as either random or systematic. The internal validity of a clinical trial reflects the lack of systematic error in the estimate of the treatment effect; precision pertains to the degree of lack of random error. The principles of study design emerge from consideration of approaches to reducing both types of measurement error.

Internal validity

Internal validity implies unbiased effect measurement within the context of the clinical trial itself. Internal validity is based on the following three issues (Miettinen, 1985):
1. *Comparability of extraneous effects* concerns the question as to whether the group comparison truly reflects the treatment comparison of interest. In the assessment of a new treatment, the attention is usually focused on the new treatment itself. Possible placebo effects, however, must be considered extraneous to the comparison at issue. The conventional method to solve this problem is the use of placebo treatment for the reference treatment. 2. *Comparability of prognosis* pertains to the 'scientific ideal' (Vandenbroucke, 1986) to control all factors which might influence the observation one wants to make. This ideal cannot be achieved in clinical trials; the best alternative to accomplish comparable treatment groups with respect to prognosis at entry in the trial is the use of random treatment assignment. 3. *Comparability of information* concerns the principle that the disease outcomes in individual patients must be obtained in a way that is identical for the treatment groups, especially in trials where the disease outcome is subject to interpretive observation. The conventional way to assure comparability of information is to collect the data unaware of the treatment allocation. The use of blinding may also be indicated to maintain comparability of extraneous effects.

Precision

Observed effects in a clinical trial are subject to random error. Whether the sources of variation that we cannot explain are actually due to chance or not makes little difference: we treat such variation as being due to chance (Rothman, 1986). Precision relates to the magnitude of random variation in the observed effect estimate. The primary means to reduce random error in the observed treatment effect, i.e. to increase precision, is to enlarge the size of the trial. The size of a clinical trial is usually determined to achieve sufficient precision for a certain expected clinically relevant effect. A usual approach for assessing the desirable size, and thereby the precision, of the trial is the use of power calculations. An alternative approach to estimating

precision of a trial is to postulate the study data and calculate the precision of the effect estimate as in data analysis, by using confidence intervals (Rothman, 1986). Except for the level of confidence, this approach requires assumptions about the magnitude of the treatment effect. If halfway in the trial the treatment effect turns out to be much greater than expected, less precision might be required, and hence less patients are required. In that case the trial might be terminated before its planned end. Statistical rules to support the eventual decision of stopping the trial early, involving the magnitude and the precision of the effect estimate during the course of the trial, are known as stopping rules. The design of such stopping rules is described in chapter 3 of this dissertation. A more extensive discussion of internal validity and precision is provided in the text books by Pocock (1983) and Rothman (1986).

CONDUCT

A clinical trial is intended to provide accurate assessment of the treatment effect while ensuring that each patient's individual needs are cared for. The design of a clinical trial needs to fulfil scientific, ethical, and organizational requirements so that the trial may be conducted efficiently. The end-result of the design of a clinical trial is a written documentation of the trial plan and is called the protocol. The protocol documents, in all relevant detail, the procedures in the trial (Spilker, 1984). Thus, its essence is stipulation of the procedures. In this context the case record form is part of the protocol. To be effective, the protocol also provides for an understanding of the nature of the intended procedures. Therefore the protocol should start by stating the objective of the trial.

The use of the accumulating information in long-term clinical trials is called monitoring and is an essential element in the conduct of a clinical trial. Monitoring without knowledge of the treatment allocation and monitoring with knowledge of the treatment allocation are distinguished (Enas et al., 1989).

The process of scrutinizing the logistics of a clinical trial so that the intended plan of the trial protocol realized is can be performed in a completely blinded manner. It embraces issues related to protocol compliance and individual patient safety. It should be performed in a deterministic manner and it should be clearly spelled out in the protocol.

The regular assessment of treatment difference, requiring the unblinded data, is of crucial importance if clinical trials are to be ethically acceptable. Investigators have the responsibility to notify patients as well as the medical community of the better

treatment once the choice is clear. In addition, one wishes to be efficient in the sense of avoiding undue prolongation of a trial once the main treatment comparisons are reasonably clear-cut. This proces of monitoring for efficacy and safety is called interim analysis.

Interim analyses are usually performed by an independent Data Monitoring Committee (DMC) that meets periodically to review the unblinded data as it accumulates. A DMC consists of experts from different fields, such as clinicians, epidemiologists, statisticians, and ethicists. To obtain a full understanding of the treatment effect(s) under study a complete statistical report should be prepared for the DMC, containing information regarding both efficacy and safety of the treatments under study (Meinert, 1986). On the basis of this information and all other external information concerning the experimental treatments it can be decided whether to stop or to continue the trial. General guidelines supporting this decision should be formulated in the protocol. With respect to the statistical considerations a more explicit decision rule should be formulated in the protocol. This decision rule is generally referred to as a stopping rule. Stopping rules are extensively discussed in the chapters 3 and 4 of this dissertation.

DATA ANALYSIS

The purpose of data analysis is to summarize the results of a particular trial, thereby enabling readership to draw their conclusions from the trial independently.It must be stressed that the purpose of a data analysis cannot be taken as that of reaching a conclusion about the investigated treatment effect. The eventual result of a clinical trial, a view (opinion) about this treatment effect, will not be based on the findings of the trial alone.

The summarization of the evidence consists of (1) a description of the design of the trial, (2) a summary presentation of the observations themselves, and (3) estimates of the treatment effect(s) in combination with an appropriate indication of its precision (Tijssen and Lubsen, 1987b). In chapter 5 of this dissertation effect estimation methods in clinical trials with interim analyses are described. With regard to effect estimation of a fixed size trial the so-called crude and stratified estimation procedures are distinguished. Crude (or un-stratified) estimation procedures apply, when it is not necessary to take into account any factors beyond the exposure of interest. Although it is not unusual to see the effect estimates presented solely in this form, the investigators should explore the data using multivariate methods. This dissertation focusses on crude estimation procedures.

REFERENCES

Enas GG, Dornseif BE, Sampson CB. Monitoring versus interim analysis of clinical trials: a perspective from the pharmaceutical industry. *Controlled Clinical Trials,* 10, 57-70, 1989.

Meinert CL. *Clinical trials: design, conduct, and analysis.* New York: Oxford University Press, 1986.

Miettinen OS. *Theoretical epidemiology.* New York: Wiley, 1985.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. introduction and design. *British Journal of Cancer,* 34, 585-612, 1976.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II. analysis and examples. *British Journal of Cancer,* 35, 1-39, 1977.

Pocock SJ. *Clinical trials: a practical approach.* Chichester: Wiley, 1983.

Rothman KJ. *Modern epidemiology.* Boston: Little Brown and Company, 1986.

Spilker B. *Guide to clinical studies and developing protocols.* New York: Raven Press, 1984.

Tijssen JGP, Lubsen J. Principles of intervention research. *European Heart Journal (supplement H),* 8, 17-22, 1987a.

Tijssen JGP, Lubsen J. Data analysis. *European Heart Journal (supplement H),* 8, 49-69. 1987b.

Vandenbroucke JP. Is the randomized controlled trial the real paradigm in epidemiology? [letter]. *Journal of Chronic Disease,* 39, 572-574, 1986.

chapter 3

OBJECTIVES AND DESIGN

INTRODUCTION

The general objective of a clinical trial is to measure a specified treatment effect with a minimum of systematic error (or bias) and with sufficient precision. The unbiasedness of the measured effect depends on the internal validity of the trial, and must be anticipated in the design of the trial. The precision of the measured effect can be improved by increasing the number of subjects in the trial. The gain in precision by enlarging the trial, however, should be weighed against expending greater efforts. A formal analysis of this cost-benefit problem is not feasible; only rough estimates as to a cost-efficient size for a clinical trial are feasible (Tijssen and Lubsen, 1987).

In many instances, however, ethical restrictions to the size of the trial are more important than financial and other practical restrictions. Investigators have the ethical obligation to notify the patients as well as the medical community of the preferred course of treatment once the choice is clear, especially when the disease under study is life-threatening. If, for example, a clinical trial reaches a conclusive result half-way and this is (or can be) known by the investigators, it is unethical to continue the trial until its planned end.

The conclusiveness of a trial result depends on the observed treatment effect and its precision in relation to the so-called minimal clinically relevant effect (E_{rel}), which is defined as the minimal effect of the index treatment relative to the reference treatment such that clinicians would be inclined to change their treatment policy from the reference treatment to the index treatment. The amount of precision that is required to reach a conclusive result depends on the treatment effect as it is expected by the investigators relative to this minimal clinically relevant effect. More precision is needed for an effect close to E_{rel} than for an effect far away from it. In situations that the investigators are convinced that the magnitude of the treatment effect is much bigger than E_{rel}, a smaller trial with less precision might suffice. In the following it is assumed that a clinical trial is designed to achieve sufficient precision given E_{rel}.

In a clinical trial with staggered patient entry and with long-term patient follow-up data slowly accumulate over time. This provides an opportunity for the investigators to

monitor the data: at every interim analysis estimates are available of the magnitude and the precision of the treatment effect so far. If the interim analysis suggests a beneficial effect close to E_{rel}, the trial should continue to obtain its planned precision. On the other hand, when interim analysis suggests a much greater beneficial effect than E_{rel}, its less precision is required and the trial might be stopped before its planned end. The same applies if an interim analysis suggests an effect in the opposite direction, an effect that might even be harmful. The monitoring of the data can be seen as an ethical safeguard against continuing a trial too long, when the interim results indicate a treatment effect that substantially differs from what was expected by the investigators (also laid down in the original trial size determination).

The main purpose of this chapter is to describe the so-called stopping rules, which are statistical decision rules that may serve as an aid in the above mentioned process of monitoring the data. Before that, some measures of treatment effect are described for clinical trials with long-term follow-up, and the concept information time is introduced.

MEASURES OF TREATMENT EFFECT

In clinical trials with long-term follow-up, the outcome of interest usually is an untoward clinical event (e.g. death) which may or may not occur in a certain period of time. In the following, without loss of generality, the clinical event of interest is denoted as death. In this section some measures of treatment effect are described. A detailed discussion about the choice of the correct effect measure is beyond the scope of this thesis, and it is referred to text books (Pocock, 1983; Miettinen, 1985; and Rothman, 1986).

Cumulative incidence

The proportion of patients who acquire a certain disease outcome in a stated period of time is called the cumulative incidence rate. Like any proportion, the value of cumulative incidence rate ranges from zero to one, and is dimensionless. It is uninterpretable, however, without specification of the time period to which it refers. An absolute effect measure in terms of the cumulative incidence rate is the rate difference (RD) which is the difference of the cumulative incidence rates in the index group and in the reference group. A relative effect measure in terms of cumulative incidence rate is the rate ratio (RR) which is the ratio of the cumulative incidence rates in the two treatment groups.

The data from a clinical trial measuring cumulative incidence rates is summarized in table 3.1. Outcome event occurrence at the treatment group level is measured as the occurrence rates in both treatment groups: the index rate r_1 equals a/n_1, and the

Table 3.1: A summary of the results of a clinical trial measuring cumulative incidence rate of a certain disease outcome.

Treatment	Disease outcome yes	no	total	rate
Index	a	c	n_1	$r_1 = a/n_1$
Reference	b	d	n_0	$r_0 = b/n_0$

reference rate r_0 equals b/n_0. The estimated rate difference rd is the difference of the rates in the index group and the reference group, respectively

$$rd = r_1 - r_0.$$

The distribution of the statistic rd is approximately Gaussian with mean RD, and with estimated variance

$$Var(rd) = \frac{r_1(1-r_1)}{n_1} + \frac{r_0(1-r_0)}{n_0}.$$

The estimated rate ratio rr is the ratio of the two rates

$$rr = \frac{r_1}{r_0}.$$

The distribution of the statistic ln(rr) is approximately Gaussian with mean ln(RR), and with estimated variance

$$Var[ln(rr)] = \frac{(1-r_1)}{r_1 n_1} + \frac{(1-r_0)}{r_0 n_0}.$$

Incidence density

The number of disease onsets in a certain population divided by the sum of the time periods of observation for all individuals in that population is called the incidence density rate. Because the incidence density rate is a quotient with a frequency in the numerator and an amount of time in the denominator, its dimensionality is time^{-1}, that is, the reciprocal of time. If the risk of death (i.e. the hazard) over time is approximately constant in both treatment groups, the incidence density rate is a suitable measure of disease frequency. An absolute effect measure in terms of incidence density rates is the

Table 3.2: A summary of the results of a clinical trial measuring incidence density rate of a certain disease outcome.

Treatment	Disease outcome Yes	person-time	rate
Index	a	T_1	$r_1 = a/T_1$
Reference	b	T_0	$r_0 = b/T_0$

incidence density rate difference (IRD) which is the difference of the incidence density rates in the index group and in the reference group. A relative effect measure in terms of incidence density rates is the incidence density rate ratio (IRR) which is the ratio of the incidence density rates in the two treatment groups.

The data of a clinical trial measuring incidence density rates can be summarized as in table 3.2. Outcome event occurrence at the treatment group level is measured as the disease incidence rates in both treatment groups: the index rate r_1 equals a/T_1, and the reference rate r_0 equals b/T_0. The estimated incidence density rate difference ird is the difference of the rates in the index group and the reference group, respectively

$$ird = r_1 - r_0.$$

The distribution of the statistic ird is approximately Gaussian with mean IRD, and with estimated variance

$$Var(ird) = \frac{r_1}{T_1} + \frac{r_0}{T_0}.$$

The estimated incidence density rate ratio irr is the ratio of the two rates

$$irr = \frac{r_1}{r_0}.$$

The distribution of the statistic ln(irr) is approximately Gaussian with mean ln(IRR), and with estimated variance

$$Var[ln(irr)] = \frac{1}{a} + \frac{1}{b}.$$

Table 3.3: A summary of the results in the time interval $(t_{i-1}, t_i]$ of a clinical trial, measuring instantaneous incidence density of a certain disease outcome.

Treatment	Disease outcome yes	no	at risk
Index	o_{i1}	$n_{i1}-o_{i1}$	n_{i1}
Reference	o_{i0}	$n_{i0}-o_{i0}$	n_{i0}
Total	o_i	n_i-o_i	n_i

Proportional hazards

The hazard is the instantaneous death rate in a short period of time. If the hazard can be assumed constant over time in both treatment groups, the incidence density rate is an appropriate measure of disease frequency. Frequently the hazard is not constant over time, thereby invalidating the incidence density rate as measure of disease frequency. This would be the case in a clinical trial if, for example, the hazard decreases as the time since exposure increases, or vice-versa. In many survival trials, however, the ratio of the hazards in both treatment groups is approximately constant. In this case the hazard ratio (HR) is a suitable effect measure. No 'direct' estimate of HR is available, but the logrank statistic is an efficient summary of the data (Kalbfleiss and Prentice, 1980) from which the hazard ratio (hr) can be estimated approximately. This approximation is based on the asymptotic Gaussian distribution of the logrank statistic.

At a particular moment in the trial d patients have died, m_1 on the index treatment and m_0 on the reference treatment. The times since randomization of all these deaths are known. The distinct values taken by uncensored survival times will be denoted $t_1, \ldots, t_k$, ranked in ascending order; t_1 being the smallest value. The data from the time-interval $(t_{i-1}, t_i]$ are summarized in table 3.3. The number of patients that died at t_i will be denoted by o_i $(i=1, \ldots, k)$. The number of patients at risk of dying in the time interval just before t_i will be denoted by n_i. Of these n_i patients, n_{i1} are on the index treatment and n_{i0} are on the standard. Let $A_{i1} = n_{i1}/n_i$ be the proportion of patients at risk who are on the index treatment, and $A_{i0} = n_{i0}/n_i$ be the proportion of patients at risk who are on the reference treatment. With these definitions the logrank statistic can be calculated:

$$s_{LR} = m_1 - \sum_{i=1}^{k} o_i A_{i1} .$$

The distribution of the logrank statistic s_{LR} is approximately Gaussian with mean S_{LR}, and with estimated variance

$$Var(s_{LR}) = \sum_{i=1}^{k} \frac{o_i(n_i-o_i)}{n_i-1} A_{i1} A_{i0} .$$

As long as the number of patients at risk in each treatment group remain fairly balanced, i.e. the hazard ratio is close to one, the s_{LR} is approximately Gaussian distributed with mean (d/4)ln(HR) and with sample variance d/4. In this case the logarithm of the hazard ratio can be estimated from the quotient of s_{LR} and $Var(s_{LR})$

$$\ln(hr) = \frac{s_{LR}}{Var(s_{LR})},$$

with an approximate variance of 4/d (Jennison and Turnbull, 1984).

INFORMATION TIME

A clinical trial can be seen as an exercise to gather information on a certain treatment effect. In a clinical trial with staggered patient entry and with long-term follow-up, information accumulates over time. Usually, the progress of the trial is measured in terms of calendar time. However, in nearly all cases the amount of information does not accumulate regularly over time. In the ASPECT trial information is approximately proportional to the number of deaths (Tsiatis, 1982). Figure 3.1 shows the number of patients admitted to the trial and the number of deaths in terms of calendar time. Patient entry started in april 1986 (see chapter 1). It can be seen that the accumulation of information (i.e. the number of deaths) is relatively slow in the beginning of the trial. Only few patients have entered, and have been at risk only for a short period of time. Information accumulates faster as the trial progresses and more patients have entered the trial and are on the average longer at risk.

In this thesis the progress of a clinical trial is reported in terms of information time instead of calendar time. Information is quantified by the reciprocal of the estimated variance of the effect estimate. This definition of the amount of information is similar to the quantification of the precision. Information and precision, which both depend on the effect measure, increase if the number of patients increases. At the i-th interim analysis (during the course of the trial) the amount of information I_i accumulated so far can be estimated. Define I_T as the total amount of information the trial should have at the planned end, then $t_i = I_i / I_T$ represents the proportion of the planned total of information observed at the i-th interim analysis. This proportion t_i is called information time (Lan et al., 1984). Information time is a monotone function of calendar time; at the start of the trial the information time is 0; at the end of the trial the information time is 1. It is clear from the above example there is not necessarily a linear relationship between information time and calendar time.

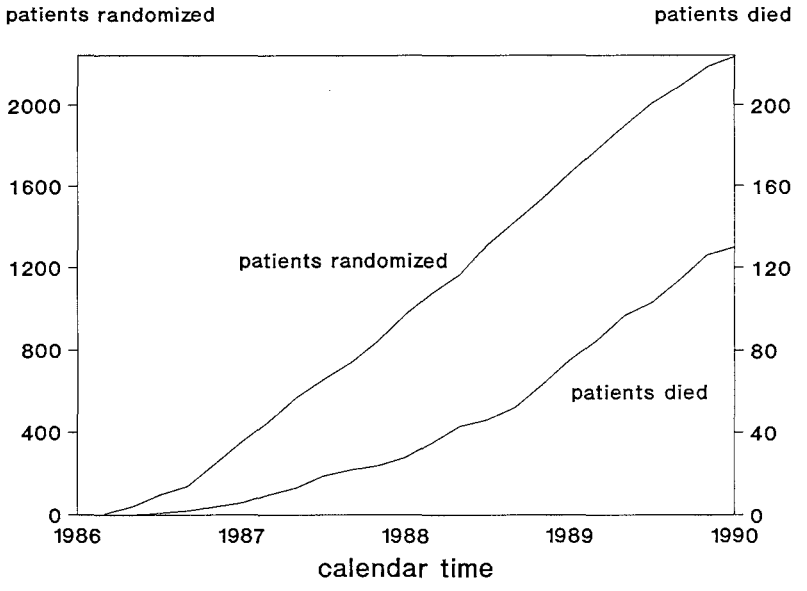


Figure 3.1: The number of patients that entered the ASPECT trial until 1 december 1989, and the number of patients that died until 1 january 1990 since the start of the trial.

In a survival trial information time at a certain moment can be estimated by the number of patients died before that moment divided by the total number of deaths at the end of the trial. The total number at the end of the trial is unknown, however, unless the design of the trial requires trial continuation until D events have occurred. If D is unknown its value must be estimated (DeMets and Lan, 1984). An overestimate of D will result in an underestimate of t_i. During the course of the trial the estimate of D can be updated.

The planned times of interim analyses in the ASPECT trial were based on the expected number of deaths in the placebo group at the respective interim analyses. The total number of deaths in the placebo group was estimated to be 200, the numbers of placebo deaths at the respective interim analyses were estimated to be 15, 40, 80, 120, and 165 (ASPECT Policy Board, 1986). Accordingly, the respective moments of the planned interim analysis were $t_1 = 0.075$, $t_2 = 0.20$, $t_3 = 0.40$, $t_4 = 0.60$, and $t_5 = 0.825$.

During the course of the ASPECT trial the total number of deaths from both treatment groups were used to estimate the information times. The total number of deaths was estimated to be 350. At the fourth interim analysis, after a modification of the protocol concerning patient accrual and patient follow-up (see chapter 1), D=350 was still a good estimate. However, the planned times of interim analyses have changed. Using the total

number of deaths at the first four interim analyses as shown in table 1.1 the estimated information times of these analyses are 0.05, 0.07, 0.19, and 0.37 respectively. The differences between the planned and the actual moments of interim analysis (especially in the first two interim analyses not much information has accumulated yet)are due to the slow patient accrual and the low overall event rate as described in chapter 1.

STOPPING RULES

At any moment during a clinical trial an interim analysis can be performed. A stopping rule is defined as an explicit decision rule, whether to continue or to stop a clinical trial, that is based on the data collected so far. In literature one will find stopping rules that are defined in terms of different summary functions of the data (i.e. statistics), such as the P-value, the Z-value, or the effect estimate. A general definition of a stopping rule is:

> At the i-th interim analysis (i=1, .., k-1) a function of the data, the statistic S_i, is calculated. A stopping rule defines the critical boundaries a_i and b_i such that the trial stops when
>
> $$S_i > a_i,$$
>
> or
>
> $$S_i < b_i,$$
>
> with $a_i > b_i$. Otherwise, if the boundaries are not crossed, the trial will be continued. If the trial is not stopped at one of the interim analyses, the trial will be stopped at its planned end.

The series of a_i and b_i (i=1, .., k-1) are called the upper and the lower boundary respectively. It is not required that the total number of analyses k is specified in advance.

The stopping boundaries for the ASPECT trial were defined in terms of (one-sided) P-values. Five interim analyses were planned to take place every 6 months after start of the trial. The lower boundaries were 0.005, 0.005, 0.005, 0.014, and 0.023 at the respective interim analyses, and the critical boundary at the final evaluation was 0.032. The upper boundaries were 0.95, 0.88, 0.81, 0.74, and 0.76, respectively. If one of the lower boundaries is crossed during the trial the data indicate a positive effect of oral anticoagulant therapy and the trial will be stopped; if one of the upper boundaries is crossed the trial will be stopped because the data indicate a negative or no effect of oral anticoagulant therapy (van Es et al., 1987). For a more detailed description one is referred to chapter 4 of this thesis.

Any choice of the boundaries according to the above definition defines a stopping rule. It goes without saying that these boundaries are not arbitrarily chosen. The choice boundaries depends on the specific circumstances of a particular trial and should reflect the considerations of the investigators involved. Various statistical approaches are available to 'translate' these considerations into a statistical stopping rule. In this chapter one of the approaches is described extensively. For the determination of the stopping boundaries according to this method a general statistical framework needs to be defined first.

Statistical framework

Consider a clinical trial which is planned to estimate a certain treatment effect utilizing the statistic S, which is assumed to be Gaussian distributed. At the i-th interim analysis S_i has expectation μ and variance σ_i^2 (i=1, .., k-1). Given the precision $1/\sigma_T^2$, that is expected at the trials' planned end, the information time t_i is defined as σ_T^2/σ_i^2, with

$$0 < t_1 < \ldots < t_i < \ldots < t_k = 1.$$

The statistic W_i (Lan and Wittes (1988) call this statistic the B-value) is defined as

$$W_i = \frac{S_i}{\sigma_i}\sqrt{t_i} = Z_i\sqrt{t_i},$$

where Z_i is the normalization of S_i. W_i is a Gaussian distributed statistic with expectation

$$h_i(\mu) = \sqrt{t_i}\,\frac{\mu}{\sigma_i} = \frac{\mu}{\sigma_T}t_i,$$

and variance t_i. Clearly, the expectation and the variance of W_i change linearly with the increasing information. In the following, three assumptions about W_i are used:

(1) W_{i-1} and (W_i-W_{i-1}) are Gaussian distributed and independent;
(2) $E(W_i-W_{i-1}) = h_i(\mu) - h_{i-1}(\mu)$;
(3) $Var(W_i-W_{i-1}) = t_i - t_{i-1}$.

These assumptions are correct for the effect measures that are described in this thesis (Lan and Wittes, 1988). For the logrank statistic this was demonstrated by Tsiatis (1982). The stopping rule described in the previous section can be re-defined in terms of W_i The critical boundaries a_i and b_i are defined such that the trial stops when

$$W_i > a_i,$$

or

$$W_i < b_i,$$

with $a_i > b_i$; otherwise the trial will be continued. The final results of a clinical trial that stopped at t_i (and with former interim analyses at $t_1,..,t_{i-1}$) can be represented by the bivariate random vector (W,t_i). The probability density function of (W,t_i) can be deduced from the following considerations. Let

$$g(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

denote the probability density function of a normal random variable with expectation μ and variance σ^2. Through recursive integration we first define the auxiliary functions

$$f^*_\mu(w;t_1) = g[w;h_1(\mu),t_1],$$

and

$$f^*_\mu(w;t_j) = \int_{b_{j-1}}^{a_{j-1}} f_{j-1}(x;t_{j-1})\, g[w-x;\, h_j(\mu)-h_{j-1}(\mu),\, t_j-t_{j-1}]\, dx, \qquad j=2,...,k.$$

Then the density function is

$$f_\mu(w;t_i) = \begin{cases} f^*_\mu(w;t_i) & \text{if } i=k,\ w_i \geq a_i \text{ or } w_i \leq b_i \\ 0 & \text{otherwise} \end{cases}$$

From this density function the following absorption probabilities, which can be utilized for the determination of the boundaries, can be derived. The probability of crossing on of the boundaries at or before the i-th look is calculated as

$$P_i(\mu;a_i,b_i) = 1-\int_{b_i}^{a_i} f_\mu(u;t_i)du. \qquad (3.1)$$

The probability of being absorbed by the upper boundary at the i-th observation is

$$Q_i^+(\mu;a_i) = \int_{a_i}^{\infty} f_\mu(u;t_i)du \qquad (3.2.a)$$

and similarly for the lower boundary

$$Q_i^-(\mu;b_i) = \int_{-\infty}^{b_i} f_\mu(u;t_i)du. \qquad (3.2.b)$$

To calculate $f_\mu(w;t_i)$, $P_i(\mu;a_i,b_i)$, $Q_i^+(\mu;a_i)$, and $Q_i^-(\mu;b_i)$ numerical integration methods must be used. For a detailed description it is referred to Armitage, McPherson and Rowe (1969); McPherson and Armitage (1971); and Tsiatis, Rosner and Metha (1984). Note that

$$P_i(\mu;a_i,b_i) = \sum_{j=1}^{i} [Q_j^+(\mu;a_j) + Q_j^-(\mu;b_j)].$$

Note. In this section the statistic S has been assumed to be Gaussian distributed with known variance. In practice, the variance is usually unknown. If the S is Gaussian with unknown variance, then replacing the true variance by the estimated variance would have minimal impact on the results, since the size of the clinical trials considered in this paper is usually moderate or large (see also Pocock, 1977).

Stopping boundaries

The theory of interim analyses with stopping boundaries originates from the statistical theory of significance testing. In this setting the stopping boundaries also serve as critical boundaries. If one of the boundaries is crossed at a certain interim analysis the null-hypothesis is rejected and the trial is stopped. At the final analysis the trial is always stopped, but significance testing still requires critical boundaries. Therefore, if the purpose of a trial is to test a certain hypothesis $\mu=\mu_0$, also boundaries should be defined for the final analysis. These stopping rules are usually defined in terms of Z-values, which are statistics with a standardized Gaussian distribution, or the corresponding P-values. In the following it is considered that $\mu_0=0$, unless indicated otherwise.

Consider a rule, which says that if the test statistic z is larger than the conventional critical value $z_{\alpha/2}$ (= 1.96) at the 0.05 α level, the trial is stopped and the null-hypothesis is rejected. Otherwise, the trial is continued. If this critical value were naively used for each repeated test, the actual type I error rate, which is the false positive rate of testing the null-hypothesis, would escalate as shown in table 3.4. If a single test were made, the error would be the desired 0.050. If the data were tested twice, the error would be 0.083; for five tests the error would be 0.142; and for ten tests, the error is 0.193. For other values of α and $z_{\alpha/2}$, these values would change, but regardless of α, repeatedly testing data using the same critical value increases the false positive rate to levels higher than α. This problem of repeated significance testing was first described in thesis by Armitage et al (1969) and McPherson and Armitage (1971). The determination of the type I error rate of the testing procedures is based on calculations that are described in the statistical framework. Embodied in this statistical framework this repeated testing procedure leads to the following boundaries:

Table 3.4: Type I error rate of the repeated significance testing (RST) and the Haybittle-Peto (H-P) stopping procedures for k = 1,...,10.

k	RST	H-P	k	RST	H-P
1	0.050	0.050	6	0.155	0.054
2	0.083	0.051	7	0.166	0.055
3	0.107	0.052	8	0.176	0.055
4	0.126	0.053	9	0.185	0.056
5	0.142	0.053	10	0.193	0.056

$$a_i = -b_i = 1.96\sqrt{t_i}, \qquad \text{for } i=1,...,k;$$

Utilizing the absorption probability (3.1) the type I error rate of the procedure can be determined for a given value of k and α (see table 3.4)*.

An ad hoc rule, which was first proposed by Haybittle (1971) and subsequently advocated by Peto et al (1976), attempted to achieve conservatism in the interpretation of interim results. This rule, henceforth referred to as the Haybittle-Peto rule, suggests a conservative critical value (in terms of the Z-value) of ± 3 times the standard deviation for all but the last analysis and then the conventional ± 1.96 times the standard deviation critical value to obtain an approximate overall 5% α level. This rule, though simple to apply and conservative in interpreting interim results, does have an inconsistency. For example a test statistic of 2.9 times the standard deviation at the penultimate evaluation would not suggest early termination yet would be quite impressive in the final analysis. Embodied in the statistical framework, which allows for calculations on the actual type I error rate, this leads to the following boundaries:

$$a_i = -b_i = 3\sqrt{t_i}, \qquad \text{for } i=1,...,k-1;$$

and

$$a_i = -b_i = 1.96, \qquad \text{for } i=k.$$

Utilizing the absorption probability (3.1) the type I error rate of the procedure can be determined for a given value of k and α. Although the actual type I error rate is close to the nominal type I error rate (see table 3.4), the Haybittle-Peto rule does not guarantee a type I error rate.

* program SPEND, available from the author upon request.

Based on the notion that the type I error rate should be controlled, many stopping rules have been described in the literature. At first they were defined under the assumption that interim analyses take place after equal increments of information has accumulated. The two best known stopping boundaries are those defined by Pocock (1977), and by O'Brien and Fleming (1979). These methods were only defined for clinical trials with immediate outcomes. Due to work of Tsiatis (1982) these methods also could also be applied to survival studies. The assumption of equal increments of information could be discarded after work of Lan and DeMets (1983), who generalized the procedures of controlling the type I error rate.

Equal increments of information. A reasonable way to plan the interim analyses of a clinical trial in advance is to look after equal intervals of information time. In a trial with instantaneous response this can be easily established by looking after a fixed number of new patients have entered the trial; in a survival trial this means looking after a fixed number of observed deaths (Tsiatis, 1982). In the literature many stopping rules are based on the assumption that interim analyses take place after equal increments of information. The information time between two analyses is then determined by specifying the number of interim analyses, and vice versa. (For k-1 interim analyses this is $1/k$)

The first method, proposed by Pocock (1977), is a modification of the repeated significance testing procedure. The constant and symmetrical critical boundaries, defined in terms of Z-values, are increased to Z_P such that the overall type I error rate of the procedure is a pre-specified value of α. In terms of P-values this means the boundaries are decreased, for example if $k=5$, from 0.05 to 0.016 to reach a significant result on the 0.05 type I error rate. These decreased boundaries are called the nominal significance level, in contrast to the overall significance level α. Embodied in the statistical framework as described in this thesis the procedure entails the following boundaries:

$$a_i = -b_i = C_P \sqrt{t_i}, \qquad \text{for } i=1,\ldots,k.$$

Utilizing the absorption probability (3.1) the value of C_P can be determined through numerical iteration methods. The value of C_P depends on the number of interim analyses and on the value of α. For $\alpha=0.05$ and for different values of k the values of C_P ($= Z_P$) are shown in table 3.5*.

O'Brien and Fleming (1979) introduced different boundaries, also defined in terms of Z-values. While the Pocock boundaries remain constant, the O'Brien and Fleming

* program RST, available from the author upon request.

Table 3.5: Constants of the Pocock (C_P), the O'Brien and Fleming (C_{OF}), and the Segmental (C_S) boundaries to obtain $\alpha=0.05$ for $\mu=0$ and $k=1,...,10$.

k	C_P	C_{OF}	C_S	k	C_P	C_{OF}	C_S
1	1.96	1.96	1.96	6	2.45	2.05	2.18
2	2.18	1.98	2.00	7	2.49	2.06	2.23
3	2.29	2.00	2.04	8	2.51	2.07	2.29
4	2.36	2.03	2.09	9	2.54	2.08	2.34
5	2.41	2.04	2.13	10	2.56	2.09	2.43

boundaries change during the k tests. At first this stopping rule is quite conservative in the sense of requiring large critical values but as the study progresses these critical values decrease. Specifically, they proposed that the upper and the lower boundaries respectively are

$$Z_i = Z_{OF}\sqrt{t_i}, \qquad \text{for } i=1,...,k.$$

While O'Brien and Fleming determined the constant Z_{OF} for a given value of α and k using simulation methods, the solution can also be arrived by using the integration methods described in the statistical framework. Embodied in the statistical framework as described in this thesis the procedure entails the following boundaries:

$$a_i = -b_i = C_{OF}, \qquad \text{for } i=1,...,k.$$

Utilizing the absorption probability (3.1) the value of C_{OF} can be determined through numerical iteration methods. The value of C_{OF} depends on the number of interim analyses and on the value of α. For $\alpha=0.05$ and for different values of k the values of C_{OF} are shown in table 3.5*. The O'Brien and Fleming rule is initially very conservative but provides at the k-th and final test a critical value that is close to that of a conventional fixed sample experiment.

In the ASPECT trial an intermediate stopping rule, defined in terms of one-sided P-values, was selected (van Es et al., 1987). The basic idea behind this so-called segmental rule is to overcome the extreme conservativeness of the O'Brien and Fleming boundaries at the first interim analyses. This procedure requires the trial to be stopped and the null-hypothesis to be rejected if the P-value at the i-th analysis is less than g_i, with

$$g_i = 0.005, \qquad \text{for } i=1,2,\ldots, j,$$

and

$$g_i = (P_S-0.005)\frac{(i - j)}{(k- j)} + 0.005, \qquad \text{for } i=j+1,\ldots,k,$$

where j is chosen arbitrarily between 1 and k-1. The constant P_S was determined to obtain an overall level α (=0.05). Van Es et al. (1987) determined P_S using simulation methods for one-sided boundaries. The solution can also be arrived by using the numerical integration methods. Embodied in the statistical framework as described in this thesis, and for two-sided boundaries, the procedure entails the following boundaries:

$$a_i = -b_i = \Phi^{-1}(0.995)\sqrt{t_i}, \qquad \text{for } i=1,\ldots,j,$$

and

$$a_i = -b_i = \Phi^{-1}\{0.995 - [\Phi(-C_S) - 0.005]\frac{(i - j)}{(k- j)}\}\sqrt{t_i}, \qquad \text{for } i=j+1,\ldots, k,$$

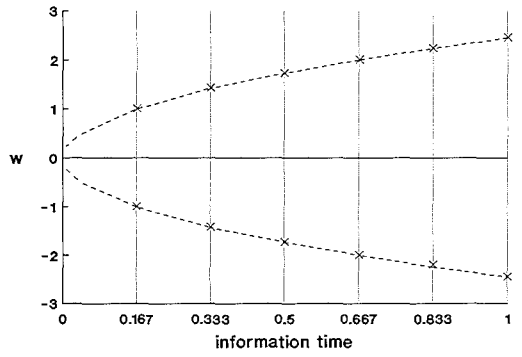
where Φ is the standard Gaussian cumulative distribution function. Utilizing the absorption probability (3.1) the value of C_S can be determined. For α=0.05, for k=1,...,10, and for j=(k-1)/2, if k is odd, and for j= k/2, for k is even, the values of C_S are shown in table 3.5*. In figure 3.2 the three stopping boundaries described above are displayed for μ_0=0 and k=6. It can be clearly seen that the so-called Segmental boundaries (van Es et al., 1987) are an intermediate of the 'constant' boundaries as defined by O'Brien and Fleming (1979) and the 'square root' boundaries as defined by Pocock (1977).

Unequal increments of information. The stopping rules described above require that the number of interim analyses be specified in advance, and that the interim analyses be equally spaced in information time. Lan and DeMets (1983) proposed a method for constructing stopping boundaries by using a so-called type I error rate or α-spending rate function which does not require these assumptions. In (Lan et al., 1984) they state that in contrast to a non-sequential design which spends the α-level at the end of the trial, a continuous sequential procedure can be described as spending the α-level over a period of information time t. If $\alpha(t)$ is defined to be the boundary crossing probability for μ=0 by time t, then $\alpha(t)$ can be interpreted as the probability of type I error rate spent by t. Clearly

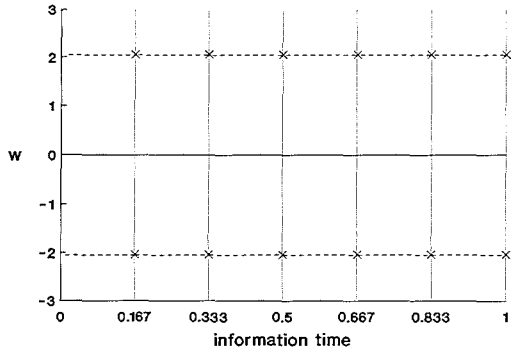
(1) $\alpha(0) = 0$;
(2) $\alpha(t)$ is non-decreasing; and
(3) $\alpha(1) = \alpha$.

---

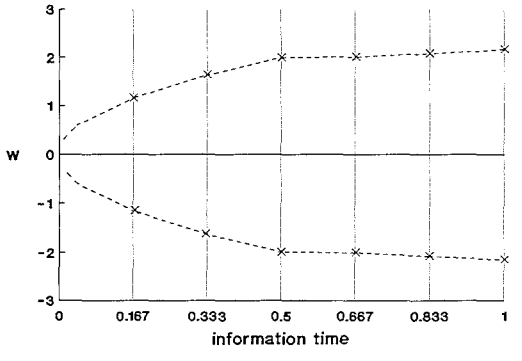
* program RST, available from the author upon request.



Pocock



O'Brien and Fleming



segmental

Figure 3.2: Symmetrical stopping boundaries in terms of the W-statistic of a clinical trial with 5 interim analyses, and with $\alpha=0.05$, according to: a. Pocock; b. O'Brien and Fleming; and c. van Es et al (segmental rule).

For discrete boundaries $\alpha(t)$ can be calculated at the planned interim analysis times. If at time t_i the i-th interim analysis takes place, then

$$\alpha(t_i) = P_i(0;a_i,b_i),$$

with $P_i(0;a_i,b_i)$ as defined by (3.1) in the statistical framework. For the boundaries of the ASPECT trial these boundary crossing (or absorption) probabilities at $t_i=i/6$ (i=1,...,6) are 0.005, 0.009, 0.012, 0.022, 0.035, and 0.050, respectively. Conversely, it is intuitively clear that if a function $\alpha(t)$ is chosen to satisfy the above 3 conditions, we can infer a particular boundary. The following α-spending functions roughly correspond with the Pocock and the O'Brien and Fleming boundaries respectively (Lan and DeMets, 1983):

$$\alpha(t) = \alpha \ln[1 + (e-1)t], \qquad 0 \le t \le 1$$

and

$$\alpha(t) = \begin{cases} 0 & t = 0 \\ 2[1 - \Phi(\frac{z_{\alpha/2}}{\sqrt{t}})] & 0 < t \le 1, \end{cases}$$

where $\Phi(.)$ is the standard Gaussian cumulative distribution function and $z_{\alpha/2}$ is the $(1 - \alpha/2)$ percentile of the standard Gaussian distribution. Kim and DeMets (1987) consider both the Pocock and the O'Brien and Fleming α-spending functions as extremes; the Pocock spending function sustaining early stopping and the O'Brien and Fleming spending function sustaining a high level of power in detecting a certain difference. Like the ASPECT boundaries, the straight line spending function

$$\alpha(t) = \alpha t, \qquad 0 \le t \le 1,$$

is one of the many possible spending functions that will result in intermediate stopping boundaries between these two extremes. The three spending functions mentioned above and the 'interpolated spending function' of the ASPECT boundaries are displayed in figure 3.3.

The ASPECT boundaries were planned for the 5 interim analyses to take place at information times 0.075, 0.20, 0.40, 0.60 and 0.825, respectively. However, the actual interim analyses did not take place at these planned times, as was described in one of the previous sections. The actual boundaries can be obtained through interpolation of the absorption probabilities mentioned above; it is not necessary to convert these discrete values into a continuous spending function. The boundaries for the factual situation of the ASPECT trial at the fourth interim analysis can be obtained as follows. The fourth interim analysis of the ASPECT trial takes place at the information time

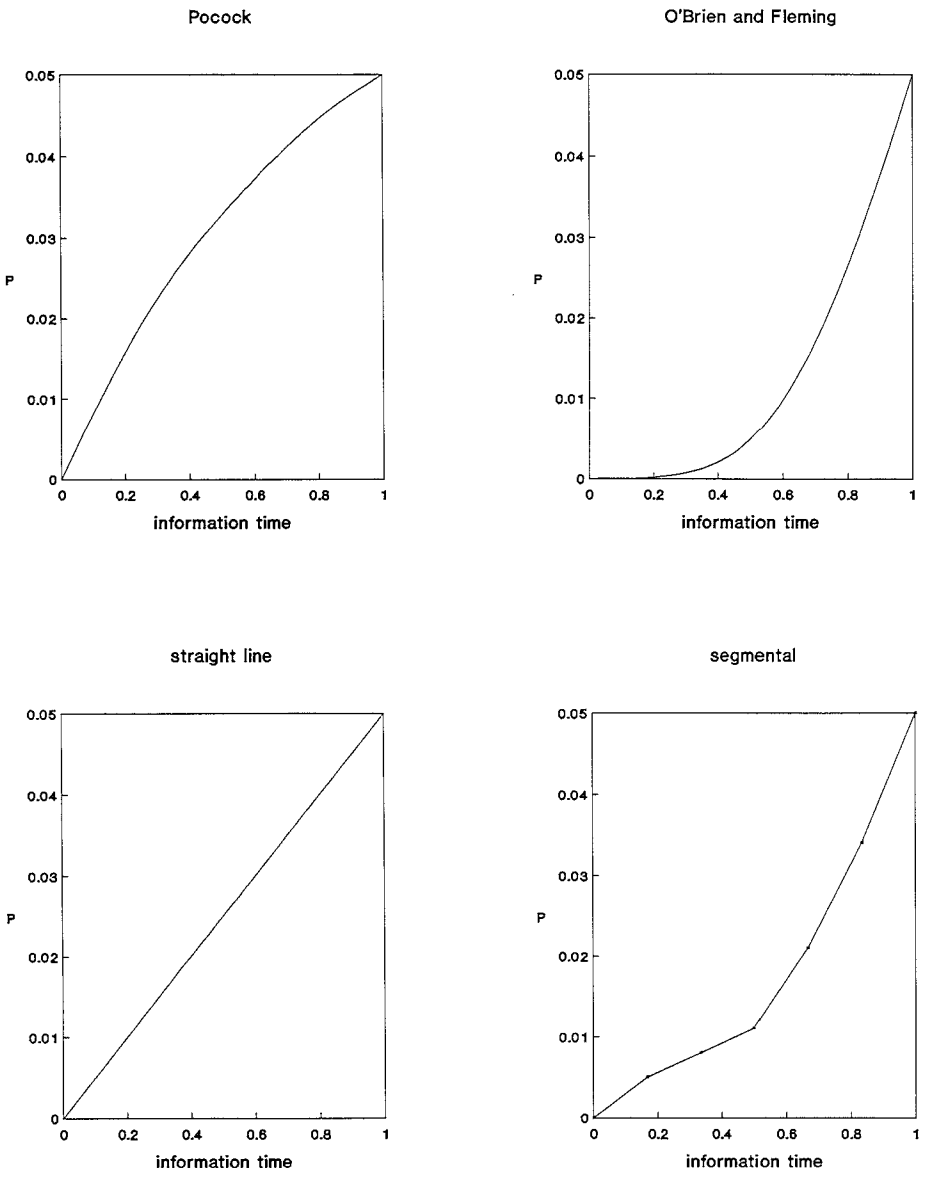


Figure 3.3: Type I error spending rate functions of the stopping boundaries of a clinical trial with 5 interim analyses, and with $\alpha = 0.05$, according to a. Pocock, b. O'Brien and Fleming, c. straight line, and d. van Es et al (segmental rule).

t_4=0.37. The amount of α to be spent until t_4 can be read from figure 3.3.d, and is 0.0091. Similarly the amount of α to be spent until the first three interim analyses were determined: 0.0015 at t_1=0.05, 0.0021 at t_2=0.07, and 0.0051 at t_3 =0.19.

Once the absorption probabilities are determined the boundaries can be obtained by utilizing the equations (3.2.a) and (3.2.b) as defined in the statistical framework. The upper boundary a_i and the lower boundary b_i (i=1,...,k) are defined to satisfy

$$Q_i^+(\mu;a_i) = \tfrac{1}{2}[\alpha(t_i)-\alpha(t_{i-1})],$$

and

$$Q_i^-(\mu;b_i) = \tfrac{1}{2}[\alpha(t_i)-\alpha(t_{i-1})],$$

with t_0=0. The boundary values a_i and b_i can be solved iteratively from these equations (Kim, 1987)*.

The actual ASPECT lower stopping boundaries (for a beneficial effect) were calculated accordingly. In figure 3.4 the planned and the actual stopping boundaries of the ASPECT trial are displayed. It should be noted that these boundaries are defined to be one-sided. The actual boundaries of the first two interim looks lie outside the planned boundaries.

COMMENTS

It is advisable to plan the moments of interim analyses and the shape of the stopping boundaries in advance. Although the effect of data-provoked changes the frequency of future data monitoring on the significance level and power is very small (Lan and DeMets, 1989), interim results might influence the investigators in making their choices.

Number of interim analyses

The frequency, and thereby the number of interim analyses usually depends practical considerations, such as the patient accrual rate, the time lag between entry an response evaluation, the administrative delays (especially in multicenter trials), and on the arrangements for DMC meetings (DMC meetings are usually schedule to coincide with meetings of the trial organizers so that necessary action can follow promptly). In choosing the number of interim analyses, the benefits of more frequent inspections of the data must be balanced against the effort required to perform additional analyses. This choice can be clarified at the planning stage of a clinical trial by presenting a

* program TEST, available from the author upon request.

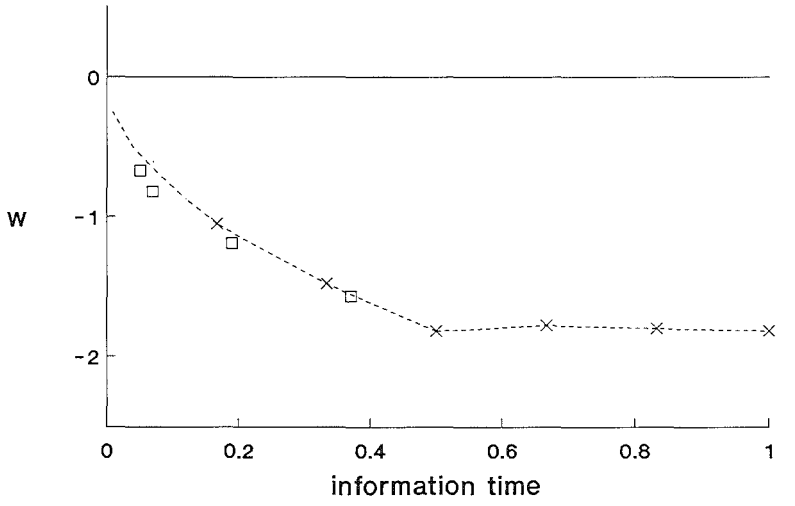


Figure 3.4: The actual lower stopping boundary values (squares) of the first four interim analyses and the planned (at equal increments of information) lower stopping boundaries (crosses) of the ASPECT trial.

summary of properties such as average sample number and power, with different numbers on interim analyses. For a discussion regarding the effect of increasing the number of interim analyses on power, average sample number, and maximal sample number it is referred to McPherson (1982) and Pocock (1982).

Power and average sample number

The power and the average sample number, as mentioned above, can be derived from the absorption probabilities as described in the statistical framework. Given the planned number of interim analyses, the stopping boundaries, and total amount of information the power (P) and the average sample number (ASN) for any given can be determined*:

$$P(\mu) = P_k(\mu;a_k,b_k)$$

and

$$ASN(\mu) = \sum_{i=1}^{k} [Q_i^+(\mu;a_i) + Q_i^-(\mu;b_i)]\, n_i,$$

where n_i is the expected number of patients in the trial at the i-th interim analysis.

* program DESIGN, available from the authors upon request.

REFERENCES

Armitage P, McPherson CK, Rowe BC. Repeated significance testing on accumulating data. *Journal of the Royal Statistical Society A,* 132, 235-244, 1969.

ASPECT Policy Board. *Anticoagulants in the Secondary Prevention of Events in Coronary Thrombosis: ASPECT- study (protocol).* Rotterdam, 1986.

DeMets DL, Lan KKG. An overview of sequential methods and their application in clinical trials. *Communications in Statistics Theory and Methods,* 13, 2315-2338, 1984.

Es GA van, Tijssen JGP, Lubsen J, Strik R van. Early termination of clinical trials with prolonged observation of individual patients: a case study. *Statistics in Medicine,* 6, 927-937, 1987.

Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials. *Statistics in Medicine,* 2, 167-174, 1983.

Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials incorporating clinical opinion. *Biometrics,* 40, 575-586, 1984.

Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology,* 44, 793-797, 1971.

Jennison J, Turnbull BW. Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials,* 5, 33-45, 1984.

Kalbfleiss JD, Prentice RL. *The statistical analysis of failure time data.* New York: Wiley, 1980.

Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika,* 74, 149-154, 1987.

Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika,* 70, 659-663, 1983.

Lan KKG, DeMets DL, Halperin M. More flexible sequential and non-sequential designs in long-term clinical trials. *Communications in Statistics Theory and Methods,* 13, 2339-2353, 1984.

Lan KKG, DeMets DL. Changing frequency of interim analysis in sequential monitoring. *Biometrics,* 45, 1017-1020, 1989.

Lan KKG, Wittes J. The B-value: a tool for monitoring data. *Biometrics,* 44, 579-55, 1988.

McPherson CK, Armitage P. Repeated significance testing on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society A,* 134, 15-25, 1971.

McPherson K. On choosing the number of interim analyses in clinical trials. *Statistics in Medicine,* 1, 26-36, 1982.

Miettinen OS. *Theoretical epidemiology.* New York: Wiley, 1985.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics,* 35, 549-556, 1979.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. introduction and design. *British Journal of Cancer,* 34, 585-612, 1976.

Pocock SJ. *Clinical trials: a practical approach.* Chichester: Wiley, 1983.

Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics,* 38, 153-162, 1982.

Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika,* 64, 191-199, 1977.

Rothman KJ. *Modern epidemiology.* Boston: Little Brown and Company, 1986.

Tijssen JGP, and Lubsen J. Principles of intervention research. *European Heart Journal (supplement H),* 8, 17-22, 1987.

Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association,* 77, 855-861, 1982.

Tsiatis AA, Rosner GL, Metha CR. Exact confidence intervals following a group sequential test. *Biometrics,* 40, 797-803, 1984.

Whitehead J. *The design and analysis of sequential clinical trials.* Chichester: Ellis Horwood, 1983.

chapter 4

A STOPPING RULE FOR THE ASPECT TRIAL

The text of this chapter has been reproduced from 'Es GA van, Tijssen JGP, Lubsen J, Strik R van. Early termination of clinical trials with prolonged observation of individual patients: a case study. Statistics in Medicine, 6, 927-937, 1987'. Few editorial adjustments have been made.

SUMMARY

Stopping rules for a placebo controlled clinical trial of anticoagulants after acute myocardial infarction were evaluated by means of computer simulation for the case of five interim analyses. The trial will be terminated and the null hypothesis of no treatment effect rejected when the one-sided P-value (logrank test) is lower than 0.005, 0.005, 0.005, 0.014, and 0.023 at the respective interim analyses, and 0.032 at final evaluation. This implies a total size $\alpha = 0.05$ and a power close to that of fixed sample size testing. The trial will also be terminated, without rejecting the null hypothesis, when the one-sided P-value exceeds 0.95, 0.88, 0.81, 0.74, and 0.67 at the respective interim analyses. This modification hardly affects size and power.

INTRODUCTION

The federation of Dutch Thrombosis Services is currently planning a clinical trial of Anticoagulants in the Secondary Prevention of Events in Coronary Thrombosis (ASPECT). The objective of ASPECT is to determine whether treatment with oral anticoagulants (OAC) after acute myocardial infarction (AMI) leads to a substantial decrease in total mortality. Because ASPECT is a trial with prolonged patient entry and long-term follow-up, interim analyses must be performed at regular time intervals for ethical reasons. This paper deals with the statistical aspects of selecting a stopping rule for ASPECT, with due emphasis on clinical applicability.

In the last ten years stopping rules in randomized clinical trials have been discussed frequently (Armitage et al., 1969; Armitage, 1975; O'Brien and Fleming, 1979; Pocock, 1982; DeMets and Lan, 1984; Halperin et al., 1982; and Canner, 1977). The literature

does not yield a clear-cut optimal stopping rule. Repeated significance testing (Armitage et al, 1969; Armitage, 1975) and O'Brien and Fleming's rule (1979) are usually advocated (Pocock, 1982; and DeMets and Lan, 1984). The above rules are designed for trials with immediate treatment response, thus yielding independent increments of the data. However, these methods are also used in trials with prolonged observation, that is, trials in which the new data are not independent of the data collected already. To evaluate the above stopping rules within the specific context of ASPECT a series of computer simulations was performed. In addition, two modified versions of these rules were considered.

We believe that early termination of the ASPECT study is not a symmetric issue. Therefore we consider stopping rules based on one-sided P-values. In addition we propose an extension of the one-sided stopping rule, without affecting the one-sidedness, which precipitates early termination of the trial when the probability of eventually obtaining a significant result within the trial has become very small.

THE ASPECT STUDY

Cardiological studies indicate that long-term treatment with OAC, compared to treatment with placebo, may produce a reduction in mortality after myocardial infarction of 10-20 per cent (Sixty Plus Reinfarction Study Research Group, 1980; Mitchell, 1981; and EPSIM Research Group, 1982). However, in view of risks associated with the therapy, the role of OAC in the secondary prevention of myocardial infarction is still under debate. These drugs are not used in most countries; France and the Netherlands are exceptions. The Federation of Dutch Thrombosis Services decided to conduct a clinical trial to determine whether institution of long term treatment with OAC immediately after AMI leads to a substantial decrease in total mortality. Given this objective we believe that one-sided hypothesis testing is warranted. If an interim analysis suggests deleterious effects of treatment, then it is unlikely that the final results will show a benefit; a harmful effect must be anticipated although its magnitude is irrelevant. On the other hand, if an interim analysis suggests some benefit from treatment we want to estimate its magnitude and avoid the possibility of stopping too early, that is, having an insufficiently convincing result.

The study is designed as a double blind, placebo controlled, randomized clinical trial with an average follow-up of two years. The outcome measure to compare the two treatment regimens is total mortality; the comparison will be made on an intention-to-treat basis.

Patients who have suffered from AMI will be eligible for entry within 2 to 6 weeks after onset, provided they satisfy the inclusion criteria and no exclusion criterion applies. A total of 4000 patients will be admitted to the study over a period of two years. The required number of patients is based on computations assuming fixed sample size testing (FST). Follow-up is continued until one year after admission of the last patient, thus the total duration of the trial is at most three years. All policy decisions in the course of the trial are made by an independent Policy Board of experts in cardiology, haemostasis, epidemiology, and biostatistics. The Data Monitoring Committee (a subcommittee of the Policy Board) will meet at least twice a year, that is six times in three years, to assess the interim results. For this purpose the committee will consider mortality and morbidity data for each treatment group and will act according to a pre-defined set of guidelines, specifying when to recommend that the Policy Board terminates the trial. The final decision will be taken by the Policy Board alone.

STOPPING RULES

In ASPECT the null hypothesis (H_0) is that under either treatment (OAC or placebo) survival times have the same distribution. It is tested against the alternative (H_1) that (H_1) that OAC treated patients have a longer expected survival time than placebo treated patients. Hypothesis testing is performed using the one-sided logrank test (Peto and Peto, 1972).

It is assumed that at K equal time intervals all follow-up information on patients enrolled in the trial is available. At each of the time points it must be decided whether H_0 is rejected and the trial terminated, or whether the trial is continued (at the Kth evaluation the trial is terminated anyway). The decision will be based on a stopping rule which is defined as a series of present cutoff points g_1, g_2,..., g_k for the P-values. When the P-value at the ith interim analysis is smaller than g_1, the trial will be terminated and H_0 rejected. The values of g_1, g_2, ..., g_k are subject to the condition that the overall level of the testing procedure does not exceed the present significance level (α). Four stopping rules are examined: repeated significance testing; O'Brien and Fleming's rule; and two modifications of these termed the linear increasing cutoff points rule and the segmental rule.

With repeated significance testing (RST) the values of g_i, $i=1,...,K$ are defined by:

$$g_i = C_1, \qquad (i=1,2,...K).$$

For O'Brien and Fleming's (OBF) rule the values of g_i are defined by:

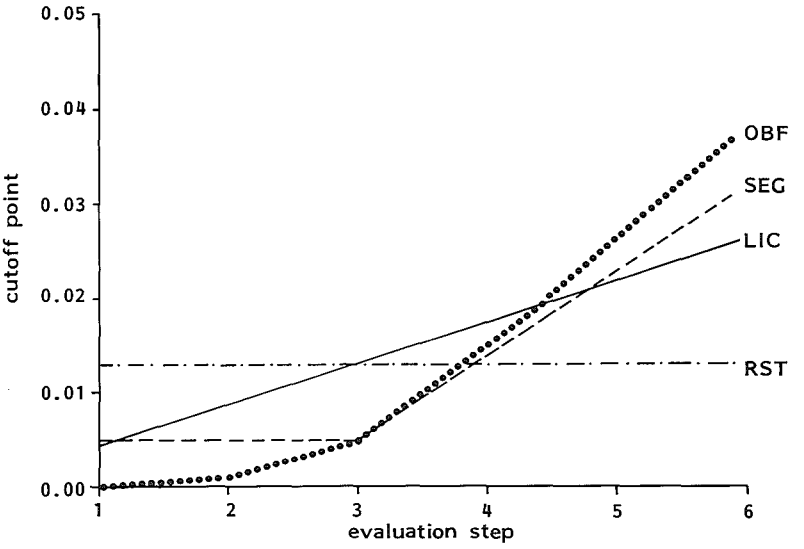


Figure 4.1: Nominal levels at each analysis for four stopping rules (K=6, j=3, α=0.05, C_1=0.013, C_2=1.08, C_3=0.026, and C_4=0.032): Repeated significance testing (RST), O'Brien and Fleming (OBF), linear increasing cutoff points (LIC) and Segmental (SEG)

$$g_i = 1 - \Phi(C_2\sqrt{\frac{K}{i}}\cdot\ \chi_{1-\alpha}), \qquad (i=1,2,...,K),$$

where Φ is the cumulative Gaussian distribution function and $X_{1-\alpha}$ is its (1-α)-fractile.

For the linear increasing cutoff points (LIC) rule the values of g_i are defined by:

$$g_i = C_3\frac{i}{K}, \qquad (i=1,2,...,K).$$

With the segmental (SEG) rule these values are defined by:

$$g_i = 0.005, \qquad (i=1,2,...,j),$$

$$g_i = (C_4 - 0.005)\frac{(i-j)}{(K-j)} + 0.005, \qquad (i=j+1,...,K),$$

where j is chosen arbitrarily between 1 and K-1.

C_1, C_2, C_3 and C_4 are constants defined implicitly by the overall significance level. For the ASPECT study K is equal to 6, with j chosen as 3 for the segmental rule. The four

stopping rules with selected values for C_1,C_2,C_3, and C_4 are illustrated in figure 4.1.

The selection of a convenient stopping rule for ASPECT is based on considerations of power and on its perspicuity to clinicians. This implies a stopping rule that is conservative at the early interim analyses, and has a critical value at the final evaluation which is close to FST. Peto et al (1976) suggest conservative nominal levels of about 0.0014 (that is, three standard deviations) at each interim analysis and of 0.05 at the final evaluation. For interpreting the evidence, we consider a rule like Peto's to be inconsistent. For example, a P-value of 0.01 at the penultimate evaluation would not reject H_0, whereas a P-value of 0.03 at the final evaluation, implying that the additional data tend to H_0, would. In the light of this, O'Brien and Fleming's rule is favoured over repeated significance testing. However, in the ASPECT study OBF has the one drawback that the first cutoff point is too extreme. Suppose that there are no deaths in the OAC group, but that in the placebo group there are, as expected, 15 deaths. Following OBF's rule the trial would not be terminated. We consider this too conservative and therefore two stopping rules that lie "between" OBF and RST were examined.

CURTAILED TESTING

In monitoring the accumulating data, it might become very unlikely that, given the current data, statistical significance will be obtained. This might be a good reason to terminate the trial early. A modification of one-sided FST is introduced and applied subsequently to the one-sided stopping rules.

This modification is represented as a series of upper cutoff points $h_1,h_2,...h_{k-1}$: when the P-value at the ith interim analysis is greater than h_i the trial will be terminated, and H_0 will not be rejected. At the final evaluation FST will be performed. The values of hi are subject to the condition that, under the assumption of H_0 being true, the probability of eventually rejecting H0 is between $\alpha - \epsilon$ and $\alpha(0 < \epsilon << \alpha)$. In addition, there should be no major loss of power under alternatives within the region of interest. From computer simulations it was apparent that a straight line would satisfy these criteria. The series of values of hi that represents this line has the form:

$$h_i = 0.95 - D(i - 1), \qquad (i=1,2,...,K-1),$$

where D is a constant defined implicitly by the choice of ϵ.

The principle of curtailed testing will also be applied to the other stopping rules. The

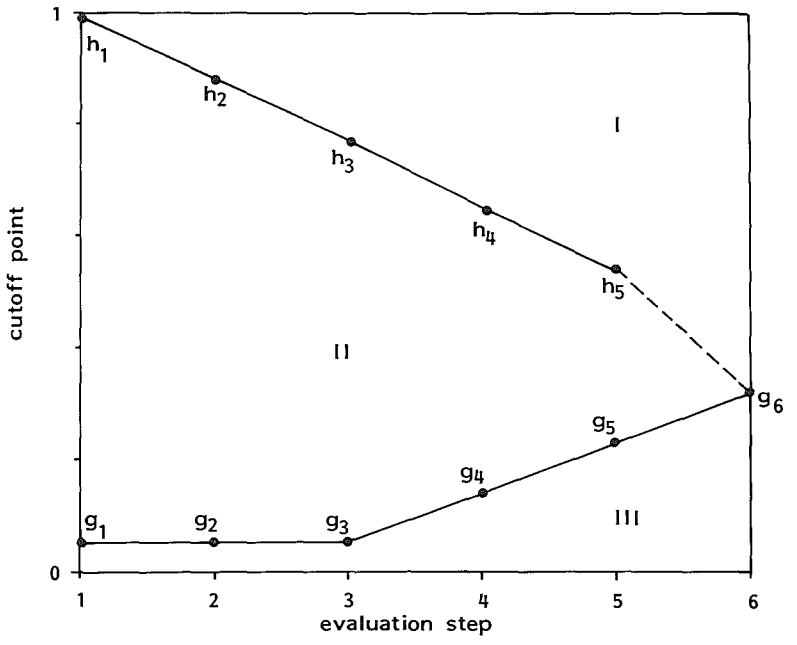


Figure 4.2: Schematic representation of a decision process for a one-sided stopping rule together with a curtailed testing procedure: I. terminate trial and do not reject H_0; II. continue trial; III. terminate trial and reject H_0

decision process applied in such a situation is shown in figure 4.2. As a first approximation, the upper cutoff points obtained from FST are applied to the stopping rules and the appropriateness of this procedure will be evaluated.

METHODS

Computer simulations of the ASPECT study were executed under various assumptions. To generate the data the following steps were executed independently: recruitment to the study was modelled as a Poisson process with a total accrual of 4000 patients in two years; treatment assignment was random in equal proportions and balanced in blocks of 12 consecutive patients; for each randomized patient survival time was generated using a Weibull distribution with parameters dependent on the specified hypothesis and the assigned treatment.

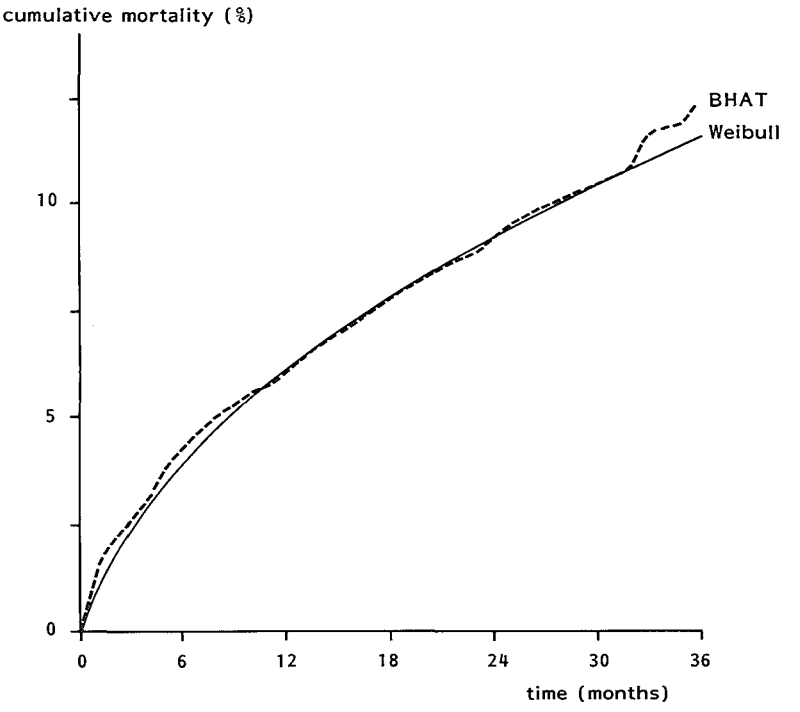


Figure 4.3: Cumulative mortalities of the placebo group in the Beta-Blocker Heart Attack Trial (BHAT) and the fitted two-parameter Weibull distribution (a=24789 and b=0.665)

The parameters a and b of the two-parameter Weibull distribution (Kalbfleisch and Prentice, 1980)

$$F(t) = 1 - \exp[(\frac{-t}{a})^b], \qquad a, b > 0; t \geq 0,$$

were fitted to the mortality data for the placebo group in the Beta-Blocker Heart Attack trial (Beta-Blocker Heart Attack Trial Research Group, 1982), giving values of 24789 and 0.665 respectively, with t expressed in days (see figure 4.3). In the remainder of this paper the parameter b is fixed at this value, whereas the parameter a is changed according to the cumulative mortality specified at two years. From now on the mortality distributions are expressed in terms of two-year mortality. The values of the parameter a of the Weibull distribution with two-year mortality of 5, 10 and 15 per cent are 63590, 21538 and 11223, respectively.

For each simulation run, approximately 4000 patients were accrued over a period of two years by generating the intervals (in days) between two consecutive patient entries according to an exponential distribution with parameter 0.1825 (that is, 730/4000). Survival times were censored at three years after the entry of the first patient. The generated data were evaluated every 6 months, that is, six times (K=6). The numbers of simulation runs were the same as those used by Canner (1977) for similar types of simulations.

Under H_0 survival times were simulated using mortality distributions with a two-year mortality of 5.0, 7.5, 10.0, 12.5, and 15.0 per cent, respectively for both treatment groups with 5000 independent simulation runs. For each series of runs, the constants C_1, C_2, C_3 and C_4 for RST, OBF, LIC, and SEG, respectively, were determined (Appendix I). For the curtailed testing procedure the constant D was determined with $\epsilon = 0.001$ (Appendix I). The values of h thus obtained were also used in RST, OBF, LIN, and SEG. The resulting modified stopping rules were evaluated in the same simulation series. The effect of misspecification of two-year mortality was studied by evaluating the simulation series with a two-year mortality of 5 per cent and 15 per cent with the stopping rules based on a two-year mortality of 10 per cent.

Under H_1 survival times were simulated with the two-year mortality of the placebo group fixed at 10 per cent and a two-year mortality in the OAC group of 7.0, 7.5., 8.0, 8.5, 9.0, 9.5, 10.0, 11.0, and 12.0 %, respectively, with 2000 independent simulation runs. For each series of runs, power and average study duration were calculated, with and without the curtailed testing procedure.

To gain some insight into the behaviour of the proposed stopping rules we calculated for each evaluation step the expected number of patients admitted, the expected number of patients deceased in the placebo group, and for each stopping rule the number of deaths in the OAC group required to precipitate early termination.

RESULTS

The cutoff points for the four stopping rules are described in table 4.1. They do not change appreciably with mortality levels, as shown in table 4.2. The consequences of misspecification of two-year placebo mortality are described in table 4.2 for the case when the two-year placebo mortality of 5 per cent and 15 per cent, respectively, is specified as 10 per cent: the overall levels do not differ much from 0.05.

The power of the four stopping rules and of fixed sample size testing under several

Table 4.1: Nominal levels for four stopping rules; K=6, α=0.05, 5000 simulations under H_0 at various levels.

	Nominal level	Two-year mortality 5%	7.5%	10%	12.5%	15%
(a)	Repeated significance testing (RST)					
	g_1	0.014	0.014	0.013	0.013	0.012
	g_2	0.014	0.014	0.013	0.013	0.012
	g_3	0.014	0.014	0.013	0.013	0.012
	g_4	0.014	0.014	0.013	0.013	0.012
	g_5	0.014	0.014	0.013	0.013	0.012
	g_6	0.014	0.014	0.013	0.013	0.012
(b)	O'Brien and Fleming (OBF)					
	g_1	0.000	0.000	0.000	0.000	0.000
	g_2	0.001	0.001	0.001	0.001	0.001
	g_3	0.006	0.007	0.006	0.005	0.005
	g_4	0.016	0.016	0.015	0.014	0.013
	g_5	0.027	0.028	0.026	0.024	0.024
	g_6	0.040	0.040	0.038	0.036	0.035
(c)	Linear increasing cutoff points (LIC)					
	g_1	0.004	0.004	0.004	0.004	0.004
	g_2	0.009	0.009	0.009	0.009	0.008
	g_3	0.013	0.013	0.013	0.013	0.012
	g_4	0.017	0.017	0.017	0.017	0.016
	g_5	0.021	0.022	0.022	0.022	0.021
	g_6	0.026	0.026	0.026	0.026	0.024
(d)	Segmental (SEG) (J=3)					
	g_1	0.005	0.005	0.005	0.005	0.005
	g_2	0.005	0.005	0.005	0.005	0.005
	g_3	0.005	0.005	0.005	0.005	0.005
	g_4	0.015	0.014	0.014	0.012	0.014
	g_5	0.025	0.022	0.023	0.019	0.022
	g_6	0.035	0.031	0.032	0.026	0.031

(see page 42 for definition of g_i, i=1,...,6)

Table 4.2: Overall levels of the stopping rules with nominal levels for two-year mortality of 10 per cent; K=6, j=3, α=0.05. 5000 simulations.

Stopping rule	Two-year mortality 5%	15%
Repeated Significance testing (RST)	0.049	0.055
O'Brien and Fleming (OBF)	0.048	0.053
Linear increasing cutoff points (LIC)	0.050	0.054
Segmental (SEG)	0.049	0.057

Table 4.3: Power of stopping rules under several alternatives for two-year mortality of 10 per cent in the placebogroup; K=6, j=3, α=0.05, 2000 simulations.

Stopping rule	Two-year mortality in the OAC group						
	7%	7.5%	8%	8.5%	9%	9.5%	10%
RST	0.912	0.783	0.574	0.382	0.229	0.099	0.050
OBF	0.959	0.859	0.690	0.480	0.275	0.129	0.049
LIC	0.944	0.837	0.655	0.446	0.258	0.115	0.049
SEG	0.951	0.848	0.671	0.460	0.266	0.118	0.049
FST	0.962	0.872	0.710	0.497	0.291	0.137	0.049

Repeated significance testing (RST), O'Brien and Fleming (OBF), linear increasing cutoff points (LIC), segmental (SEG) and fixed sample-size testing (FST), oral anticoagulants (OAC).

alternatives (with a two-year mortality of 10 per cent in the placebo group) is shown in table 4.3. The powers of O'Brien and Fleming's rule, of the segmental rule, and of the linear increasing cutoff points rule hardly differ for the mortality percentages considered here. Under all alternatives the power of repeated significance testing is clearly lower than the power of fixed sample size testing.

The average study duration, expressed as number of interim analyses, for the four stopping rules under several alternatives (two-year mortality of 10 per cent in the placebo group) is described in table 4.4. Under the alternatives investigated repeated significance testing has the shortest and O'Brien and Fleming's rule the longest average study duration. The average study duration of the linear increasing cutoff points rule hardly differs from that of repeated significance testing.

The constant D for the fixed sample size curtailed testing procedure, with $\epsilon = 0.001$, was calculated as 0.069. This means that the upper cutoff point decreases by 0.069

Table 4.4: Average number of interim analyses for stopping rules under several alternatives for a two-year mortality of 10 per cent in the placebo group; K=6, j=3, α=0.05, 2000 simulations.

stopping rule	two-year mortality in the OAC group						
	7%	7.5%	8%	8.5%	9%	9.5%	10%
RST	3.6	4.1	4.7	5.2	5.5	5.7	5.8
OBF	4.0	4.5	5.0	5.4	5.6	5.9	5.9
LIC	3.7	4.2	4.8	5.2	5.5	5.8	5.9
SEG	3.9	4.4	4.9	5.3	5.6	5.8	5.9

Repeated significance testing (RST), O'Brien and Fleming (OBF), linear increasing cutoff points (LIC), and segmental (SEG), oral anticoagulants (OAC).

Table 4.5: A Possible outcome of the ASPECT study with the cutoff points of the stopping rules expressed as the number of deaths in the OAC group.

Analysis	Number of patients	Number of deaths in placebo group	Cutoff points RST	OBF	SEG	LIC	Curtailed testing
1	1000	15	5	-	3	3	26
2	2000	40	22	17	20	21	51
3	3000	80	54	51	51	54	92
4	4000	120	88	89	89	90	130
5	4000	165	128	132	131	131	173
Final 6	4000	200	159	167	166	164	-

Repeated significance testing (RST), O'Brien and Fleming (OBF), linear increasing cutoff points (LIC), and segmental (SEG), oral anticoagulants (OAC).

at each evaluation step. The loss of power for this procedure was less than 0.009 compared to the power values in table 4.3. Additionally the simulations showed that if the same curtailed testing procedure was applied to the four stopping rules, the loss in α was always less than ϵ (=0.001), with the power loss never being greater than 0.011. Curtailed testing leads to a decrease in average number of interim analyses of about 1.4 (that is, 8 months) for all stopping rules under H_0.

In table 4.5 the cutoff points of the stopping rules are expressed as number of deaths in the OAC treatment group. If the expected number of deaths in the placebo group at the first evaluation (=15) does indeed occur, OBF would not terminate the trial, even if no deaths had occurred in the OAC group.

DISCUSSION

In the literature, stopping rules are described and investigated only for clinical trials with immediate treatment response. In this paper, O'Brien and Fleming's rule, repeated significance testing, and two modified versions, were investigated for a clinical trial with prolonged observation, that is, a trial in which at each interim analysis new data are not independent of those acquired earlier. The stopping rules were evaluated and compared by means of computer simulations.

The results of the simulations show that the cutoff points of the stopping rules are slightly dependent on two-year mortality hardly affects the overall level of the stopping rules (table 4.2).

For trials with immediate treatment response the cutoff points of repeated significance testing obtained by numerical integration procedures for a trial with five interim evaluations are 0.014 (by interpolation) (Armitage et al., 1969; and Pocock, 1982). This is in agreement with other results obtained by computer simulation (Canner, 1977). Our simulations of the chronic disease model yielded a similar outcome. Therefore the cutoff points for repeated significance testing in the ASPECT study could have been obtained by reference to stopping rules for trials of acute disease (Armitage et al., 1969; and Pocock, 1982).

For clinical trials of treatments that have an immediate response the cutoff point of O'Brien and Fleming's rule at the final evaluation is 0.042 (by interpolation) (O'Brien and Fleming, 1979). Our results suggest that the cutoff point at the final evaluation should be somewhat lower (0.038) for studies with prolonged observation. This corroborates other findings (Seigel and Milton, 1983) where, using O'Brien and Fleming's original cutoff points for chronic disease models, it was found that, close to our region of interest, the overall level is about 0.06.

By contrast to repeated significance testing, O'Brien and Fleming's rule has almost the same power as fixed sample size testing (table 4.3). This is in agreement with other findings (Pocock, 1982). As might be expected, both the linear increasing cutoff points rule and the segmental rule lead to a power close to that of O'Brien and Fleming's rule (table 4.3), but to a shorter average study duration (table 4.4).

The curtailed testing procedure for fixed sample size testing described in this paper yields satisfactory results. Application of the same series of upper cutoff points to the stopping rules gives equally satisfactory results. It is not necessary, therefore, to calculate these upper cutoff points for each stopping rule separately.

The results (table 4.5) illustrate the extremeness of OBF at the first interim analysis, and therefore it is doubtful whether the Policy Board would stick to the stopping rule in practice.

We have recommended that the Policy Board of the ASPECT study adopt the segmental rule together with the curtailed testing procedure (Appendix II). This procedure implies a size of 0.05, and a power of 0.84 when two year mortality in the placebo and OAC groups is 10 per cent and 7.5 per cent, respectively. It is emphasized that the choices for this stopping rule have been made in the particular context of the ASPECT study. It is not our aim to propose merely another stopping rule. When faced with the challenge of designing a stopping rule for a particular trial

one cannot rely on optimally criteria, which do not exist, therefore, one cannot avoid making choices.

It is stressed that a stopping rule should not be used as an absolute criterion in deciding whether to discontinue a clinical trial. This decision also requires the weighing of other relevant issues.

APPENDIX I: DETERMINATION OF THE CONSTANTS C_1, C_2, C_3, C_4, AND D

The four stopping rules, repeated significance testing (RST), O'Brien and Fleming (OBF), linear increasing cutoff (LIC) and segmental (SEG), are defined by a series of preset cutoff points $g_1,g_2,...,\ g_k$ for the P-values $P_1,\ P_2,...P_k$ at the respective interim analyses. When P_i is smaller than g_i, the trial will be terminated and H_0 rejected. The values of $g_1,g_2,...,g_k$ are subject to the condition that the overall level of the testing procedure does not exceed the preset significance level α. In the formulation of a stopping criterion this condition can be represented by a constant C that is defined implicitly by α. A function gi from the real line, R, to the interval, [0,1] can be defined, such that

$$g_i = g_i(C), \qquad (i=1,2,...,K)$$

For the four stopping rules RST, OBF, LIC, and SEG, the functions g_i, are given on page 929; the constants C_1, C_2, C_3, and C_4, respectively, have been substituted for C. For OBF we substitute 1/C'2 for C_2 as follows:

$$g_i(C_2) = 1 - \Phi\left(\frac{1}{C_2'}\sqrt{\frac{K}{i}}\,\chi_{1-\alpha}\right), \qquad (i=1,2,...,K),$$

where Φ is the cumulative Gaussian distribution function and $X_{1-\alpha}$ is its $(1-\alpha)$-fractile. The inverse functions, g_i^{-1} are given by:

(RST) $$g_i^{-1}(g_i) = g_i, \qquad (i=1,2,...,K),$$

(OBF) $$g_i^{-1}(g_i) = \Phi^{-1}(g_i)\sqrt{\frac{i}{K}}\chi_{1-\alpha}^{-1}, \qquad (i=1,2,...,K),$$

(LIC) $$g_i^{-1}(g_i) = \frac{K}{i}g_i, \qquad (i=1,2,...,K),$$

(SEG) $$g_i^{-1}(g_i) = (g_i - 0.005)\,\frac{(K - j)}{(i - j)} + 0.005, \qquad (i=1,2,...,K).$$

To determine the values of C_1, C'_2,C_3, and C_4, respectively, 5000 computer simulations under H_0 are executed (see Methods). From every simulation the minimum value G of $g_i^{-1}(P_i)$ is taken (i=1,2,...,K). The resulting 5000 values of G are sorted in ascending order, with $G_{(1)}$ being the smallest value. Given that α = 0.05. $H_{(250)}$ is the appropriate value for C.

In the curtailed testing procedure (applied to FST) the determination of D is similar. The upper cutoff points $h_1,h_2,...,h_k$ are written as a function h_i from the real line, R to the interval, [0,1] such that:

$$h_i = h_i(D) = 0.95 - D(i - 1), \qquad (i=1,2,...,K-1),$$

where D is defined implicitly by ϵ (see Curtailed Testing). The inverse function h_i^{-1} is given by:

$$h_i^{-1}(h_i) = \frac{0.95 - h_i}{i - 1}, \qquad (i=1,2,...,K-1).$$

To determine the value of D, 5000 computer simulations under H_0 are executed (see Methods). Under FST and α = 0.05 , we expect H_0 to be rejected in 250 of the 5000 simulations. From each of these 250 simulations the minimum value H, of $h_i^{-1}(P_i)$ is taken (i=1,2...,K-1). The resulting 250 values of H are sorted in ascending order, with $H_{(1)}$ being the smallest value. Given that ϵ = 0.001, $H_{(5)}$ is the appropriate value for D.

APPENDIX II: STOPPING RULE FOR THE ASPECT STUDY

	Analysis	Terminate and reject H_0	Continue	Terminate and do not reject H_0
Interim	1	P ≤0.005	0.005<P<0.95	P ≥0.95
	2	P ≤0.005	0.005<P<0.88	P ≥0.88
	3	P ≤0.005	0.005<P<0.81	P ≥0.81
	4	P ≤0.014	0.014<P<0.74	P ≥0.74
	5	P ≤0.023	0.023<P<0.67	P ≥0.67
Final	6	P ≤0.032		P ≥0.032

P is calculated from the one-sided logrank test, and H_0 is the hypothesis that under either treatment (OAC or placebo) the survival times have the same distribution.

REFERENCES

Armitage P. *Sequential Medical Trials, second edition.* New York: Wiley, 1975.

Armitage P, McPherson CK, Rowe BC. Repeated significance testing on accumulating data. *Journal of the Royal Statistical Society, Series A,* 132, 235-244, 1969.

Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *Journal of the American Medical Association,* 247 (2), 1704-1714, 1982.

Canner PL. Monitoring treatment differences in long-term clinical trials. *Biometrics,* 33, 603-615, 1977.

DeMets DL, Lan KKG. An overview of sequential methods and their applications in clinical trials. *Communications in Statistics,* 13, 2315-2338, 1984.

EPSIM Research Group. A controlled comparison of aspirin and oral anticoagulants in prevention of death after myocardial infarction. *New England Journal of Medicine,* 307, 701-708, 1982.

Halperin M, Lan KKG, Ware JH, Johnson NJ, Demets DL. An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials,* 3, 311-323, 1982.

Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* New York: Wiley, 1980.

Mitchell JRA. Anticoagulants in coronary heart disease-retrospect and prospect, *Lancet,* I, 257-262, 1981.

O'Brien PC, Fleming TR. A Multiple testing procedure for clinical trials, *Biometrics,* 35, 549-556, 1979.

Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A,* 135, 185-206, 1972.

Peto R, Pike M, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PJ. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *British Journal of Cancer,* 34, 585-612, 1976.

Pocock SJ. Interim analysis for randomized clinical trials: The group sequential approach. *Biometrics,* 38, 153-162, 1982.

Seigel D, Milton RC. Further results on a multiple testing procedure for clinical trials. *Biometrics,* 39, 921-928, 1983.

Sixty Plus Reinfarction Study Research Group. A double blind trial to assess long-term oral anticoagulant therapy in elderly patients after myocardial infarction. *Lancet,* II, 989-994, 1980.

chapter 5
EFFECT ESTIMATION

INTRODUCTION

The major objective of data analysis is to obtain a valid estimates of the treatment effects and a indication of their precision (confidence intervals). In the final reports of many recent clinical trials with interim analyses point and interval estimates are presented as if no interim analyses had been performed (HINT Research Group, 1986; CONSENSUS Trial Study Group, 1987; AIMS Trial Study Group, 1988; and CAST investigators, 1989). Estimation procedures that ignore the fact that interim analyses were performed during the course of a clinical trial are called 'naive' procedures. The consequences of the use of these naive estimation procedures is investigated. Various methods to determine point and confidence intervals, that do take into account the fact that interim analyses have been performed, have been proposed. Two of these methods, based on different orderings of the outcome space, are described. The methodology described in this chapter utilizes the statistical framework as described in chapter 3.

NAIVE ESTIMATION METHODS

Application of naive estimation procedures in clinical trials with interim analyses generally leads to biased point estimates and incorrect confidence intervals (Siegmund, 1978; Whitehead, 1983; and Kim, 1988). This section explores naive estimation methods in long-term clinical trials through simulation studies, based on the model of the CONSENSUS (COoperative North-Scandinavian ENalapril SUrvival Study) trial (CONSENSUS Trial Study Group, 1987). The CONSENSUS trial was a randomized, double blind, placebo-controlled clinical trial to study the effects on mortality of enalapril in severe congestive heart failure. A trial size of 400 patients was calculated on the assumption that the six-month mortality would be 40 percent in the placebo group and would be lowered to 24% by enalapril (α=0.05, power 0.90). Interim analyses were performed every 3 months by an independent Ethical Review Committee; no formal stopping criterion was defined. Based on the data from 244 patients the committee came to the conclusion that continuation was no longer justifiable on ethical grounds and of limited scientific value (Julian et al., 1987). On

recommendation of the Ethical Review Committee the trial was terminated. In the final report of the CONSENSUS trial naive estimation methods were used.

Apart from the effect of (possibly) early stopping on the estimation procedures, the effect of overrunning on the estimation procedures is investigated. Overrunning is the phenomenon that data will continue to accumulate after it is decided to stop a trial (Whitehead, 1986). In many cases there will be patients who have already been admitted to the trial but whose responses are not yet known. Also some extra patients will enter the trial because of the delay between the moment the data for the final interim analysis were retrieved, and the moment participating clinical centers receive instruction to stop recruitment.

Methods

For the simulations of the CONSENSUS trial a formal stopping criterion is defined which is based on the likelihood of the minimal clinically relevant effect E_{rel} (see chapter 3), and can be seen as a modification of the stopping rule described by Freedman et al (1983). The following stopping criterion, which is not aimed to preserve the type I error rate, was specified:

> Given a naive 95% confidence interval for the effect measure E at a certain interim analysis the trial will be continued if the confidence interval covers a pre-selected clinically relevant effect E_{rel}, and the trial will be stopped otherwise. If the trial has not been stopped at any interim analysis, the trial will be terminated at its planned end.

Different versions of the stopping criterion are applied, varying the nominal level, the number of interim analyses, and the timing of the interim analyses.

Clinical trials with gradual patient entry, and an event of interest (death) which occurs with constant rate and with follow-up data censored at each interim analysis are considered. The effect measure chosen is the incidence rate ratio IRR as defined in chapter 3. The point estimate of IRR and its confidence interval were determined according to the method described by Rothman (1986). The boundaries of the (1-α) confidence interval for IRR can be obtained as

$$irr\ \exp[\pm z_{\alpha/2}\sqrt{Var(\ln(irr))}],$$

where irr and Var(ln(irr)) are the point estimate and the estimated variance, respectively, as described in chapter 3.

For the model used in the simulations the clinical trial has a planned duration of 3 years; 2 years of patient entry and 1 year of follow-up after entry of the last patient. A series of k interim analyses is specified; at each interim analysis irr and the 95% confidence interval are calculated. The stopping criterion described above, with $E_{rel}=0.9$ (unless specified otherwise), is to be applied. To generate data according to this clinical trial model the following steps were simulated independently: recruitment was modelled as a Poisson process with an expected number of n patients in 2 years; treatment assignment was random with equal probabilities; event times were generated using an exponential distribution ($F(t)=\exp(-a.t)$, $a>0$) with different values for the parameter a in both treatment groups. For the reference group this parameter was fixed; for the index group the parameter was varied within a range of discrete values of the theoretical IRR within a range of practical interest.

From the CONSENSUS trial, the simulation parameters were obtained as follows: a mean recruitment rate of 155 patients/year and a mean mortality rate (the major outcome event was death) in the reference group of 1.022/year, corresponding to a cumulative six months mortality rate of 0.40.

For each of the following six models 5000 simulations were performed for values of IRR ranging from 0.5 to 1.0 with increments of 0.1:

model A: interim analyses are performed every six months (k=4) after the start of the trial;
model B: interim analyses are performed at 12, 18, and 24 months (k=3) after the start of the trial, i.e. omitting the first interim analysis;
model C: interim analyses are performed after every 50 deaths until 6 months before the end of the trial, i.e. equal intervals in terms of information time;
model D: model A (k=4), with overrunning, as if the trial continued for 3 months after the last interim analysis when the stopping decision was made;
model E: model A (k=4), with a stopping criterion based on 98.42% confidence interval, according to Pocock (1982), to preserve the overall type I error rate of 0.05; and
model F: model A (k=4), with a doubled rate of patient intake.

For each simulation irr and confidence interval were determined at the time of stopping. Coverage rates of the 95% confidence interval for IRR were determined separately for simulated trials that stopped at the first, the second, etc. interim analysis. Median values of the point estimates were determined separately for simulated trials stopped because the 95% confidence interval was above E_{rel} and for those because

Table 5.1: Coverage percentages of 95% confidence intervals at the time of stopping (E_{rel}=0.9) under model A, based on 5000 simulated trials.

IRR	interim analysis 6 months	12 months	18 months	24 months	final 36 months	overall
0.5	93.4% (944+0)	98.4% (2174+0)	100.0% (1385+0)	93.4% (410+0)	48.3% (87)	**96.6%**
0.6	85.2% (575+8)	97.0% (1348+1)	99.7% (1425+0)	100.0% (871+0)	84.6% (745)	**95.1%**
0.7	68.9% (342+24)	91.8% (656+6)	97.4% (799+2)	99.5% (789+0)	94.9% (2382)	**93.7%**
0.8	37.9% (195+58)	68.8% (291+29)	81.5% (273+13)	90.8% (267+4)	96.8% (3870)	**90.9%**
0.9	0.0% (110+116)	0.0% (94+95)	0.0% (66+83)	0.0% (66+55)	97.2% (4315)	**83.9%**
1.0	39.6% (62+218)	60.4% (21+262)	75.0% (14+262)	92.3% (7+280)	97.1% (3874)	**90.3%**

model A: interim analyses are performed every six months (k=4) after the start of the trial; **IRR** is the true incidence rate ratio; **overall** is the overall coverage percentage, irrespective of the stopping moment; the numbers between brackets indicate the number of simulated trials that were stopped because the 95% confidence interval was below or above E_{rel}, respectively.

the 95% confidence interval was below this value. Also, overall coverage percentages were obtained, taking all simulations together.

Results

In table 5.1 the coverage percentages under model A are specified according to the time of stopping. At IRR=E_{rel}=0.9 by definition none of the confidence intervals of the trials that stop early cover IRR. For IRRs close to E_{rel} the coverage rates in trials that stop at the first interim analysis are considerably lower than those in trials that stop at a later analysis; the majority of those trials continue to the planned end, and have coverage percentages of 95% and higher. For IRR=E_{rel}=0.9 overall coverage is 83.9%. When IRR moves away from E_{rel} the coverage percentage increases to 95% or more. The results are in fact symmetrical on a logarithmic scale around IRR=E_{rel}=0.9.

The overall coverage percentages were also calculated for various other values of E_{rel} and showed a similar pattern; minimal coverage percentages close to 85% are reached when IRR and E_{rel} are equal. For IRR moving away from E_{rel} more trials stop before their planned end, and higher coverage percentages are reached.

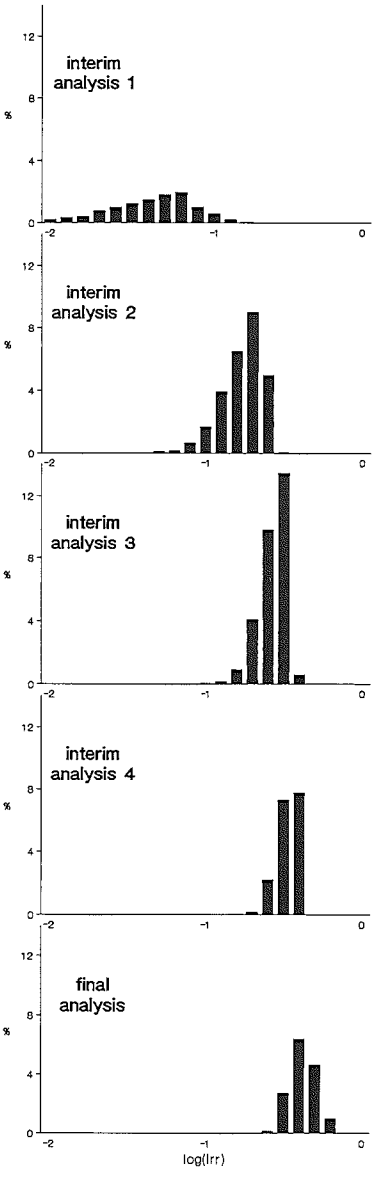


Figure 5.1: Distribution of the observed incidence rate ratio (logarithmic scale) at the 4 interim analyses and the final analysis. The simulated true incidence rate ratio in the 5000 simulated trials is 0.6; ln(0.6) = -0.51.

Table 5.2: Coverage percentages of 95% confidence intervals at the time of stopping (E_{rel}=0.9) under model B, based on 5000 simulated trials.

IRR	interim analysis 12 months	18 months	24 months	final 36 months	overall
0.5	96.6% (2997+0)	100.0% (1483+0)	98.7% (429+0)	48.4% (91)	**96.5%**
0.6	94.2% (1764+1)	99.7% (1545+0)	100.0% (924+0)	84.6% (766)	**95.5%**
0.7	87.2% (843+8)	97.3% (854+2)	99.5% (830+0)	94.9% (2463)	**94.8%**
0.8	63.8% (361+39)	81.8% (295+13)	90.0% (283+6)	96.9% (4003)	**92.9%**
0.9	0.0% (119+120)	0.0% (68+91)	0.0% (71+56)	97.3% (4475)	**87.1%**
1.0	58.2% (29+339)	74.3% (14+286)	92.5% (7+298)	97.1% (4027)	**92.6%**

model B: interim analyses are performed at 12, 18, and 24 months (k=3) after the start of the trial; **IRR** is the true incidence rate ratio; **overall** is the overall coverage percentage, irrespective of the stopping moment; the numbers between brackets indicate the number of simulated trials that were stopped because the 95% confidence interval was below or above E_{rel}, respectively.

Table 5.3: Coverage percentages of 95% confidence intervals at the time of stopping (E_{rel}=0.9) under model C, based on 5000 simulated trials.

IRR	interim analysis 50 cases	100 cases	150 cases	200 cases	final 36 months	overall
0.5	91.3% (3110+1)	99.6% (1363+0)	100.0% (401+0)	83.3% (96+0)	51.7% (29)	**93.9%**
0.6	85.8% 1872(3+)	98.6% (1423+1)	100.0% (781+0)	100.0% (447+0)	81.6% (473)	**92.5%**
0.7	76.2% (1010+13)	93.5% (848+3)	98.7% (677+1)	100.0% (540+0)	93.8% (1909)	**91.5%**
0.8	49.6% (500+59)	67.9% (359+27)	85.1% (292+11)	92.1% (265+5)	97.1% (3487)	**88.5%**
0.9	0.0% (198+164)	0.0% (139+130)	0.0% (86+69)	0.0% (67+55)	98.6% (4092)	**80.7%**
1.0	39.4% (80+362)	60.9% (39+334)	73.7% (25+249)	90.4% (11+208)	98.2% (3692)	**88.6%**

model C: interim analyses are performed after every 50 deaths until six months before the end of the trial (k=4); **IRR** is the true incidence rate ratio; **overall** is the overall coverage percentage, irrespective of the stopping moment; the numbers between brackets indicate the number of simulated trials that were stopped because the 95% confidence interval was below or above E_{rel}, respectively.

Table 5.4: Coverage percentages of 95% confidence intervals at the time of stopping (E_{rel}=0.9) under model D, based on 5000 simulated trials.

IRR	interim analysis 6 months	12 months	18 months	24 months	final 36 months	overall
0.5	90.8% (944+0)	97.8% (2174+0)	99.5% (1385+0)	91.7% (410+0)	48.3% (87)	**95.6%**
0.6	86.3% (575+8)	95.3% (1348+1)	99.3% (1425+0)	99.9% (871+0)	84.6% (745)	**94.6%**
0.7	77.6% (342+24)	87.5% (656+6)	95.8% (799+2)	98.9% (789+0)	94.9% (2382)	**93.4%**
0.8	67.6% (195+58)	72.8% (291+29)	76.2% (273+13)	85.6% (267+4)	96.8% (3870)	**92.0%**
0.9	62.4% (110+116)	51.9% (94+95)	50.3% (66+83)	43.0% (66+55)	97.2% (4315)	**91.2%**
1.0	69.6% (62+218)	72.8% (21+262)	81.5% (14+262)	87.1% (7+280)	97.1% (3874)	**92.7%**

model D: 3 months of overrunning, interim analyses are performed every six months (k=4) after the start of the trial; **IRR** is the true incidence rate ratio; **overall** is the overall coverage percentage, irrespective of the stopping moment; the numbers between brackets indicate the number of simulated trials that were stopped because the 95% confidence interval was below or above E_{rel}, respectively.

Figure 5.1 shows the distribution of the point estimates irr on a logarithmic scale for simulated trials with IRR=0.6 separately for trials that stop because the data suggest a beneficial effect and for trials that continue until the final analysis. At the first interim analysis all 575 of the 5000 (i.e. 11.5%) simulated trials stop with an irr less than 0.6. At the final analysis the distribution is shifted towards the right: the median value of irr is 0.64. Figure 5.2.A presents the median values of irr for those trials that were stopped under model A because of a suspected beneficial effect, and for those trials that continued until the planned end. In combination with table 5.1 it can be seen that extreme cases, i.e. irr shifted far away from IRR with low coverage rates of the confidence interval, occur less often.

The results concerning model B (omitting the first interim analysis) are shown in table 5.2 and figure 5.2.B. The minimum overall coverage at IRR=E_{rel} is 87.1%. The median point estimates irr under model B hardly differ from those in model A; the only difference can be observed at the interim analysis at 12 months: under model B there is less bias at the lower values of IRR. With regard to the coverage rates of confidence intervals, model A and B are also similar.

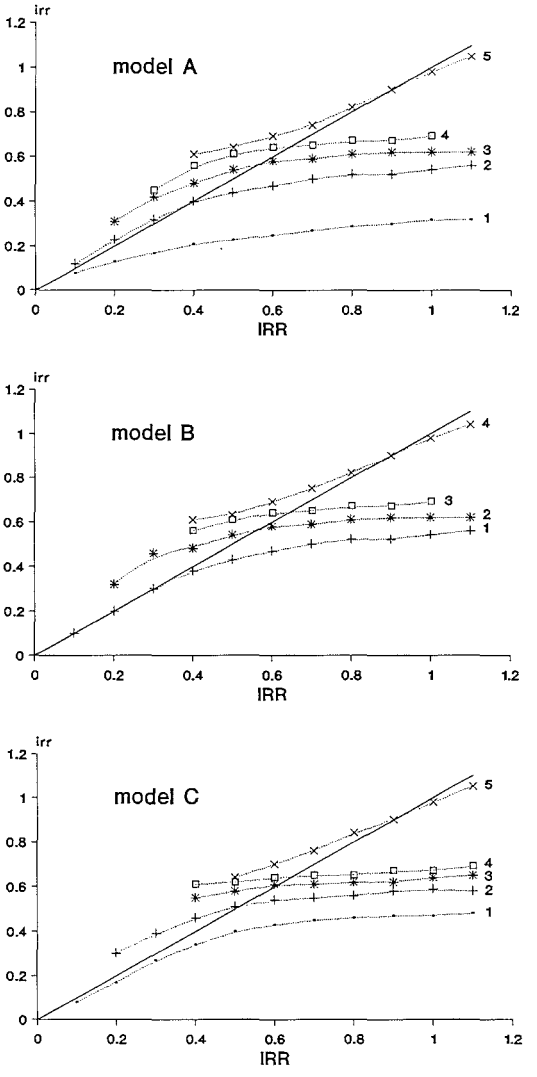


Figure 5.2.a-c: Median values of the observed incidence rate ratio (irr) for different values of the true incidence rate ratio (IRR) at the time the simulated trials were stopped (1 - stopped at first analysis, 2 - stopped at second analysis, 3 - stopped at third analysis, 4 - stopped at fourth analysis, 5 - stopped at fifth analysis). For each indicated value of IRR 5000 trials were simulated.

5.2.a - model A, interim analyses are performed every six months (k=4) after the start of the trial;

5.2.b - model B, interim analyses are performed at 12, 18, and 24 months (k=3) after the start of the trial;

5.2.c - model C, interim analyses are performed after every 50 deaths until six months before the end of the trial (k=4).

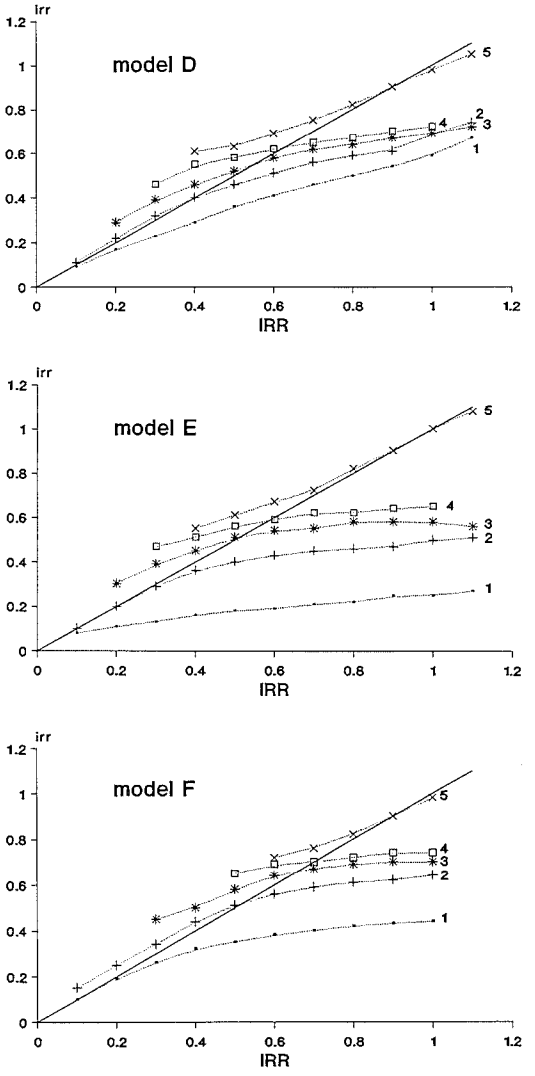


Figure 5.2.d-e: Median values of the observed incidence rate ratio (irr) for different values of the true incidence rate ratio (IRR) at the time the simulated trials were stopped (1 - stopped at first analysis, 2 - stopped at second analysis, 3 - stopped at third analysis, 4 - stopped at fourth analysis, 5 - stopped at fifth analysis). For each indicated value of IRR 5000 trials were simulated.

5.2.d -model D, 3 months of overrunning, interim analyses are performed every six months (k=4) after the start of the trial;

5.2.e -model E, Pocock boundaries, interim analyses are performed every six months (k=4) after the start of the trial;

5.2.f - model F, doubled patient intake, interim analyses are performed every six months (k=4) after the start of the trial.

Table 5.5: Coverage percentages of 95% confidence intervals at the time of stopping (E_{rel}=0.9) under model E, based on 5000 simulated trials.

IRR	interim analysis 6 months	12 months	18 months	24 months	final 36 months	overall
0.5	78.6% (281+0)	96.2% (1829+0)	99.8% (1856+0)	99.9% (791+0)	67.1% (243)	**95.7%**
0.6	54.2% (165+1)	91.7% (867+0)	98.9% (1321+0)	100.0% (1157+0)	91.1% (1489)	**94.1%**
0.7	11.0% (95+5)	75.4% (348+2)	90.6% (488+0)	98.0% (684+0)	96.2% (3378)	**92.7%**
0.8	0.0% (49+11)	6.5% (98+10)	52.3% (149+6)	67.7% (156+2)	96.5% (4519)	**91.1%**
0.9	0.0% (29+26)	0.0% (27+39)	0.0% (26+25)	0.0% (15+26)	96.1% (4787)	**92.0%**
1.0	0.0% (14+52)	2.6% (8+109)	33.3% (5+148)	64.9% (2+132)	96.7% (4530)	**90.4%**

model E: Pocock boundaries, interim analyses are performed every six months (k=4) after the start of the trial; **IRR** is the true incidence rate ratio; **overall** is the overall coverage percentage, irrespective of the stopping moment; the numbers between brackets indicate the number of simulated trials that were stopped because the 98.42% confidence interval was below or above E_{rel}, respectively.

The results for model C (interim analyses after every 50 deaths) are shown in table 5.3 and figure 5.2.C. As expected it can be seen that, compared to model A, estimation only is improved for those trials that stop at the first interim analysis.

Table 5.4 and figure 5.2.D give the results for model D (3 month period of overrunning). It can be seen that overrunning leads to a shift of the median point estimate towards the true value. Trials that stop at 12, 18 or 24 months with IRR<0.8 have decreased coverage rates compared to model A.

The effects of applying more rigid critical boundaries (model E) are shown in table 5.5 and figure 5.2.E. In nearly all cases coverage rates are less than under model A. As expected overall coverage in the region around IRR=0.9 are higher compared to model A, due to the fact that less trials stop early. The median bias is not much different from that under model A.

The results for model F (doubled patient intake) are shown in table 5.6 and figure 5.2.F. More trials stop early, and have better coverage, than under model A; overall coverage, however, is comparable. The median bias is a little less than under model A.

Table 5.6: Coverage percentages of 95% confidence intervals at the time of stopping (E_{rel}=0.9) under model F, based on 5000 simulated trials.

IRR	interim analysis 6 months	12 months	18 months	24 months	final 36 months	**overall**
0.5	94.8% (1836+0)	99.1% (2636+0)	88.2% (510+0)	16.7% (18+0)	0.0% (0)	**96.1%**
0.6	88.6% (1073+7)	98.9% (2150+0)	100.0% (1314+0)	94.2% (395+0)	37.7% (61)	**95.9%**
0.7	74.8% (574+17)	95.8% (1137+0)	99.8% (1257+0)	100.0% (922+0)	90.6% (1093)	**93.9%**
0.8	45.4% (286+53)	76.6% (399+15)	91.3% (491+6)	97.0% (462+2)	96.7% (3286)	**91.1%**
0.9	0.0% (129+142)	0.0% (99+93)	0.0% (70+69)	0.0% (55+57)	97.5% (4286)	**83.5%**
1.0	40.8% (53+268)	78.2% (11+379)	90.4% (8+397)	96.8% (1+406)	96.7% (3477)	**91.1%**

model F: doubled patient intake, interim analyses are performed every six months (k=4) after the start of the trial; **IRR** is the true incidence rate ratio; **overall** is the overall coverage percentage, irrespective of the stopping moment; the numbers between brackets indicate the number of simulated trials that were stopped because the 95% confidence interval was below or above E_{rel}, respectively.

On the whole the results show that the median point estimates irr underestimate the true value of IRR in trials that stop at the first interim analysis in all investigated cases; in trials that continue until the planned end IRR is overestimated for IRR < E_{rel}. For trials that stop in between the identity line in figure 5.2 is always crossed at some point between 0 and E_{rel}.

The estimation procedures used in this study were also applied to a fixed sample clinical trial model, resulting in median unbiased point estimates and coverage rates of the 95% confidence intervals very close or equal to 95%.

Discussion

In a fixed sample trial irr asymptotically is an unbiased estimator. However, when applying a stopping criterion, that is based on the data collected until the actual interim analysis, in general leads to biased estimates. The results show that especially early stopping can lead to a considerable shift of the point estimate. Equivalent to sequential trials (Whitehead, 1983) irr is pushed away from IRR when a suspected treatment difference has been the reason for stopping the trial at the first interim analysis: irr

underestimates IRR. If the trial continues beyond the first interim analysis the magnitude of the bias becomes smaller; in some cases irr even overestimates IRR. The results are similar to those from Hughes and Pocock (1988), who studied a clinical trial model with a short-term effect measure and with more statistical power.

Application of the stopping criterion as described in this chapter, which is based on the naive 95% confidence interval, often results in a coverage close to 95%. However, in (the relatively rare) situations where considerable bias occurs, the coverage percentage also decreases substantially. This happens when trials are stopped at an early interim analysis, and/or when trials with IRR=E_{rel} stop before the final analyses. The latter coincides with results from Armitage et al (1969). The magnitude of the bias at a certain interim analysis is mainly dependent on the information available at the time the interim analysis is performed: the earlier the stop, the bigger the bias of the point estimate will be. With respect to the timing of the (first) interim analysis of a clinical trial it should therefore be guarded that enough information is available.

At the time of the final evaluation of a clinical trial that stopped before its planned end additional data may have come available (overrunning). The use of these data, which are not influenced by the decision to stop and which can be combined with the data received prior to termination, leads to a natural bias reduction (see figures 5.2.A and 5.2.D). However, this improvement does not automatically result in better coverage percentages in all situations (compare tables 5.1 and 5.4).

The notion that 'naive' estimation methods cannot directly be applied to experiments with interim analyses of the data is strongly corroborated in the preceding sections. At the conclusion of a clinical trial with interim analyses, the naive estimates are no longer valid because of the dependence of the stopping rule on the nature of the evidence collected. However, if a trial is not stopped at an interim analysis but at its planned end the 'naive' point estimate and its confidence interval generally will lead to satisfactory results.

Although the computer simulations do show the weaknesses of naive estimation procedures they do not offer a solution to the estimation problem (van Es et al., 1988). The results show that, besides being dependent of the stopping rule, the bias depends on the true treatment effect. Practical implementation of the above results is impossible because they concern estimation properties of group sequential designs where treatment effect is known; in no clinical trial is the true effect ever known.

EXACT ESTIMATION METHODS

Given the stopping rule, i.e. the times at which interim analyses are performed, and the corresponding boundaries, the outcome of a clinical trial can be characterized by the moment of stopping and the magnitude of the estimated treatment effect. In case the trial is early terminated, this treatment effect must have exceeded one of the boundary values at that time. The set of all possible outcomes, which is called the outcome space, is therefore uniquely defined by the timing of the interim analyses and the stopping rule that has been applied. By defining an order relationship in the outcome space it is possible to calculate effect estimates, that are adjusted for the fact that a (selective) stopping criterion has been used to obtain the final result. Different estimation procedures described in the literature (Siegmund, 1978; Tsiatis et al., 1984; Rosner and Tsiatis, 1988; Chang, 1989) are based on this notion. Although these methods are constructed to be exact, they differ depending on which ordering is used to relate the outcome space to the parameter space. In the following sections two different orderings of the outcome space are defined, followed by a discussion of procedures to calculate the point estimate of IRR together with its confidence interval, and the P-value based on these orderings.

Ordering of the outcome space

The results of a clinical trial with interim analyses can be summarized by the bivariate vector (W,τ), where W denotes the Gaussian distributed test statistic as described in the statistical framework (chapter 3), and τ is the moment at which the trial is stopped expressed in units of information time. The outcome space is defined as the set of all possible outcomes (i.e. values of (W,τ)) of the trial. In figure 5.3 the outcome space of a clinical trial with 5 interim analyses equally spaced in information time and with Pocock stopping boundaries is shown. Clearly the outcome space is defined by the stopping criterion that is used: the moment of looking and the stopping boundaries determine the outcome space. By inducing an order relationship on the outcome space the probability $Pr_{\mu}[(W,\tau) > (w,t_m)]$ can be defined with respect to this outcome space. If it is assumed that this probability is a continuous and monotonically increasing function of μ, then there exists a unique solution to the equation

$$Pr_{\mu}[(W,\tau) > (w,t_m)] = p,$$

for any (w,t_m) and $p \in (0,1)$. In the following sections two order relationships are considered. Other order relationships have been described by Jennison and Turnbull (1983), and Rosner and Tsiatis (1988).

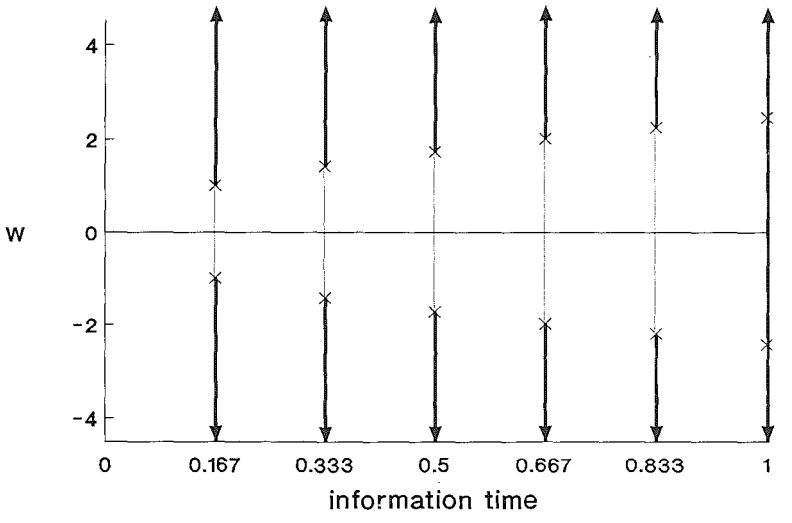


Figure 5.3: Outcome space from clinical trial with 5 interim analyses at times 0.167, 0.333, 0.5, 0.667, and 0.833, and with two-sided Pocock boundaries ($\alpha = 0.05$).

Intuitive ordering. Siegmund (1978) developed an ordering of the sample space, introduced by Armitage (1957), which corresponds to the notion that the larger the value of w, the estimate of W, the earlier one will stop. The order relationship firstly depends on the boundary crossed, then on the stopping time t, and secondly on the value of w. This so-called intuitive ordering, induced on the same outcome space as in figure 5.3, is illustrated in figure 5.4.

More formally, the order relationship on the sample space of (W,τ) is defined as follows: the value (w,t) is greater than (w^*,t^*) if and only if one of the following is true:

(1) $w \geq a$ and $w^* \leq b^*$;
(2) $t < t^*$ when $w \geq a$ and $w^* \geq a^*$;
(3) $t > t^*$ when $w \leq b$ and $w^* \leq b^*$; or
(4) $w > w^*$.

Here a, b and a^*, b^* are the upper and lower boundaries at t and t^*, respectively. In (Kim and DeMets, 1987) the proof of monotonicity is given for this intuitive order relationship. Embodied in the statistical framework and under this ordering of the

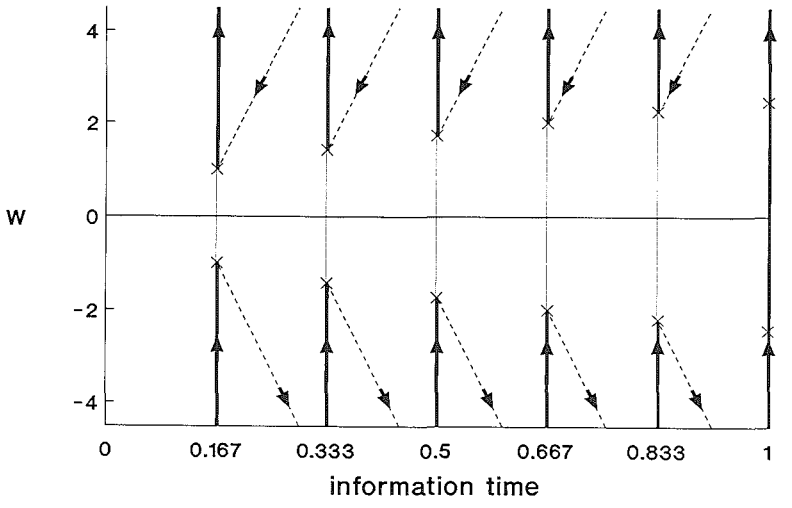


Figure 5.4: Intuitive ordering defined on an outcome space from clinical trial with 5 interim analyses at times 0.167, 0.333, 0.5, 0.667, and 0.833, and with two-sided Pocock boundaries ($\alpha=0.05$).

outcome space, the probabilities $Pr_\mu[(W,\tau) > (w,t_m)]$ and $Pr_\mu[(W,\tau) < (w,t_m)]$ can be expressed in terms of the absorption probabilities (3.2.a) and (3.2.b) as follows:

$$Pr_\mu[(W,\tau) > (w,t_m)] = \sum_{i=1}^{m-1} \left(Q_i^+(\mu;a_i)\right) + Q_m^+(\mu;w),$$

and

$$Pr_\mu[(W,\tau) < (w,t_m)] = \sum_{i=1}^{m-1} \left(Q_i^-(\mu;b_i)\right) + Q_m^-(\mu;w).$$

These probabilities are a function of the stopping boundaries prior to stopping and not any future stopping criteria. It can be seen that, for clinical trials that have been stopped at the first interim analysis, this intuitive ordering is identical to the usual fixed sample ordering, and therefore will result in the same estimates.

Likelihood ratio ordering. Rosner and Tsiatis (1988) and Chang (1989) described an order relationship of the outcome space, that is based on a standardized measure of distance from the data to a certain value of the parameter of interest μ. The basic idea behind the ordering is that all pairs of (w,t) that have the same distance to μ are of equal order in the outcome space. This so-called standardized distance, which is

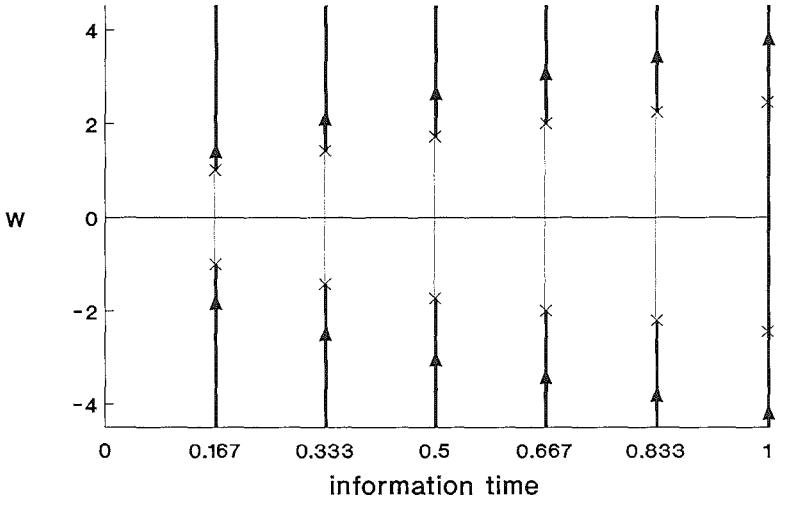


Figure 5.5: Likelihood ratio ordering defined on an outcome space from clinical trial with 5 interim analyses at times 0.167, 0.333, 0.5, 0.667, and 0.833, and with two-sided Pocock boundaries (α=0.05).

equivalent to the likelihood ratio test statistic for the pair (w,t) evaluated at μ, is

$$D(\mu,w,t) = \sqrt{t}\left|\frac{w}{t} - \frac{\mu}{\sigma_T}\right|.$$

The order relationship, based on this statistic, is defined as follows: the value (w,t) is greater than (w^*,t^*) if and only if

$$\sqrt{t}\left|\frac{w}{t} - \frac{\mu}{\sigma_T}\right| > \sqrt{t^*}\left|\frac{w^*}{t^*} - \frac{\mu}{\sigma_T}\right|.$$

This inequality can be written, for each t, as

$$w > t\frac{\mu}{\sigma_T} + \sqrt{tt^*}\left(\frac{w^*}{t^*} - \frac{\mu}{\sigma_T}\right),$$

and

$$w < t\frac{\mu}{\sigma_T} - \sqrt{tt^*}\left(\frac{w^*}{t^*} - \frac{\mu}{\sigma_T}\right),$$

if (w,t) is smaller than (w^*,t^*). The likelihood ratio order relationship for μ=0 is displayed in figure 5.5. There are some theoretical problems, however, concerning the proof of monotonicity using the likelihood ratio ordering. Although this was investigated by

Chang (1989), no theoretical proof has been given yet.

Embodied in the statistical framework as described in chapter 3, and under the assumption of monotonicity of this order relationship, the probabilities $Pr_{\mu}[(W,\tau) > (w,t_m)]$ and $Pr_{\mu}[(W,\tau) < (w,t_m)]$ can be expressed in terms of the absorption probabilities (3.2.a) and (3.2.b) as follows:

$$Pr_{\mu}[(W,\tau) > (w,t_m)] = \sum_{i=1}^{k} Q_i^{+}\left[\mu;\tau\frac{\mu}{\sigma_T} + \sqrt{\tau t_m}\left(\frac{W}{t_m} - \frac{\mu}{\sigma_T}\right)\right],$$

and

$$Pr_{\mu}[(W,\tau) < (w,t_m)] = \sum_{i=1}^{k} Q_i^{-}\left[\mu;\tau\frac{\mu}{\sigma_T} - \sqrt{\tau t_m}\left(\frac{W}{t_m} - \frac{\mu}{\sigma_T}\right)\right],$$

These probabilities are a function of all possible stopping times, irrespective the actual moment of stopping.

Point estimates

With an order relationship induced on the outcome space, the median unbiased point estimates μ^{hat} can be determined. For a given value (w,t_m), with $w \geq a_m$, μ^{hat} can be solved iteratively from

$$Pr_{\mu}[(W,\tau) > (w,t_m)] = \tfrac{1}{2},$$

or from

$$Pr_{\mu}[(W,\tau) < (w,t_m)] = \tfrac{1}{2},$$

when $w \leq b_m$. If none of the boundaries is crossed and the trial is terminated at the final analysis, both formula's can be used.* Point estimates obtained this way will be referred to as the median unbiased point estimates.

In table 5.7 the model of the ASPECT trial (as it was planned) is considered in order to illustrate the point estimation procedures. The interim analyses were planned at the information times 0.075, 0.20, 0.40, 0.60, and 0.82 (see chapter 3); the planned lower boundaries were -0.71, -1.15, -1.63, -1.70, and -1.81. The naive, intuitive and likelihood ratio point estimates of the incidence rate ratio for some hypothetical outcomes of the ASPECT trial are given. Due to the stopping rule, trial outcomes at the first interim analysis lead to very low point estimates, far away from E_{rel}. The intuitive point estimate (0.20) is by definition equal to the naive point estimate (0.20); the likelihood ratio point estimate (0.22) leads to a marginal 'adjustment' towards E_{rel}. If the trial is stopped at a later interim analysis, for instance interim analysis number 5, the likelihood ratio

* program ESTIM, available from the author upon request.

Table 5.7: Median unbiased estimates for the incidence rate ratio IRR corresponding to some hypothetical outcomes of the (planned) model of the ASPECT trial and based on different orderings.

i	W_i	naive	intuitive	LR
1	-0.84	0.20	0.20	0.22
2	-1.22	0.50	0.51	0.54
3	-1.85	0.60	0.62	0.63
	-2.45	0.50	0.55	0.53
4	-1.97	0.70	0.71	0.72
	-3.67	0.60	0.64	0.52
5	-2.69	0.70	0.75	0.72
	-5.01	0.50	0.73	0.52
6	0.00	1.00	1.00	1.00
	-0.98	0.90	0.90	0.90
	-3.28	0.70	0.80	0.72
	-6.11	0.50	0.79	0.52

i is the analysis number; $\mathbf{W_i}$ is the (hypothetical) outcome (see chapter 3); **naive** is the naive point estimate; **intuitive** is the point estimate based on the intuitive order relationship; **LR** is the point estimate based on the likelihood ratio order relationship. The times of interim analyses are $t_1=0.075$, $t_2=0.20$, $t_3=0.40$, $t_4=0.60$, and $t_5=0.82$, and the lower boundaries are $b_1=-0.71$, $b_2=-1.15$, $b_3=-1.63$, $b_4=-1.70$, and $b_5=-1.81$.

ordering leads to a marginal adjustment of the naive estimate (from 0.5 to 0.52 and from 0.70 to 0.72); the intuitive ordering leads to a more outspoken adjustment if the outcome W_5 moves away from the boundary value b_5: at $W_5=-5.01$ (and $b_5=-1.81$) the naive estimate (=0.50) and the intuitive estimate (=0.73) are substantially different. It should be mentioned, however, that such an extreme outcome ($W_5=-5.01$) is highly unlikely to occur. If the trials stops at its planned end and W is not far away from 0 than both the likelihood ratio and the intuitive estimates are the same as the naive estimates.

Confidence intervals

The basic idea behind the determination of a confidence interval for μ is that the lower limit of the (1-α) confidence interval is the smallest value of μ for which an event at least as extreme as the observed one has a probability of at least ½α, and similarly for the upper limit. Therefore, by inducing an ordering of the sample space, for any given (w,t_m), the upper and lower limits μ_U and μ_L can be determined similarly to the determination of the median unbiased point estimate. From

Table 5.8: 95% confidence intervals for the incidence rate ratio IRR corresponding to some hypothetical outcomes of the (planned) model of the ASPECT trial and based on different orderings.

i	W_i	naive	intuitive	LR
1	-0.84	0.07, 0.56	0.07, 0.56	0.07, 0.64
2	-1.22	0.30, 0.82	0.31, 0.87	0.33, 0.92
3	-1.85	0.43, 0.85	0.43 0.92	0.44, 0.91
	-2.45	0.35, 0.71	0.37, 0.90	0.37, 0.75
4	-1.97	0.53, 0.92	0.54, 0.97	0.54, 0.98
	-3.67	0.38, 0.67	0.45, 0.94	0.39, 0.69
5	-2.69	0.55, 0.89	0.58, 1.00	0.57, 0.93
	-5.01	0.39, 0.64	0.55, 1.00	0.40, 0.66
6	0.00	0.81, 1.23	0.81, 1.23	0.81, 1.23
	-0.98	0.73, 1.11	0.73, 1.12	0.73, 1.12
	-3.28	0.57, 0.87	0.63, 1.03	0.58, 0.91
	-6.11	0.40, 0.62	0.61, 1.03	0.41, 0.64

i is the analysis number; $\mathbf{W_i}$ is the (hypothetical) outcome (see chapter 3); **naive** is the naive point estimate; **intuitive** is the point estimate based on the intuitive order relationship; **LR** is the point estimate based on the likelihood ratio order relationship. The times of interim analyses are $t_1=0.075$, $t_2=0.20$, $t_3=0.40$, $t_4=0.60$, and $t_5=0.82$, and the lower boundaries are $b_1=-0.71$, $b_2=-1.15$, $b_3=-1.63$, $b_4=-1.70$, and $b_5=-1.81$.

$$Pr_{\mu}[(W,\tau) > (w,t_m)] = \tfrac{1}{2}\alpha,$$

and

$$Pr_{\mu}[(W,\tau) < (w,t_m)] = \tfrac{1}{2}\alpha,$$

μ_L and μ_U can be determined respectively, through numerical iteration methods (Chang, 1989)*. In the absence of monotonicity, the calculated confidence intervals would not be guaranteed to achieve their nominal level.

Based on the same hypothetical outcomes as in table 5.7 of the ASPECT trial the naive, intuitive, and likelihood ratio based 95% confidence limits for the incidence rate ratio IRR are calculated, as shown in table 5.8. The intuitive confidence intervals are mainly determined by the moment of stopping, and are (by definition) identical to the naive intervals for trials that stop at the first interim analysis. The likelihood ratio based confidence intervals are more dependent on the magnitude of the outcome at the moment of stopping. Regardless of the stopping time, as the value of W_i increases, the

* program ESTIM, available from the author upon request.

Table 5.9: One-sided P-values corresponding to some hypothetical outcomes of the (planned) model of the ASPECT trial and based on different orderings.

i	W_i	naive	intuitive	LR
1	-0.84	0.001	0.001	0.003
2	-1.22	0.003	0.008	0.011
3	-1.85	0.002	0.011	0.005
	-2.45	*	0.010	*
4	-1.97	0.005	0.017	0.019
	-3.67	*	0.013	*
5	-2.69	0.001	0.024	*
	-5.01	*	0.024	*
6	0.00	0.500	0.500	0.500
	-0.98	0.163	0.169	0.169
	-3.28	0.001	0.037	0.001
	-6.11	*	0.037	*

i is the analysis number; $\mathbf{W_i}$ is the (hypothetical) outcome (see chapter 3); **naive** is the naive point estimate; **intuitive** is the point estimate based on the intuitive order relationship; **LR** is the point estimate based on the likelihood ratio order relationship. The times of interim analyses are $t_1=0.075$, $t_2=0.20$, $t_3=0.40$, $t_4=0.60$, and $t_5=0.82$, and the lower boundaries are $b_1=-0.71$, $b_2=-1.15$, $b_3=-1.63$, $b_4=-1.70$, and $b_5=-1.81$.

resulting confidence interval becomes closer and closer to the naive interval (Rosner an Rosner, 1988). The likelihood ratio intervals seem somewhat shorter than the intuitive confidence intervals.

P-values

Given an order relationship as described in the above section, P-values can be calculated. Assume a null hypothesis of no treatment effect ($\mu_0=0$). For a given value (w,t_m), with $w \geq a_m$ or with $w \leq b_m$, one-sided P-values can be determined* by

$$p = Pr_0[(W,\tau) > (w,t_m)],$$

and

$$p = Pr_0[(W,\tau) < (w,t_m)],$$

respectively. If none of the boundaries is crossed and the trial is terminated at the final

* program ESTIM, available from the author upon request.

analysis, both formula's can be used. In a paper by Fairbanks and Madsen (1982) tables are presented for the intuitive ordering on an outcome space defined by Pocock boundaries.

In table 5.9 the P-values for the same hypothetical outcomes of the ASPECT trial as in the previous sections are given. The adjustments of the naive P-values by the two orderings are comparable as for the point and interval estimates.

COMMENTS

The estimation procedures described in this chapter are defined according to an order relationship that is induced on the outcome space. Although these methods are constructed to have exact overall coverage probability, they differ depending on which ordering is used to relate the outcome space to the parameter space. This leads to different estimates if one looks at the separate stopping moments (see tables 5.7, 5.8, and 5.9).

Effect estimates based on the likelihood ratio ordering become closer and closer to the corresponding naive effect estimates as the value of W increases, regardless of the stopping time. Furthermore, the later the trial stops, the closer these estimates will be to the naive estimates. This corroborates with the findings of this chapter: naive estimates, obtained at the trials planned end, lead to satisfactory results. A difficulty with the likelihood ratio ordering, however, is its dependence on future interim analyses. The outcome of a clinical trial is thereby partly determined by eventualities. Although effect estimates, based on the intuitive ordering, do not depend on future interim analyses and are therefore identical to the naive estimates for those clinical trials that stop at the first interim analysis, it is exactly this latter property does not corroborate with the simulation results of this chapter, which indicate that especially early stopping leads to biased estimates.

With regard to confidence intervals, Rosner and Tsiatis (1988) and Chang (1989) investigated differences between these two order relationships. Their (numerical) results, concerning coverage and width of the confidence intervals, slightly favored the estimation procedures based on the likelihood ratio ordering. This is corroborated by the findings in table 5.8. Kim (1988; and 1989) investigated point estimation procedures based on the intuitive order relationship. For different shapes of the type I error rate spending functions and for different analysis time schedules, both the median unbiased and the midpoint estimates lead to a substantial reduction of bias compared to the naive estimates.

REFERENCES

AIMS Trial Study Group. Effect of intravenous APSAC on mortality after acute myocardial infarction: preliminary report of a placebo-controlled clinical trial. *Lancet,* 12, 545-549, 1988.

Armitage P. *Sequential medical trials.* New York: Wiley, 1957.

Armitage P, McPherson CK, Rowe BC. Repeated significance testing on accumulating data. *Journal of the Royal Statistical Society Series A,* 132, 235-244, 1969.

CAST investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppressed after myocrdial infarction. *New England Journal of Medicine,* 321, 406-412, 1989.

Chang MN. Confidence intervals for a normal mean following a group sequential test. *Biometrics,* 45, 247-254, 1989.

Chang MN, O'Brien PC. Confidence intervals following group sequential tests. *Controlled Clinical Trials,* 7, 18-26, 1986.

CONSENSUS Trial Study Group. Effects of enalapril on mortality in severe congestive heart failure. *New England Journal of Medicine,* 316, 1429-1435, 1987.

Es GA van, Lubsen J, Strik R van. Point and interval estimation in clinical trials with interim analyses. *Controlled Clinical Trials,* 9, 253, 1988.

Fairbanks K, Madsen R. P values for tests using a repeated significance testing design. *Biometrika,* 69, 69-74, 1969.

Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials. *Statistics in Medicine,* 2, 167-174, 1983.

HINT Research Group. Early treatment of unstable angina in the coronary care unit: a randomised double blind, placebo controlled comparison of recurrent ischaemia in patients treated with nifedipine or metoprolol or both. *British Heart Journal,* 56, 400-413, 1986.

Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Statistics in Medicine,* 7, 1231-1242, 1988.

Jennison C, Turnbull BW. Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics,* 25, 49-58, 1983.

Julian DG, Levine B, Lubsen J, McFate Smith W. Monitoring methods, considerations and statement of the CONSENSUS ethical review committee. *New England Journal of Medicine, 1987,* 6 pages.

Kim K. Point estimation following group sequential tests. *Biometrics,* 45, 613-617, 1989.

Kim K, DeMets DL. Confidence intervals following group sequential tests in clinical trials. *Biometrics,* 43, 857-864, 1987.

Kim K. Improved approximation for estimation following closed sequential tests. *Biometrika,* 75, 121-128, 1988.

Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics,* 38, 153-162, 1982.

Rosner GL, Tsiatis AA. Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika,* 75, 723-729, 1988.

Rothman KJ. *Modern epidemiology.* Boston: Little Brown and Company, 1986.

Siegmund D. Estimation following sequential tests. *Biometrika,* 65, 341-9,1978.

Tsiatis AA, Rosner GL, Metha CR. Exact confidence intervals following a group sequential test. *Biometrics,* 40, 797-803, 1984.

Whitehead J. *The design and analysis of sequential trials.* Chichester: Ellis Horwood, 1983.

Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika,* 73, 573-581, 1986.

chapter 6
GENERAL DISCUSSION

In this chapter the usefulness of statistical methods in long-term clinical trials with interim analyses is discussed. A distinction is made between the decision whether or not to stop a trial and the estimation of the treatment effect after termination of the trial. Furthermore, the application of stopping rules and estimation methods is discussed in the light of some recent clinical trials with interim analyses.

STOPPING RULES

In long-term clinical trials interim analyses are usually performed by an independent Data Monitoring Committee (DMC) to assess early evidence of unexpected treatment differences or harmful side-effects. After every interim analysis the DMC advises whether to continue or to stop the trial. Prior knowledge, evolving knowledge, statistical considerations, medical judgement and ethical principles are involved in this decision. It is important that the investigators and the DMC define their strategy for interim analyses in advance and thereby avoid the influence of already accumulated data. However, owing to the intricacy and the incomparability of the factors involved it is not feasible to design an decision rule that will be adhered to in all situations. Experience shows us that almost always unpredicted circumstances occur which obviate the pre-defined decision rule. ASPECT is no exception: a much lower intake rate and a much lower overall mortality rate than expected. Nevertheless, a predefined stopping rule that pertains to the main treatment effect can and should be a major guidance for the DMC in its complicated task.

For funding purposes and in view of logistics it is usually necessary to set an upper limit on the number of patients to be accrued in a clinical trial. The determination of this upper limit is a trade-off between costs and time on the one hand and the precision desired by the organizers on the other hand. The amount of precision that is required to reach a conclusive result depends on the treatment effect as it is expected by the investigators relative to this minimal clinically relevant effect. More precision is needed for an effect close to E_{rel} than for an effect far away from it. In situations that the investigators are convinced that the magnitude of the treatment effect is much bigger than E_{rel}, a smaller trial with less precision might suffice. This implies the following considerations in the

design of a stopping rule:

- If the interim data suggest an effect close to E_{rel} the trial should be continued until its planned end and size.
- If the interim data suggest an effect that is substantially greater than E_{rel} less precision is required to prove that the treatment is beneficial and the trial might be terminated early.
- If the interim data suggest a detrimental effect or no effect the precision of the treatment effect is less relevant and the trial should not be continued to prove that the experimental treatment is either harmful or ineffective.

On a treatment effect scale this should lead to an asymmetrical stopping rule that is centered around E_{rel}. This strategy also holds when a 'range of equivalence', as proposed by Freedman et al. (1983), is considered instead of the single point E_{rel}.

Investigators are usually reluctant to stop a clinical trial early. The reduction of the number of patients by early stopping may not compensate for the loss of precision. Furthermore the investigators often have a very strong prior belief in the magnitude of the true treatment effect. Therefore early stopping because of a deviation from their expectations in the beginning of the trial is usually undesirable; the investigators' prior belief will overrule the scanty information from the trial. Stopping boundaries that are conservative in the early stages of the trial reflect this attitude (Freedman, 1988; Freedman and Spiegelhalter, 1989).

Various approaches are available to design a stopping rule for a clinical trial, their differences mainly being concerned with the underlying statistical assumptions. The group sequential designs as described in chapter 3 of this dissertation were originally defined to control both the type I error rate and the type II error rate of the trial. The shape of the boundaries is determined according to the requirements and insights of the investigators. Usually the number and timing of the interim analyses are specified in advance; modifications are often desirable, however. The conversion of the stopping boundaries to the type I error spending rate as described in chapter 3 allows for these modifications, although mis-specification of the total amount of information might jeopardize the type I error rate.

The general requirements mentioned above could also be accomplished using classical sequential methods. Stopping rules based on the theory of sequential experimentation were originally designed for continuously monitoring of the trial, thereby reducing the expected sample size of the trial while controlling for both the type I and the type II error rate (Wald, 1947; and Armitage, 1975). Recently these methods were adapted to allow

for discrete monitoring extending their applicability to survival trials and other long-term clinical trials (Whitehead et al., 1983a and 1983b).

Most stopping rules adopt statistical models based on hypothesis testing. A recent publication by Jennison and Turnbull (1989) and the subsequent discussion indicates a gradual shift to estimation. In this dissertation a (long-term) clinical trial is viewed as a tool to collect evidence concerning the magnitude of clinical benefit of a certain experimental treatment relative to a standard treatment and not merely to show a statistically significant difference between treatments. Effect estimation better corresponds to this objective. With regard to the design of a stopping rule, this implies that the emphasis is no longer on controlling the type I error rate.

Bayesian methods conceptually incorporate prior information in the design of the stopping rule. A pre-trial prior probability distribution of the magnitude of the treatment effect is specified, and the data collected are used to modify the prior distribution into a posterior distribution. The posterior distribution may be calculated at any stage of the trial and used to make a decision with regard to the future of the trial. This approach was advocated by Cornfield (1966a, and 1966b) in the sixties and seventies and was recently supported by Freedman (1988, and 1989) and Spiegelhalter (1988) amongst others. Although the concept is appealing, the quantification of the prior information is not straightforward and remains one of the major stumbling blocks for adoption of these methods. Freedman (1983) and Spiegelhalter (1988) are developing techniques aimed at the quantification of prior belief.

No matter which statistical approach is used, the utility of a stopping rule should always be evaluated with respect to the question whether it rightly reflects the whole of considerations of the investigators. A stopping rule for a particular trial should therefore be evaluated by the investigators in 'non-statistical' terms. This can be realized through the use of simulations of some intelligible hypothetical situations of the planned trial. Furthermore the effect of varying the number of looks and the shape of the boundaries can be illustrated through the same simulations, enabling the investigators to make intelligent decisions in the planning stage of the trial.

The stopping rule of the ASPECT trial, as described in chapter 5, was designed to control the type I error rate and to some lesser extent the type II error rate. The stopping boundaries were defined in terms of one-sided P-values, expressing the asymmetry of the treatment-placebo comparison. The shape of the boundaries expressed the conservatism of the ASPECT Policy Board related to the very early stopping of the trial. The times of the interim analyses were planned in advance. However, due to an unforeseen slow patient intake the timing of interim analyses differed from the plan. By

interpolation on the 'α-spending scale' (see chapter 3) the boundaries could now be determined for the actual moments of interim analysis.

The boundaries of the ASPECT stopping rule, originally (in 1983) determined through computer simulations, can now be obtained by the numerical procedures described in chapter 3. The statistical framework that is used (described in chapter 3) assumes the effect estimate E (or a transformation of E) to be approximately Gaussian distributed. Although the Gaussian approximation provides exact results only in the limiting situation of very large trials, it is well known that asymptotic statistics tend to be quite adequate in clinical trials with moderate or large size as described in this dissertation. The computer simulation results of the ASPECT trial (chapter 4) confirm the appropriateness of the approximation in this case. Nevertheless caution and judgement should always be exercised, especially in the early phase of a clinical trial when not much information has yet accumulated. For binary data, exact methods are available in these situations, for survival data currently more effective approximations are being investigated (Jennison and Turnbull, 1989).

EFFECT ESTIMATION

After termination of a clinical trial at one of the interim analyses or at its scheduled end, an estimate of the treatment effect must be generated from the data. The estimation procedure used should ideally lead to a valid point estimate with maximal precision.

The option to stop a trial prematurely alters the distribution of the final effect estimates. Whitehead (1986) investigated this for the maximum likelihood estimate after a sequential trial: the value of a maximum likelihood estimate will not be altered by the stopping rule, but the distribution of the maximum likelihood estimates will be affected, sometimes introducing substantial bias. The results in chapter 4 showed similar findings for the naive effect estimate (i.e. the usual fixed size estimate) of the incidence density rate ratio. Given the magnitude of the true effect, the median bias of the naive estimate depends on both the stopping rule and the stopping time. Simulations at a true treatment effect not far from minimal clinically relevant effect showed that the naive estimate is pushed away from the true treatment effect.

In this dissertation two methods of obtaining effect estimates, that take account of the interim analyses into account, are described. These methods are based on so-called orderings of the outcome space. The outcome of a clinical trial with interim analysis is represented by the effect estimate and the moment that the trial is stopped; the outcome space is the set of all possible outcomes of the trial. When an order relation is defined

among all points of the outcome space, the probability of a deviation more (or less) extreme in this order relation than the observed one can be computed. Consequently, within this ordering of the outcome space, a median unbiased point estimate and a confidence interval that has exact overall coverage can be obtained.

The two order relationships, which are described and illustrated, are the so-called intuitive ordering and the likelihood ratio ordering. The intuitive ordering, as introduced by Armitage (1957) and worked out by Siegmund (1978) and Tsiatis et al. (1984), considers only the actually observed outcome space at the moment of stopping. For trials that stop at the first interim analysis this means that the effect estimates will be identical to the naive estimates. The later the trial stops the more the estimates based on the intuitive ordering will be pulled away from the naive estimates towards a less extreme treatment effect. The likelihood ratio ordering, as defined by Rosner and Tsiatis (1988) and Chang (1989), is based on the entire outcome space of the trial, consisting of the interim analyses already performed and the analyses still to come. Regardless of the stopping time, the effect estimates based on the likelihood ratio ordering become closer and closer to the corresponding naive estimates as the value of the statistic W increases. Furthermore, the later the trial stops the closer these estimates will be to the naive estimates.

The results of the estimation methods as described depend on the order relationship that is defined on the outcome space. The fact that the choice of the order relationship seems haphazard undermines the credibility of these methods. Therefore criteria for the determination of confidence intervals in sequential trials should be established, such that choice of the order relationships can be evaluated according to these criteria. A drawback of these methods is that they require an explicit definition of the outcome space. For a clinical trial that stops early the specification of the outcome space for the likelihood ratio ordering is partly based on possible future events. Moreover, even the definition of actual outcome space is not straightforward. It cannot be verified if the predefined stopping boundaries would have been adhered to if other trial outcomes had occurred at an earlier stage. In the light of the foregoing the use of the term 'mathematically exact' is misleading. Further research concerning the robustness of these methods, with respect to mis-specifications of the outcome space, is needed.

The estimation methods described in this dissertation are based on the repeated sampling principle (Cox, 1974). According to this principle, the statistical procedures are to be assessed by their behavior in hypothetical repetitions under the same conditions. Another approach based on this repeated sampling principle (Cox, 1974) is the repeated confidence interval method described and advocated by Jennison and Turnbull (1989). This method combines aspects of estimation and testing in clinical trials with interim

analyses. However, the final point estimate will be biased and the confidence interval is unduly conservative. The suggestion by Hughes and Pocock (1988), which seems to utilize both the repeated sampling principle and the likelihood principle, that the Bayes posterior interval is 'adjusting' for early stopping is delusive. The Bayes posterior interval only 'adjusts' for the prior; the adjustment based on the likelihood ratio function is independent of the stopping rule (Cox, 1974).

In contrast to the repeated sampling principle, inferences based on the likelihood principle solely consider the data actually obtained. The likelihood principle takes into account the ratio of the probabilities of obtaining the observed results under various plausible hypotheses. Based on this principle problems concerning the sequential nature of the trial do not exist. Tijssen et al. (1987) suggest that a redefinition of the confidence interval as the posterior interval from an uninformative prior distribution will lead to a workable solution of the estimation problem. This approach, however, might lead to confusion, because it mixes properties of both the repeated sampling principle and the likelihood principle. Therefore, more theoretical work needs to be done to support these 'likelihood-based' estimates.

Although 'sequential analysis' is often used as the battleground of the repeated sampling and the likelihood view (Anscombe, 1963; Armitage, 1963; Cornfield, 1966a; Berry, 1985; and Spiegelhalter and Freedman, 1988), it is not the purpose of this dissertation to enter into this debate. However, it is the author's point of view that this debate is very fruitful for the development of statistical methodology in clinical trials with interim analyses. Lakatos (1974) put this viewpoint into words in a general context: 'The history of science has been and should be a history of competing research programmes, but it has not been and must not become a succession of periods of normal science: the sooner competition starts, the better for progress'.

RECENT CLINICAL TRIALS WITH INTERIM ANALYSES

In this dissertation a general approach of designing a stopping rule in terms of the effect estimate and its precision is described. Various methods are available in the statistical literature; one method is extensively described in this dissertation. However, the choice of a stopping rule should not solely depend on the statistical approach that is used but on all considerations involved in the decision making process. Because these considerations differ from trial to trial no unequivocal stopping rule can be dictated as is illustrated by some examples of recent clinical trials with interim analyses.

After completion of a clinical trial the results should be summarized, preferably by

estimates of the treatment effects with an indication of their precision. The fact that the usual estimation methods are affected by the interim analyses seems to be ignored, in the final reports of the clinical trials that are discussed.

Stopping rules

In long-term clinical trials it is obligatory to perform interim analyses at regular intervals to assess early evidence of unexpected treatment differences or side effects. Interim analyses should be performed by an independent Data Monitoring Committee (DMC) that has access to the unblinded data. In certain instances an interim analysis may lead to early termination of the trial. The decision to stop or to continue a trial is a cost-benefit problem involving precision, costs, and a surmise of the utility of the result. A stopping rule, based on the primary effect estimate and its precision, constitutes a major contribution to the complicated task of the DMC in executing the interim analyses. It must be acknowledged, however, that the formal stopping rule is only a useful guide in the decision-making process. It is not to be seen as a rigid rule for stopping the trial. A stopping rule cannot be designed to anticipate all contingencies. Considerations such as side effects or toxicity of treatment or evidence from an extraneous study may override the formal statistical stopping rule. This is illustrated by a clinical trial in patients with acute myocardial infarction with the objective to determine whether an invasive strategy with a thrombolytic agent (alteplase, rt-PA) and immediate percutaneous transluminal coronary angioplasty (PTCA) would be superior to a noninvasive strategy with the same medical treatment but without PTCA (Simoons et al., 1988). At interim analysis, which entailed 344 of the 400 planned patients, the trial was terminated early on recommendation of its DMC. The primary reason to terminate the trial was an unexpected higher mortality rate, a high rate of early recurrent myocardial ischemia, and other complications in the experimental group. Furthermore, the results concerning the endpoints of the trial, enzymatic infarct size and global left ventricular function indicated that immediate PTCA was not beneficial. In such unforeseen circumstances, which do often occur, one should rely on the wisdom and the experience of the DMC members, and not on a pre-defined stopping rule.

To avoid the influence of already accumulated data at least a formal plan for interim analyses, including a stopping rule, should be defined in advance. Furthermore, the definition of a stopping rule before the start of the trial is a very useful exercise in writing the protocol of the trial: it pinpoints the investigators to reflect on the eventualities that might occur during the course of the trial. It should be appreciated, however, that such stopping rules can only be designed if the new (experimental) treatment has been well defined and will be applied unchanged throughout the trial. In trials with a more explorative nature this might be impossible, as illustrated by a recent clinical trial with the

objective to evaluate the effect of thrombolysis compared to conventional treatment in patients with acute myocardial infarction (Simoons, 1985). At the start of the trial, in 1981, 'the investigators realized that thrombolytic therapy was still at its infancy. It was expected that improvements of the intervention might arise during the course of the trial. Therefore, modifications of the protocol would be allowed, if during the trial insights how to achieve optimal reperfusion would alter' (Vermeer, 1987). Interim analyses were essential for the conduct of this trial although a 'stopping rule' could not be defined. The results of the interim analyses of the ongoing trial were presented at several public meetings, and various policy decisions, such as modifications of the experimental treatment and the extension of the trial, were taken based on both the (promising) interim results and on new insights concerning recanalization procedures in patients with evolving myocardial infarction.

In general it can be said that the results of small trials reporting extreme treatment differences lack credibility. Therefore, early stopping may undermine the credibility of the trial and is not desirable. In the 1960s Meuwissen et al (1969) conducted a randomized placebo controlled trial to study the effect of oral anticoagulants on mortality in post-infarct patients. The trial was stopped early when the data showed a significant difference in mortality between the two treatment groups: 1 out of 68 patients under oral anticoagulant treatment died against 8 out of 70 under placebo treatment. Although one can, from an ethical point of view, accept the decision not to continue the trial, the results remain those from a small trial with an extreme result (an 87% mortality reduction) and cannot be taken as a 'proof' of efficacy of oral anticoagulants after myocardial infarction. In fact the role of oral anticoagulants after myocardial infarction is still under debate in the Netherlands and in other countries, and the ASPECT and the WARIS trials were designed 15 years later to a better estimate of the true treatment effect.

A predefined stopping rule or strategy might lead to a more conservative approach with regard to early stopping as illustrated by the APSAC Intervention Mortality Study (AIMS Trial Study Group, 1988). AIMS was a placebo-controlled mortality trial of a thrombolytic agent APSAC (anisoylated plasminogen streptokinase activator complex) after myocardial infarction. Interim analyses were planned according to an O'Brien and Fleming rule (1979). At the first interim analysis a clear beneficial effect of APSAC was observed, but the stopping boundaries were not crossed. At the second interim analysis, more than 1000 patients had entered the trial, it was decided to stop recruitment because the interim results showed a mortality reduction on APSAC treatment which exceeded the predefined stopping boundary. Additional investigation of the data supported this favorable trend of APSAC, while also other large trials (GISSI, 1987 and ISIS-2, 1988) had documented lower mortality after thrombolytic therapy, compared with conventional treatment. Hence, the DMC felt it would be unethical to continue to

allocation to placebo treatment. Their recommendation was ratified by the steering committee and patient entry was stopped. With respect to the termination of the AIMS trial after the second interim analysis, and not after the first interim analysis, Pocock and Hughes (1989) stated: 'Looking back we now feel it was a wise decision to continue to the second analysis, since it enhanced the trial's credibility and provided more reliable estimation of what appears to be a major treatment advance'.

Another example of a clinical trial that was stopped because the data suggested a treatment effect that was larger than anticipated is the COperative North Scandinavian ENalapril SUrvival Study (CONSENSUS Trial Study Group, 1987). The CONSENSUS trial was a randomized, double blind, placebo controlled clinical trial to study the effects on mortality of enalapril in severe congestive heart failure (CONSENSUS Trial Study Group, 1987). Based on the data of 244 patients, indicating a clear beneficial effect of enalapril compared to placebo, the trial was terminated although the protocol specified that 400 patients would be allocated. The Ethical Review Committee stated in its final report (Julian et al., 1987) that '... continuation was no longer justifiable on ethical grounds and of limited scientific value'.

The above examples illustrate that the investigators have to weigh their responsibilities concerning the patients in the trial receiving inferior treatment, which is referred to as individual ethics, against the provision of more comprehensive knowledge for treatment decisions on future patients, which is referred to as the collective ethics (Pocock and Hughes, 1989). The balance between individual and collective ethics is complex and becomes particularly difficult when the clinical trial is of major importance. In ISIS-2 (second International Study of Infarct Survival), another clinical trial of thrombolysis in myocardial infarction comparing streptokinase and placebo, the steering committee reported interim findings that showed a major reduction of in-hospital mortality on streptokinase for nearly 4000 patients treated within 4 hours of onset of symptoms (ISIS-2 Collaborative Group, 1988). They declared this was 'proof beyond reasonable doubt' that streptokinase was effective in patients treated within four hours of the onset of myocardial infarction. Nevertheless the trial coordinator allowed any investigator who wished to randomize additional patients in this group to do so. Finally, nearly 7500 patients have been randomized within 4 hours of onset of symptoms. This approach appears to indicate a shift of emphasis from individual ethics towards collective ethics.

In placebo-controlled clinical trials with an interim result indicating a detrimental or no effect of the index treatment, individual ethics should prevail. The Holland Interuniversity Nifedipine/metoprolol Trial (HINT) was a randomized, double blind, placebo controlled clinical trial of nifedipine, metoprolol, and nifedipine combined with metoprolol to estimate the effect of these four modes of treatment on recurrent ischaemia or myocardial

infarction within 48 hours in patients with unstable angina (HINT Research Group, 1986; and Tijssen et al., 1987). The trial was terminated early because the results that accumulated until the fourth interim analysis suggested that the risk of myocardial infarction was higher in patients assigned to nifedipine alone than in patients treated with the other trial medications, including placebo. The investigators considered it to be unethical to expose any further patients to mono-therapy with nifedipine only to prove that this treatment is harmful. This example illustrates the asymmetry of the stopping decision. Later analysis of the data, including those data that accumulated after the final interim analysis, demonstrated a beneficial trend of nifedipine in patients which entered the trial with beta blockers. It is to be regretted that this arm of the trial was also discontinued (Tijssen et al., 1987).

In the Cardiac Arrhythmia Suppression Trial (CAST) the above mentioned asymmetry was reflected by the stopping rule that was used. The boundaries for a harmful effect, based on stochastic curtailment, were less stringent than the boundaries for a positive effect. In CAST the flecainide and encainide treatment arms, drugs with Class IC antiarrhythmic action, were discontinued because the boundaries indicating harm were crossed (CAST Investigators, 1989). The other arm of CAST, comparing moricizine and placebo is still continuing.

During the course of a clinical trial insights concerning the treatment under investigation may change, requiring a more precise estimate of the treatment effect. If at the end of such trial the results are promising but, given the new precision requirements, not yet convincing, the trial might be prolonged. A clinical trial may also be extended if the basis of the sample size calculations turns out to be invalid. This occurs, for example, if the observed mortality in the reference or placebo group appears to be lower than expected during the course of the trial. The decision to prolong a trial, however, should be taken under comparable circumstances as to an interim analysis: an independent committee should evaluate the available results. The investigators should be kept unaware of the treatment allocation scheme. Before the data are shown to the independent committee, however, a set of guidelines, including a stopping rule, should be defined to reflect the investigators' considerations concerning the decision whether to prolong the trial or not. The Intravenous Streptokinase in Acute Myocardial infarction (ISAM) trial, comparing streptokinase and placebo therapy in myocardial infarction, was designed under the expectation of an 18% placebo mortality. Since the assumed mortality turned out to be overestimated the trial was extended from 860 to 1741 patients. In fact, the observed mortality in the placebo group appeared to be 7.1% (ISAM Study Group, 1986).

The above discussion demonstrates that a stopping rule selectively stops those trials early that show extreme outcomes relative to the investigators' expectations; the level

of extremity can be determined by the stopping rule. This also means that when a trial continues until its planned end the magnitude of the estimated treatment effect lies within the boundaries of the stopping rule and thereby more or less confirms the expectations of the investigators. Furthermore, if the trial has progressed until its planned end, there is sufficient precision to estimate this treatment effect.

Effect estimation

After completion of a clinical trial the results should be summarized, preferably by estimates of the treatment effects with an indication of their precision. According to classical statistical theory, estimation methods, that do not take account of the fact that interim analyses are performed, lead to an overestimation of the treatment effect. Some recently developed estimation methods based on classical statistical theory that 'adjust' for the fact that interim analyses were performed are described and discussed in this dissertation. At present, no definitive solution to the best method seems available.

In the final reports of the clinical trials mentioned above, the point and interval estimates as presented were calculated as if no interim analyses had been performed This approach may be justified with the so-called likelihood principle, which implies that the interpretation of the results should only be based on the data observed as such, and thus be independent of a stopping rule. In the HINT trial the use of naive estimates has been justified by implying a re-definition of effect estimates in terms of the likelihood function. However, the investigators state 'no workable solutions have yet been provided by formal statistical theory' for their 'pragmatic' approach of using naive estimates (Tijssen et al., 1987).

Early discontinuation of a clinical trial is usually related to an extreme treatment effect. In their final report the AIMS' investigators state that the striking reduction in 30-day mortality compared with the results of other trials 'requires cautious interpretation' (AIMS Trial Study Group, 1990). It should be appreciated that the treatment effect in AIMS was far greater than anticipated. In fact, the reduction in mortality in this trial is far greater than in any other trial of intravenous thrombolytic therapy (Arnold et al., 1989). Generally it was felt that the effect observed by the AIMS investigators was somewhat extreme. Apparently, prior views tend to attenuate treatment effects suggested in a trial that is terminated ahead of time. This attenuation of the treatment effect acquired through such 'inferential mechanism' does not depend on the underlying statistical principle that is used and must not be confused with the 'bias reduction' of the estimation methods described in this dissertation.

Finally, it is recommended to include a full description of the stopping rule (or policy)

of a clinical trial in the Methods section of the final report of the trial. Furthermore, discrepancies between the actual and the intended policy are to be explained. The readership of the report should be enabled to draw their own conclusions from the trial independently.

REFERENCES

AIMS Trial Study Group. Effect of intravenous APSAC on mortality after acute myocardial infarction: preliminary report of a placebo-controlled clinical trial. *Lancet,* 334, 545-549, 1988.

AIMS Trial Study Group. Long-term effects of intravenous anistreplase in acute myocardial infarction: final report of the AIMS study. *Lancet,* 335, 427-431, 1988.

Anscombe FJ. Sequential medical trials. *Journal of the American Statistical Association,* 58, 365-383, 1963.

Armitage P. *Sequential medical trials.* New York: Wiley, 1957.

Armitage P. *Sequential medical trials (second edition).* New York: Wiley, 1975.

Armitage P. Sequential medical trials: some comments on FJ Anscombe's paper. *Journal of the American Statistical Association,* 58, 384-387, 1963.

Arnold AER, Simoons ML, Lubsen J. Trombolytische therapie van het acute myocard infarct anno 1988. *Nederlands Tijdschrift voor Geneeskunde,* 133, 341-349, 1989.

Berry DA. Interim analyses in clinical trials: classical vs. Bayesian approaches. *Statistics in Medicine,* 4, 521-526, 1985.

CAST investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppressed after myocardial infarction. *New England Journal of Medicine,* 321, 406-412, 1989.

Chang MN. Confidence intervals for a normal mean following a group sequential test. *Biometrics,* 45, 247-254, 1989.

CONSENSUS Trial Study Group. Effects of enalapril on mortality in severe congestive heart failure. *New England Journal of Medicine,* 316, 1429-1435, 1987.

Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *American Statistician,* 20, 18-23, 1966a.

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

Cox DR, Hinkley DV. *Theoretical statistics.* London: Chapman and Hall, 1974.

Freedman LS. A comparison of Bayesian with frequentist methods of monitoring clinical trials data. *Proceedings of the XIVth International Biometric Conference,* 57-65, 1988.

Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials. *Statistics in Medicine,* 2, 167-174, 1983.

Freedman LS, Spiegelhalter DJ. The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician,* 32, 153-160, 1983.

Freedman LS, Spiegelhalter DJ. Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials,* 10, 357-367, 1989.

Gruppo Italiano per lo Studio della Streptochine-nase nell'Infarcto Miocardio. Long-term effects of intravenous thrombolysis in acute myocardial infarction: final report of the GISSI study. *Lancet,* 871-874, 1987.

HINT Research Group. Early treatment of unstable angina in the coronary care unit: a randomised double blind, placebo controlled comparison of recurrent ischaemia in patients treated with nifedipine or metoprolol or both. *British Heart Journal,* 56, 400-413, 1986.

Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Statistics in Medicine,* 7, 1231-1242, 1988.

ISAM Study Group. A prospective trial of intravenous streptokinase in acute myocardial infarction (I.S.A.M.). Mortality, morbidity, and infarct size at 21 days. *New England Journal of Medicine,* 314, 1465-1471, 1986.

ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17187 cases of suspected myocardial infarction: ISIS-2. *Lancet,* 349-360, 1988.

Jennison C, Turnbull BW. Interim anlyses: the repeated confidence interval approach. *Journal of the Royal Statistical Society B,* 51, 305-361, 1989.

Julian DG, Levine B, Lubsen J, McFate Smith W. Monitoring methods, considerations and statement of the CONSENSUS ethical review committee. *New England Journal of Medicine,* 1987 (6 pages).

Lakatos I. Falsifications and the methodology of scientific research programmes. In: *Criticism and the growth of knowledge.* Edited by Lakatos I and Musgrave A. Cambridge: Cambridge University Press, 1974.

Meuwissen OJAT, Vervoorn AC, Jordan FLJ, Nelemans FA. Double blind trial of long-term anticoagulant treatment after myocardial infarction. *Acta Medica Scandinavia,* 186, 361-368, 1969.

Pocock SJ, Hughes MD. Practical problems in interim analyses with regard to estimation. *Controlled Clinical Trials,* 10, 209S-221S, 1989.

Rosner GL, Tsiatis AA. Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika,* 75, 723-729, 1988.

Siegmund D. Estimation following sequential tests. *Biometrika* 65, 341-9, 1978.

Simoons ML, Serruys PW, Brand M vd, et al. Improved survival after early thrombolysis in acute myocardial infarction. *Lancet,* 578-582, 1985.

Simoons ML, Arnold AER, Betriu A, et al. Thrombolysis with tissue plasminogen activator in acute myocardial infarction: no additional benefit from immediate percutanous coronary angioplasty. *Lancet,* 197-203, 1988.

Spiegelhalter DJ, Freedman LS. Bayesian approaches to clinical trials. *Bayesian Statistics,* 3, 453-477, 1988.

Tijssen JGP, Domburg RT van, Lubsen J. Progress monitoring and termination. *European Heart Journal (supplement H),* 8, 79-85, 1987.

Tsiatis AA, Rosner GL, Metha CR. Exact confidence intervals following a group sequential test. *Biometrics,* 40, 797-803, 1984.

Van der Werf F, Arnold AER. Intravenous plasminogen activator and size of infarct, left ventricular function, and survival in acute myocardial infarction. *British Medical Journal,* 297, 1374-1379, 1988.

Vermeer F. *Thrombolysis in acute myocardial infarction.* Dissertation, Erasmus University Rotterdam, 1987.

Wald A. *Sequential analysis.* New York: Wiley, 1947.

Whitehead J. *The design and analysis of sequential trials.* Chichester: Wiley, 1983.

Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika,* 73, 573-581, 1986.

Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. *Biometrics* 39, 227-236, 1983a.

Whitehead J, Jones DR, Ellis SH. The analysis of a sequential clinical trial for the comparison of two lung cancer treatments. *Statistics in Medicine,* 2, 183-190, 1983b.

chapter 7

SUMMARY

In the final stage of the development of a new therapy usually a large-scale comparative clinical trial is used. In many clinical trials the effect of a treatment can only be assessed over a long period of time, or the length of time that is needed to recruit the required number of patients for the trial is long. During the course of such long-term clinical trials it is desirable to evaluate the accumulating results at regular time intervals. The interim results might indicate that continuation of the trial is meaningful, superfluous, or ethically unjustified. This dissertation describes statistical methods for the design and the analysis of long-term clinical trials with interim analyses within the classical statistical framework. The usefulness of these methods is investigated. The ASPECT trial, a trial that is currently being conducted in the Netherlands of the effect of anticoagulant therapy on mortality in patients after myocardial infarction, serves as a major example throughout this dissertation.

In a historical review of the application of interim analyses in clinical trials concepts and methods are introduced. The ASPECT trial, a double-blind placebo controlled trial in which 3500 to 4000 patients are recruited and followed over a period of six years, is extensively described. About once a year an independent committee conducts an interim analysis of the accumulated results. Based on the results and directed by predefined guidelines, including a statistical stopping rule, it is decided whether or not to continue the trial. The progress of the ASPECT trial is discussed by means of the interim analyses conducted sofar. (chapter 1)

Some general principles concerning the design, the conduct, and the analysis of (long-term) clinical trials are described. It is started from the principle that a clinical trial is intended to measure the effect of the experimental treatment relative to a standard treatment accurately and with sufficient precision. (chapter 2)

A statistical stopping rule constitutes a useful guide in the decision making process whether, given the interim results, to continue a clinical trial or not. A stopping rule is a decision rule, defined in terms of information time, that indicates on the basis of the magnitude of the effect estimate whether the trial is to be continued or not. First some suitable measures of treatment effect are described and the concept 'information time', a measure to quantify the information in the trial, is introduced. A statistical framework

of clinical trials with interim analyses is defined. Various stopping rules, that are known from the literature, are described according to this statistical framework. Flexibility of a stopping rule is required, because a long-term clinical trial cannot completely be planned in advance. For this reason the so-called α-spending function is defined. Application of the α-spending function is demonstrated by the ASPECT trial. (chapter 3)

The asymmetrical stopping rule for the ASPECT trial is obtained by computer simulations. The stopping rule will terminate the trial if an unexpected big effect of the experimental treatment can be demonstrated with sufficient evidence. The stopping rule also provides for the possibility to stop the trial early when the interim results indicate a negative or no treatment effect. (chapter 4)

After completion of a trial the results are summarized, preferably by estimates of the treatment effects and their precision. By computer simulations it is shown that the usual estimation methods, which do not take into account that interim analyses have been performed, will overestimate the true treatment effect. Two recently published estimation methods, that do take account of the fact that interim analyses have been performed, are described. These methods are illustrated by some postulated data of the ASPECT trial. (chapter 5)

The methods, that were investigated in this dissertation, are discussed in general and in the light of some recent clinical trials in the field of coronary heart disease. A stopping rule should be defined before the start of a clinical trial and must reflect the considerations of the investigators at that moment. Adequate statistical methods are available to design flexible stopping rules. However, it should be realized that unforeseen contingencies might occur during the course of the trial. Furthermore, a stopping rule only takes into account an incomplete measure of the treatment effect. Therefore, a stopping rule is not to be seen as a rigid rule for terminating a clinical trial. Concerning estimation of the treatment effect after a clinical trial with interim analyses no clear-cut solution seems available as yet. More research is needed. The clinical trials with interim analyses, that are discussed in this chapter, reported effect estimates as if no interim analyses were performed. (chapter 6)

SAMENVATTING

In de laatste fase van de ontwikkeling van een nieuwe therapie wordt veelal gebruik gemaakt van een grootschalig vergelijkend klinisch experiment. In veel klinische experimenten is het effect van een therapie slechts over een langere periode te beoordelen, of duurt het lang om het benodigde aantal patienten in het onderzoek te betrekken. Gedurende het verloop van zo'n langdurig experiment is het wenselijk om de resultaten enkele malen tussentijds te evalueren. Op grond van de tussentijdse resultaten is dan vast te stellen of het voortzetten van het experiment zinvol, overbodig of eventueel ethisch onverantwoord is. Dit proefschrift beschrijft statistisch methoden voor het opzetten en analyseren van langdurige klinische experimenten waarbij de resultaten tussentijds worden geevalueerd. De toepasbaarheid van de methoden wordt onderzocht. Het ASPECT onderzoek, een in Nederland lopend onderzoek naar het effect van antistolling op mortaliteit bij patienten die een hartinfarct hebben doorgemaakt, dient als voorbeeld door het gehele proefschrift.

Aan de hand van een historisch overzicht over toepassing van tussentijdse evaluaties in langdurige klinische experimenten worden begrippen en methoden geintroduceerd. Het ASPECT onderzoek, een dubbelblind placebo-gecontroleerd onderzoek, waarin gedurende zes jaar 3500 tot 4000 patienten zullen worden toegelaten en vervolgd, wordt in detail beschreven. Ongeveer eens per jaar voert een onafhankelijke commissie een tussentijdse evaluatie van de tot dan toe verzamelde gegevens uit. Op grond van de resultaten wordt hierbij mede aan de hand van vooraf opgestelde richtlijnen, waaronder een statistische stopregel, besloten of het onderzoek al dan niet zal worden voortgezet. De voortgang van het ASPECT onderzoek wordt aan de hand van de tot nu toe uitgevoerde tussentijdse evaluaties besproken. (hoofdstuk 1)

Enige algemene principes betreffende het opzetten, het uitvoeren, en het analyseren van (langdurige) klinische experimenten worden beschreven. Hierbij wordt ervan uitgegaan dat een klinisch experiment is bedoeld om de grootte van het effect van de experimentele behandeling ten opzichte van de standaard behandeling met de gewenste nauwkeurigheid te meten. (hoofdstuk 2)

Een statistische stopregel biedt nuttige richtlijnen in het besliskundige proces betreffende het al dan niet voortijdig stoppen van een klinisch experiment op basis van de tussentijdse resultaten. Een stopregel is een beslisregel, gedefinieerd in termen van informatie tijd, die op basis van de op dat moment gemeten grootte van de effect maat aangeeft of het onderzoek voortgezet dan wel gestopt moet worden. Allereerst worden enkele effect maten beschreven en wordt het begrip 'informatie tijd' geintroduceerd. Dit is een maat om de voortgang van het onderzoek uit te drukken in de hoeveelheid verzamelde informatie. Een statistisch model van een klinisch experiment met tussentijdse evaluaties wordt gedefinieerd. Enige uit de literatuur bekende stopregels

worden volgens dit model beschreven. Omdat een langdurig klinisch experiment niet vooraf te plannen, dient een stopregel flexibel te zijn. Daarom wordt de zogenaamde 'α-spending' functie gedefinieerd. Toepassing van deze 'α-spending' functie wordt geillustreerd aan de hand van het ASPECT onderzoek. (hoofdstuk 3)

Een asymmetrische stopregel voor het ASPECT onderzoek wordt bepaald met behulp van computer simulaties. De stopregel zal het onderzoek voortijdig beeindigen indien een overwacht groot positief effect van de experimentele behandeling met voldoende zekerheid kan worden aangetoond. De stopregel voorziet ook in de mogelijkheid om het experiment voortijdig te stoppen indien de tussentijdse resultaten een negatief of geen behandelings effect aantonen. (hoofdstuk 4)

Na het beeindigen van een klinisch experiment worden aan de hand van de verzamelde gegevens de grootte van het behandelingseffect met een indicatie van de nauwkeurigheid geschat. Met behulp van computer simulaties wordt aangetoond dat de gebruikelijke schattings-methoden, waarbij geen rekening wordt gehouden met het feit dat tussentijdse evaluaties zijn uitgevoerd, het werkelijke behandelingseffect overschatten. Twee recent gepubliceerde schattings-methoden, die wel gebruik maken van het feit dat tussentijdse evaluaties zijn uitgevoerd, worden beschreven. Deze methoden worden geillustreerd aan de hand van enkele hypothetische uitkomsten van het ASPECT onderzoek. (hoofdstuk 5)

De in dit proefschrift onderzochte methoden worden in het algemeen en aan de hand van een aantal recente klinische experimenten op het gebied van hart en vaatziekten besproken. Een stopregel dient voor de start het onderzoek gedefinieerd te worden en moet een reflectie zijn van de overwegingen van de onderzoekers op dat moment. Adequate statistische methoden om een flexibele stopregel te definieren zijn voorhanden. Er dient echter rekening te worden gehouden met het feit dat er tijdens de uitvoering van het experiment onverwachte gebeurtenissen kunnen optreden. Verder beschouwt een stopregel een onvolledige maat voor het behandelings effect. Een stopregel mag daarom niet gezien worden als een dwingend voorschrift om een klinisch experiment te stoppen. Wat betreft het schatten van het behandelings effect na het beeindigen van een klinisch experiment met tussentijdse evaluaties is nog geen eenduidige oplossing voorhanden. Verder onderzoek is gewenst. De klinische experimenten, die in dit hoofdstuk beschreven worden, rapporteerden effect schattingen zonder rekening te houden met het feit dat er tussentijdse evaluaties werden uitgevoerd. (hoofdstuk 6)

NAWOORD

Het schrijven van een proefschrift is te vergelijken met het maken een lange treinreis. De voorbereiding meegerekend neemt de reis een lange periode van iemands leven in beslag. Het eindoel staat de reiziger steeds voor ogen. Echter, eenmaal daar aangekomen blijkt het doel opeens veel minder belangrijk dan de weg er naar toe.

Het begon allemaal met mijn afstudeerprojekt aan de Technische Hoogeschool in Delft. Via mijn afstudeerhoogleraar Prof. Ir. J.W. Sieben en Prof. R. van Strik van de Instituut Biostatistica kwam ik op de afdeling klinische epidemiologie van het Thoraxcentrum terecht, toen nog onder leiding van Dr. J. Lubsen. Onder direkte begeleiding van Drs. J.G.P. Tijssen, Prof. R. van Strik en Dr. H.J.L. van Oorschot (de laatste vanuit de vakgroep Statistiek, Stochastiek, en Operations Research in Delft) voltooide ik mijn afstudeerscriptie 'Stopregels voor tussentijdse evaluatie bij gerandomiseerde klinische proeven'. Het cijfer dat mij bij de afronding van dit afstudeerprojekt werd toebedeeld, kreeg ik onder voorwaarde dat ik beloofde het onderzoek voort te zetten. Het idee voor de reis was er!

Mijn reisleider was Prof. R. van Strik; professor van Strik, vanaf het begin was u bij mijn wetenschappelijke ontdekkingsreis betrokken. Met al mijn problemen kon ik bij u terecht. U liet mij steeds vrij in mijn keuzes, mits ik ze kon verantwoorden. Deze manier van begeleiden heb ik altijd bijzonder op prijs gesteld. Ook de mensen van uw afdeling waren behulpzaam. Met name noem ik Dr. Th. Stijnen; Theo, jij hebt de vele versies van dit proefschrift gelezen en bekommentarieerd. Met jouw hulp kwam ik langs de soms moeilijke statistisch trajekten.

De kunst van het reizen heb ik geleerd van Dr. J.G.P. Tijssen; Jan, het begon allemaal met jouw idee. Dat idee werd eerst uitgewerkt tot mijn afstudeerscriptie, daarna volgde het daaraan gekoppelde artikel, en nu is er dit proefschrift. In al deze fases is jouw hand duidelijk zichtbaar. Ik hoop dat dit niet het einde van onze samenwerking is.

De voorbereiding van de reis geschiedde mede onder de hoede van Prof. Dr. J. Lubsen; Koos, de door jou opgezette afdeling klinische epidemiologie bood mij de unieke combinatie van praktijk en theorie als werkomgeving.

De trein was het Thoraxcentrum, altijd op volle snelheid vooruit! Van Prof. Dr. J.R.T.C. Roelandt kreeg ik de vrijheid om de afgelopen 10 maanden bijna volledig aan mijn proefschrift te mogen wijden. Prof. Dr. J. Pool was bereid om in de leescommissie plaats te nemen, en voorzag het manuscript van nuttige kommentaren. Prof. Dr. M.L. Simoons leverde een essentiele bijdrage aan hoofstuk 6 van het proefschrift. De wagons waarin ik gedurende mijn reis heb vertoefd zijn de afdeling klinische en experimentele informatieverwerking (AKEI), onder leiding van Ir. C. Zeelenberg, en de

afdeling klinische epidemiologie (KLEP), onder leiding van Dr. J.W. Deckers en Dr. J.G.P. Tijssen. Op beide afdelingen heb ik me altijd thuisgevoeld. Met name noem ik Cees Zeelenberg, die mij de ruimte liet om naast mijn werk ook aan dit proefschrift te werken; mijn collega's Ron van Domburg, Max Patijn, en Leidy Braaksma, die het afgelopen jaar veel van mijn taken hebben overgenomen; en Ale Algra, collega, kamergenoot, vriend, en paranimf.

Een andere trein die dezelfde richting opging was het ASPECT onderzoek, met Prof. Dr. J. v.d. Meer als voorzitter van de Policy Board and Drs. J.J.C. Jonker als 'Study Chairman'. Ik wens de ASPECT trein, die zich reeds in een vergevorderd stadium bevindt, veel succes toe.

Onderweg heb ik een aantal belangrijke tussenstations aangedaan, die bij het tot stand komen van dit verslag van wezenlijk belang zijn geweest. Bij het CONSENSUS en het ASPECT onderzoek mocht ik de tussentijdse evaluaties van dichtbij meemaken. De leden van de 'Ethical Review Committee' van de CONSENSUS trial (onder voorzitterschap van Prof. Dr. J. Lubsen) en de leden van de 'Data Monitoring Committee' van het ASPECT onderzoek (onder voorzitterschap van Prof. Dr. A. Hofman) boden mij het inzicht dat er bij een tussentijdse evaluatie meer komt kijken dan een stopregel. Another station along the line was the Department of Applied Statistics of the University of Reading. Dr. J. Whitehead, reader at this department and one of the world's major statistical experts in the field of sequential clinical trials, offered me the opportunity to discuss my thesis with him. John, your course on sequential clinical trials made me see the light at the end of the tunnel.

Mijn tolk was de heer R. Finch, die vele uren heeft besteed aan het verfraaien en verbeteren van de tekst. Met plezier denk ik terug aan de altijd leerzame besprekingen.

Gedurende mijn reis heb ik veel mensen ontmoet. Zij allen vormden de weg naar het einddoel van de reis.

Onderweg ben ik ook Inge tegengekomen; met haar en met Lonneke hoop ik nog veel te kunnen reizen.

april 1990 *Gerrit-Anne*

CURRICULUM VITAE

Gerrit-Anne van Es werd op 15 juni 1959 te Polsbroek geboren. In 1977 behaalde hij het eindexamen atheneum-B aan de Rijksscholengemeenschap Goeree en Overflakkee te Middelharnis. In hetzelfde jaar begon hij zijn studie aan de afdeling Wiskunde en Informatica van de Technische Universiteit in Delft. In 1984 beeindigde hij zijn studie met het behalen van het ingenieursdiploma bij de vakgroep Statistiek, Stochastiek, en Operations Research. Vanaf 1983 is hij als technisch wetenschappelijk medewerker verbonden aan het Thoraxcentrum van het Dijkzigt Ziekenhuis in Rotterdam.

LIST OF ABBREVIATIONS

DMC = data monitoring committee

E = effect measure
E_{rel} = minimal clinically relevant effect

RD = cumulative incidence rate difference
RR = cumulative incidence rate ratio
IRD = incidence density rate difference
IRR = incidence density rate ratio
HR = hazard ratio
(lower case equivalents refer to estimates)

k = total number of analyses
i = actual analysis ($i=1,...,k$)
S_i = Gaussian distributed statistic derived from E at i
σ_i^2 = variance of S_i
σ_T^2 = variance of S_k
t_i = information time at i ($= \sigma_T^2/\sigma_i^2$)
Z_i = S_i/σ_i
W_i = Z_i/t_i

a_i = upper boundary value at i
b_i = lower boundary value at i
P_i = boundary crossing (or absorption) probability at or before i

α = false positive rate (type I error rate)
Φ = standard Gaussian cumulative distribution function