STOPPING GUIDELINES VS STOPPING RULES:
A PRACTITIONER'S POINT OF VIEW

David L. DeMets, Ph.D.
Departments of Statistics and Human Oncology
University of Wisconsin-Madison
Room 6763 Medical Science Center
420 North Charter Street
Madison, Wisconsin 53706

TECHNICAL REPORT #26
AUGUST 1984

Typist: Terry Metcalf

STOPPING GUIDELINES VS STOPPING RULES:
A PRACTITIONER'S POINT OF VIEW

David L. DeMets, Ph.D.

Departments of Statistics and Human Oncology
University of Wisconsin-Madison
Room 6763 Medical Science Center
420 North Charter Street
Madison, Wisconsin  53706

ABSTRACT

Monitoring interim accumulating data in a clinical trial for evidence of therapeutic benefit or toxicity is a frequent policy, usually carried out by an independent scientific committee. While statistical methodology has been developed to assess the significance of these interim analyses, such methods should not be viewed as absolute rules but only serve as useful guides. The decision process to terminate a trial early is very complex and many factors must be taken into account. The complexity of this decision process is illustrated by reviewing the experience of several recent clinical trials.

INTRODUCTION

The randomized clinical trial has become one of the primary means of testing the benefits of a new therapy or procedure
against established practice as a control. During the past two
decades, the methodology for conducting clinical trials has
evolved, especially for multicenter trials. This methodology has
been described by authors such as Friedman, Furberg, and DeMets
(1981). Most clinical trials sponsored by the National Institutes
of Health (NIH) follow a particular model, referred to here as the
NIH model although it is used more broadly than simply NIH
activity. In general terms, the NIH model has various functioning
bodies which include: 1) the participating investigators and
clinics, 2) an executive or steering committee representing the
investigators, 3) central laboratories as needed, such as a
clinical chemistry or an electrocardiogram reading laboratory,
4) a statistical-data management center, and 5) a policy advisory
and data monitoring committee. Within this administrative
structure, each component contributes to the overall design,
conduct, quality, and analyses of the clinical trial. The purpose
of this paper is to discuss, by way of examples, how many factors
must be considered in the decision making process. At this time
many clinical trials have been conducted with a data monitoring
committee. Unfortunately, only a few trials have been published
on this experience, but those available will be the focus of this
discussion.

THE REPEATED SIGNIFICANCE TESTING PROBLEM

For the NIH clinical trials model, the data monitoring
committee is an independent group of scientists, not directly
involved with patient recruitment and care for the particular
trial, who periodically review progress of the trial. This
committee often includes physicians, epidemiologists,
statisticians and an ethicist. The responsibilities of this
committee may be broadly defined but generally include the ethical
and scientific responsibility of monitoring for evidence of
possible toxicity or early therapeutic benefit. The rationale for
such a committee has been described by Shaw and Chalmers (1970),
Chalmers (1979), and Canner (1979). One important role is to remove from the participating investigator the possible ethical dilemma of seeing interim data and yet having to recruit additional patients.

The process of repeatedly reviewing data has important implications for the evaluation of statistical analyses and tests of hypotheses. One consequence is that the Type I error or false positive rate is increased if a classical critical value is used at each test of hypothesis. For example, the probability of falsely rejecting the null hypothesis is increased to levels beyond 0.05 if a critical value of 1.96 were used at each test, the amount depending on the number of such repeated evaluations. Statisticians have long recognized the repeated testing issue and have proposed various methods for preserving traditional Type I error levels. Many of the early sequential models were found not to be used very often. A review of these issues is provided by Gail (1982) and DeMets and Lan (1984). In recent years, however, two basic ideas have been proposed which seem to be useful. Pocock (1977) and later O'Brien and Fleming (1979) introduced a data monitoring concept, usually referred to as group sequential boundaries, which adjusts the critical value used at each test to a more conservative (i.e., larger) value. This adjustment allows a prespecified α level to be achieved. Lan and DeMets (1983) have further developed this idea and Tsiatis (1981), for example, has shown how to use group sequential methods in survival analysis.

A second idea, referred to as stochastic curtailment, considers whether the interim results are so extreme, either for or against the null hypothesis, that the conclusions from test of hypothesis to be performed at the end of the trial are not likely to change. If the conclusions are unlikely to change, early termination of the trial may be considered. Since the point of reference is always what would happen at the end of the study if it were continued, the Type I error is also controlled. This
approach was in part developed by Lan, Simon, and Halperin (1982) and the application discussed in Halperin, Lan, Ware, Johnson, and DeMets (1982).

Given that such statistical models have been proposed and are available, the temptation might be to strictly adhere to them in the monitoring and decision making process. The models, in fact, assume that the trial would be terminated as soon as the specific criteria are met. However, the decision process to terminate a study early is a very complex one, involving many more factors than simply observing whether or not a boundary is exceeded, a point of view held by Meier (1979), the Coronary Drug Project Research Group (1981), DeMets, Hardy, Friedman, and Lan (1984), and Armitage (1979). To use these models as strict stopping rules would be unwise. Nevertheless, such models offer a practical guide to data monitoring committees in assessing the "significance" of observed interim results.

UNIVERSITY GROUP DIABETES PROJECT - AN EARLY EXPERIENCE

One of the earliest multicenter clinical trials sponsored by the National Institutes of Health was the University Group Diabetes Program or the UGDP (1970, 1971). This was a randomized double blind trial in adult onset diabetics using a placebo control group and four therapeutic strategies. The four therapies were: 1) a fixed dose of insulin, 2) a variable dose of insulin, 3) tolbutamide alone, and 4) phenformin alone which was added later in the study. The primary outcome variables were measures of retinal damage in addition to possible morbidity, such as infection.

However, in 1970 the data monitoring committee observed what was felt to be an excess cardiovascular mortality in the tolbutamide group compared to the placebo group (12.7% vs. 4.9%). The total mortality rate was not statistically significant, but was in the same direction (14.7% vs. 10.2%).
This treatment group was subsequently terminated. Later, the phenformin group was also terminated due to excess total mortality (15.2% vs. 9.4%) as well as excess cardiovascular mortality (12.7% vs. 3.1%). Regardless of treatment, diabetics are at higher risk of cardiovascular disease than the non-diabetic population. As summarized by Kolata (1979), the tolbutamide decision in particular has caused a great deal of controversy and, in fact, led to an NIH review of the UGDP by a special committee appointed by the Biometric Society (see the Report of the Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents, (1975)).

Among the criticisms of the decision was the fact that tolbutamide patients had more cardiovascular risk factors than did placebo patients, possibly explaining the observed excess deaths. Subsequent analyses by the special committee suggested the deaths to be excessive even after adjusting for the baseline risk factor imbalance. Another criticism was the issue of accurately determining cause of death, particularly for the tolbutamide group. This, of course, is a major concern for any mortality study and is a reason why cause specific mortality in itself is not adequate as a sole outcome measure.

While the controversy surrounding the UGDP may not be resolved, it is also unlikely that any of the data monitoring methods now available would have altered the UGDP decision process as described. For sure, the decision process was more involved than simply noting whether a test statistic was impressively large.

THE CORONARY DRUG PROJECT

Another one of the very early NIH multicenter clinical trials was the Coronary Drug Project (CDP), a randomized double blind trial to evaluate the effectiveness of lipid-lowering drugs in the secondary prevention of coronary heart disease under the direction
of the Coronary Drug Project Research Group (1970, 1972, 1973, 1975). The study evaluated five drug regimens (different drugs or varying dosages) compared to a placebo. During a three year period from 1966-1969, 53 participating clinical centers recruited over 8,000 patients who had a recent documented myocardial infarction. Patients were followed carefully from time of recruitment until 1974, a minimum follow-up of five years. At randomization, patients were classified into two risk groups, one being a more complicated myocardial infarction situation.

A data and safety monitoring committee (DSMC) was established which met periodically, approximately every six months, during the follow-up period. The DSMC membership included statisticians, physicians and basic scientists. This committee made recommendations to an NIH advisory committee and to the NIH. The structure of the CDP became a prototype for many subsequent NIH funded clinical trials, a structure referred to earlier as the NIH model. The CDP did, in fact, terminate three of the treatment regimens upon recommendation of the DSMC and that process has been described in detail by the Coronary Drug Project Research Group (1981). A brief summary is presented here.

Decision 1: Termination of High Dose Estrogen

Approximately six months after the last patient was recruited, the high dose estrogen group was observed to have increased rate of non-fatal myocardial infarction (MI) compared to placebo (6.2% vs 3.2%, z=4.11) as well as increased total mortality (8.1% vs 6.9%, z=1.33). One primary concern was the potential for bias in the diagnosis of non-fatal myocardial infarction, given that estrogen therapy has side effects which could potentially unblind the patient and physician. After extensive analysis investigating possible over or under diagnosis, no evidence was found that such biases were present. Further analyses then focused on risk subgroups. The non-fatal MI was primarily seen in the lower risk group (6.7% vs 2.9%, z=4.3) but the mortality
difference was in the higher risk group (13.9% vs 8.5%, z=3.0). In this case, the decision was reached after some debate to terminate the high dose estrogen in both the high and low risk groups, thus dropping that arm of the trial. If only a single outcome variable had been considered, different decisions might have been reached.

Decision 2: Termination of Dextrothyroxine

A year and a half after the estrogen decision, a decision was reached to discontinue to dextrothyroxine (DT4) treatment arm, due to increased mortality overall (14.8% vs 12.5%) and especially in the higher risk subgroup (22.5% vs 15.4%, z=3.1). The decision process focused on whether this result was consistent across certain subgroups; that is, could this adverse result be further isolated. After numerous subgroups had been defined, one such set (subgroups A and B) drew special attention because the mortality rate was in favor of DT4 in subgroup A (4.1% vs 7.7%, z=-2.6) and against (16.4% vs 11.2%, z=3.3) in subgroup B. Recognizing the dangers of multiple and repeated subgroup analyses, the DMSC decided to wait until the next meeting to review subsequent mortality in subgroups A and B. The results continued to suggest an adverse effect in subgroup B (5.4% vs 3.8%), but there was no trend for a beneficial effect in subgroup A. Based on that information, along with previous analyses, the decision to terminate was reached. The authors comment that if this particular subgrouping had been stated in advance, the decision process might have been less complicated. Even so, however, there is no guarantee that subgroups stated in advance will be the only ones of concern during the trial.

Decision 3: Termination of Low Dose Estrogen

In 1973, the low dose estrogen treatment regimen was dropped due to an excess of venous thromboembolism, an excess of cancer deaths (not statistically significant) and total mortality (19.9% vs 18.8%, also not statistically significant). The
decision considered that since there was a slight adverse trend, was there any chance at this stage in the trial of showing a beneficial effect at the scheduled end of the trial? This is an early application of the curtailed sampling idea. Furthermore, the DSMC did not want to prove beyond a doubt that estrogen was harmful, an example of asymmetric boundaries. After a substantial amount of analyses, it was felt very unlikely, if not impossible, for the low dose estrogen group to achieve significantly lower mortality than the placebo group. It was estimated, for example, that the future percentage of deaths would have to be 1.5% in low dose estrogen and 6.4% in placebo for the overall comparison to be just significant. Given this result and the other trends as well as the earlier high dose estrogen decision, low dose estrogen was terminated.

Decision 4: Continuation of the Clofibrate Group

One treatment group that was not terminated during the trial was the clofibrate treated group. At the end of the trial mortality was 25.5% in clofibrate vs. 25.4% in placebo (see Figure 1). During the course of the trial, however, the results were more intriguing. Early on after 20-30 months into the trial, the test statistic fluctuated around -2.0, in favor of clofibrate. By 40 months, the test statistic was slightly positive, in favor of placebo. At five years, the test statistic was again near -2.0 with the eventual test statistic to be nearly zero (see Figure 2). Interestingly, this pattern is seen but not with such wide swings if the life table results are used as a retrospective look. The DSMC did recognize the need for conservatism in the interpretation of interim analyses and used methods available at that time.

The CDP group concluded their experience by stating that "Although a number of rather sophisticated statistical tools are available in the decision making process, these are at best red flags that warn of possible treatment problems and can never be
used by themselves as hard and fast decision rules". Subsequent published experience to be described below only affirms their opinion.

 **NOCTURNAL OXYGEN THERAPY TRIAL EXPERIENCE**

The Nocturnal Oxygen Therapy Trial (NOTT) was an NIH sponsored randomized clinical trial under the direction of the Nocturnal Oxygen Therapy Group (1980) comparing two levels of oxygen therapy to individuals with advanced chronic obstructive pulmonary disease. The NOTT entered 201 patients across six clinical centers. The two groups were those who received standard continuous oxygen therapy (COT) compared with those who received only nocturnal oxygen therapy (NOT). Since COT was considered to be the standard, the hypothesis was whether less oxygen could be administered and still have the same benefits; that is, was 12 hours of oxygen as good as 24 hours? The primary outcome variables were pulmonary function and quality of life measures. As in the CDP study, a policy advisory and data monitoring committee was established to review interim data. The monitoring experience of this trial has been described previously by DeMets et al. (1982).

One aspect of the data monitoring is of particular interest. While mortality was not defined to be the primary outcome variable, during the course of the trial a mortality trend favoring COT began to emerge (see Table I). This mortality result was not anticipated as a benefit and so it drew considerable attention. At a committee meeting about one year prior to the scheduled study termination, the significance levels for mortality comparisons between therapies considering all randomized patients were in the neighborhood of 0.10, using the log rank statistic. One of the questions raised, similar to the CDP dextrothyroxine situation, was whether this mortality trend was consistent across subgroups.
A special subcommittee was asked to review the mortality data carefully before the next scheduled meeting. One such subgroup was defined on the basis of one second forced expiratory volume (FEV₁), an index of pulmonary function with lower levels indicating more severe disease. At an interim meeting in the summer of 1979, the analysis for this subgroup gave a significance level of 0.01 for survival curve comparisons. Extensive analyses were done to assure that the result was not due to baseline imbalances between the two treatment groups. The debate centered on whether the trial should be stopped entirely, stopped only for the low FEV₁ subgroup or continued. To stop only the low FEV₁ subgroup could, in practice, have had the effect of terminating the entire study and the subgroup issue would not be totally resolved. Furthermore, third party payers were quite interested in this study since oxygen therapy is very expensive and were likely to make decisions based on the NOTT results. Given that this subgroup had not been defined in advance and that several subgroups had been considered, the decision was to continue to the next full committee meeting in September of 1979, with continued interim reports (see Table I). At the next full committee meeting, the total mortality comparison p-value was 0.09 with the FEV₁ subgroup being 0.008.

While the monitoring committee continued to struggle with this, the statisticians were uncomfortable that all mortality data might not be available and up to date. A few more deaths in either group could change the results substantially. It was agreed that this matter had to be cleared up before any decision could be made. Special efforts were made to ascertain the vital status of each patient, taking care not to alarm the participating investigators. After this process had been completed, neither the overall or subgroup comparisons were near statistical significance. At the January meeting, interest in early termination decreased. As the final results of the trial emerged, mortality turned out to be quite significant (41/102 vs 23/101,
with the results for the two FEV₁ subgroups being similar, each in favor of COT. Had the trial been terminated early on the basis of a subgroup, the study might have suggested that COT was more effective in one subgroup alone.

Table I

Significance Levels for Mortality Comparisons at Various Interim Analyses as Originally Presented and in Retrospect, for the Total Group and for the Lower FEV₁ Subgroup

<table>
<thead>
<tr>
<th>Meeting Date</th>
<th>Original Analyses</th>
<th>Retrospective Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total Group</td>
<td>Subgroup</td>
</tr>
<tr>
<td>Oct, 1978</td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>March, 1979</td>
<td>.15</td>
<td></td>
</tr>
<tr>
<td>June, 1979</td>
<td>.07</td>
<td>.01</td>
</tr>
<tr>
<td>Sept, 1979</td>
<td>.09</td>
<td>.008</td>
</tr>
<tr>
<td>Oct, 1979</td>
<td>.52</td>
<td>.25</td>
</tr>
<tr>
<td>Jan, 1980</td>
<td>.15</td>
<td>.05</td>
</tr>
<tr>
<td>March, 1980</td>
<td>.06</td>
<td>.05</td>
</tr>
<tr>
<td>June, 1980</td>
<td>.01</td>
<td></td>
</tr>
</tbody>
</table>

In retrospect, had the mortality been as up to date as ideally possible, the subgroup issue would not have been as great a concern since the P-value for the low FEV₁ was .06 at the September meeting as shown in Table I. Realistically, mortality data and other outcome variable data will not be totally up to date, but the NOTT group strongly encouraged this lag to be minimized.
Another aspect of this trial is that once the hypothesis involved morbidity and quality of life, a multitude of potential outcome measures were available. It is difficult, if not impossible, to isolate a single variable to fully describe what superficially might seem to be a simple question. Hence the amount of data to evaluate can be overwhelming if some selected group of measures is not identified in advance. Even though that was done in the NOTT, the task was still not easy and as it turned out attention focused on another variable, mortality.

THE DIABETIC RETINOPATHY STUDY EXPERIENCE

The Diabetic Retinopathy Study (DRS) was a randomized controlled clinical trial which entered over 1,700 patient in 15 participating centers. Various aspects of this trial have been presented by Cornfield (1976), Ederer (1981), and the Diabetic Retinopathy Study Research Group (1976, 1978, 1981). The primary objective was to compare the effectiveness of photocoagulation with no therapy to either delay or even prevent visual loss in diabetic patients with proliferative retinopathy. A unique feature of this study is that patient's eyes were randomized, one to experimental therapy and the other to no treatment. A five year follow-up was planned and the primary outcome was the incidence of blindness. Here, as in the other trials described, a data monitoring committee was established to review interim data and part of their experience has been reported by Ederer (1981).

Shortly after recruitment of patients was completed and after two years of the scheduled five years of follow-up, a 60% reduction in the incidence of blindness was observed in the treated eyes in contrast to the untreated eyes (16.3% vs 6.4%, z=5.5). This result was highly statistically significant. The question the data monitoring committee was faced with was whether the early effect could be negated by later harmful effects. A previously published study suggested such a possibility. To stop
early if there might be long term adverse effects would not be desirable since study patients might have the untreated eye treated and other nonstudy patients also might have eyes treated. On the other hand, if treatment were continued and no long term adverse effects were observed, the therapeutic benefits would be delayed for the study patients' untreated eyes as well as deferring treatment for all other patients not in the study. Projections were made as to how adverse the long term effects would have to be in order to reverse the observed early benefit, a curtailed sampling idea. These types of calculations suggested it was unlikely that the current trend could be reversed.

As in previous studies discussed, the issue of subgroups was also a factor. Apparently, in looking at one possible set of three mutually exclusive subgroups, only one suggested a "significant" difference. In some further subdivisions, small numbers of subjects in each gave the resulting data an inconsistent or perhaps a random appearance. It was not clear whether to terminate the entire study or some part of it.

After much discussion and extensive statistical analyses, the following decisions were reached.

a) The protocol was modified to treat all untreated eyes in study patients defined to be at "high risk".
b) Results to date were published for the benefit of all nonstudy patients.
c) All study patients would be followed to see if any long term adverse effects were observed, although the comparison would be contaminated to some degree by this "later" treatment of untreated high risk eyes.

The five year results confirmed the two year results and no adverse effects were manifest. The decision process involved incorporating basic analysis of two treatment groups, subgroup analyses, other published data, and projections of possible outcomes.
BETA-BLOCKER HEART ATTACK TRIAL

Like the Diabetic Retinopathy Study, the Beta-Blocker Heart Attack Trial (BHAT) was terminated earlier than scheduled. BHAT was a randomized double blind trial of 3,837 patients with a recent myocardial infarction, conducted in over 30 centers, sponsored by NIH, and directed by the Beta-Blocker Heart Attack Trial Research Group (1981, 1982). Patients were randomized into a propranolol treated group or a placebo treated group during a two year period of recruitment with an additional two years of follow-up. Thus, the scheduled follow-up would be at least two years and at most four years, with the average being three years. Mortality from all causes was the primary outcome measure. As in the others described, the NIH clinical trial model was used and a data monitoring committee was established. This trial used group sequential boundaries and stochastic curtailment methods. The results of their experience has been presented by DeMets et al. (1984).

Approximately one year after the last patient was randomized, the data monitoring committee of BHAT recommended to the NIH that the trial be terminated and the results be made available. The mortality curves for the two treatment groups are represented in Figure 3. The logrank test statistic for comparison of these two curves is 2.92. This mortality result did not appear suddenly but rather the trend was seen very early in the life of the study as shown in Table II. As discussed earlier, repeatedly testing data increases the probability of Type I error. One way to achieve an overall Type I error is to use conservative critical values at interim analyses. The method proposed by O'Brien and Fleming (1979) and adopted by BHAT uses a large critical value at early interim analyses, relaxing this critical value at the trial progresses, with a more conventional value at the final scheduled analysis. The 5% significance level group sequential boundaries of the O'Brien-Fleming family are also shown in Table II, given
the seven scheduled data monitoring committee meetings. The early survival comparisons did not exceed these boundaries even though the test statistic did exceed nominal values. The critical value at the sixth meeting clearly exceeded this 5% significance level boundary. Moreover, at that point, the results were so strong that using stochastic curtailment ideas, it appeared highly unlikely that the results would be reversed if the trial continued, short of a major catastrophe. That was also considered unlikely since this drug has been used extensively in patients of this type for other indications.

Table II

O'Brien-Fleming Monitoring Boundaries (2α = .05) for the Beta-Blocker Heart Attack Trial, Assuming Seven Scheduled Analyses

<table>
<thead>
<tr>
<th>Meeting</th>
<th>Log Rank Statistic</th>
<th>Upper Boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>May, 1979</td>
<td>1.68</td>
<td>5.40</td>
</tr>
<tr>
<td>Oct, 1979</td>
<td>2.24</td>
<td>3.82</td>
</tr>
<tr>
<td>Mar, 1980</td>
<td>2.37</td>
<td>3.12</td>
</tr>
<tr>
<td>Oct, 1980</td>
<td>2.30</td>
<td>2.70</td>
</tr>
<tr>
<td>Apr, 1981</td>
<td>2.34</td>
<td>2.41</td>
</tr>
<tr>
<td>Oct, 1981</td>
<td>2.82</td>
<td>2.20</td>
</tr>
<tr>
<td>June, 1982</td>
<td>-</td>
<td>2.04</td>
</tr>
</tbody>
</table>

Although the committee felt that this early effect of propranolol on the reduction in mortality to be statistically and clinically significant, the decision to terminate was not easy. The issue focused on how long the drug should be given or how long was the drug effective in preventing fatal events in this patient population. This was felt by the clinicians to be an important
recommendation for the BHAT to make. The dilemma for BHAT was similar to that of the DRS. To continue for another year meant delaying the potential benefit of therapy to study patients as well as others. To terminate early meant not having another year of follow-up to ascertain whether the three year benefit persisted for at least another year, implying continued treatment for at least four years.

The statistical approach taken to help resolve this issue was similar in spirit to the DRS approach, although done independently. Projected four year life tables were constructed based on the existing three years of experience. It was argued that while some information would be obtained, it would not be very precise due to the few numbers of anticipated events, assuming the force of mortality to be quite small for years three and four in contrast to the first year. After considerable discussion, the committee voted for termination on the basis that the observed early benefits outweighed the marginal gain in information from an additional year. As it turned out, two additional large trials reported similar results, one just prior to the BHAT decision and one afterwards, confirming the early benefit.

OTHER TRIALS

While few other trials have published their data monitoring experience in detail, three recent NIH clinical trials at least suggest that studies may be continued even when interim results are quite positive or very negative. The Hypertension Detection and Follow-up Program (1979), also referred to as the HDFP, published positive results suggesting that aggressive treatment of mild hypertension can lead to a reduction in mortality from all causes (P=.01). While the mortality results as they were seen by their data monitoring committee have not been published, one can get some idea of what those results must have been by looking at
the published life table, recognizing the long period of follow-up relative to the recruitment period. The survival curves steadily separate during the five years of follow-up. It is likely that their results were "significant" somewhere in the middle of the follow-up period, yet the trial continued. One might speculate that there was concern over acceptability of the results had the trial stopped early and a need to establish toxicity information on long term usage. For whatever reasons, the committee felt justified continuing the trial, given the interim results.

Two other randomized multicenter trials, the Aspirin Myocardial Infarction Study (AMIS) and the Multiple Risk Factor Intervention Trial (MRFIT) published negative results (1980, 1982). AMIS compared aspirin vs. placebo in patients with a previous myocardial infarction. After several years of follow-up, the mortality results and the survival curves in the two groups were nearly identical. Again, one can only speculate what the rationale to continue might have been. Aspirin is certainly readily available and there already existed public sentiment about the benefits of aspirin from earlier published studies. Since no serious harm was being done other than known side effects of aspirin usage, continuation may have seemed in this case quite reasonable.

One might also speculate that MRFIT continued for quite another reason. This trial involving over 12,000 people was aimed at primary prevention of fatal cardiovascular events in people identified to be at high risk due to either high cholesterol, high blood pressure, cigarette smoking or a combination. Two forms of therapy were compared, a very aggressive special care and a "regular" care approach. After seven years of follow-up, the published results suggested no treatment group differences as far as total mortality or cardiovascular mortality was concerned. Again the interim results during the trial must have looked discouraging. One factor that may well have been argued is that
no one is sure how immediate the impact of such therapy would be, even assuming it to be effective. While early results were not impressive, one might conceive that the therapy might eventually become beneficial if given enough time to reverse years of an individual's life style.

It is, of course, dangerous to speculate on decision processes without benefit of the details for these studies' experience. Even with such details, one might not be able to totally recapture the issues and factors. The main message from these studies is, however, that studies were continued given interim data that strongly suggested the final outcome.

SUMMARY

The CDP experience led the investigators to suggest a number of issues which must be resolved before any decision for early termination can or should be reached. Those issues have been affirmed by others as well. A brief summary of the issues (with some modifications) is described below.

1) Are the results explained by a possible imbalance in the distribution of baseline patient characteristics?
2) Could the patient evaluation be biased as far as primary outcome measures is concerned?
3) Is there a consistency of results for other primary or secondary response variables?
4) Is there a consistency of results across various subgroups of the patients and across clinical centers?
5) What is the risk and benefit ratio of the therapy?
6) Could the results be explained by patient compliance to prescribed therapy or some other therapy?
7) Have the results been interpreted in view of the effect of repeated testing of accumulating data?
8) Could the current trends likely be reversed if the trial continued to the end?
9) How much additional information would be obtained by continuing?

10) What would the impact of early termination have on the acceptability of the results to medical and scientific colleagues?

11) Are the results consistent with other studies or available information?

12) Was it ethical to continue further treatment and follow-up given the early evidence of therapeutic benefit.

All of this discussion has attempted to demonstrate that early termination of any clinical trial is a complex, difficult if not at times tortuous, and not an easily defined process. Even for studies which implemented some of the recent statistical methodology, the decision process was not straightforwardly dictated by that statistical methodology. These methods clearly were not stopping rules as is often the manner in which they are described.

Nevertheless, such statistical methods are essential. Even though many of the methods assume termination to occur if a boundary were crossed, the fact that this assumption might be violated in practice does not lessen their value. Just as navigators need positional sites to help them judge their location, data monitoring committees need statistical methods to help them judge their position with regard to cautious interpretation of interim results from accumulating data. While they are not stopping rules, such methods can be useful guides in the decision making process. As the methods become further developed and can more closely reflect the clinical trial situation, the more precise guides the methods will be, but ultimately these methods will still only be practical guides and not absolute rules.
ACKNOWLEDGEMENTS

This work was supported in part by NIH Grants No. NCI P30-CA-14520-11 and 2-R01-CA18332. The author would like to thank Ms. Terry Metcalf for typing this manuscript.
BIBLIOGRAPHY


FIG. 1  Life table cumulative mortality rates, Coronary Drug Research Project Group. (Reprinted by permission of the publisher from "Practical aspects of decision making in clinical trials" by the Coronary Drug Project Research Group, Controlled Clinical Trials, 1, 323. Copyright 1981 by Elsevier Science Publishing Co., Inc.)
FIG. 2 Z-values for clofibrate-placebo differences in proportion of deaths by calendar month since beginning of study (Month 0 = March 1966, Month 100 = July 1974). (Reprinted by permission of the publisher from "Practical aspects of decision making in clinical trials" by the Coronary Drug Project Research Group, Controlled Clinical Trials, 1, 372. Copyright 1981 by Elsevier Science Publishing Co., Inc.)
FIG. 3 Cumulative mortality curves for 30 months of follow-up, comparing propranolol and placebo treated patients. (Reprinted by permission of the publisher from the "Beta-Blocker Heart Attack Trial" by the Beta-Blocker Heart Attack Study Group. J. Amer. Med. Assoc., 246(18):2073, Nov. 6, 1981.)
FIG. 1  Life table cumulative mortality rates, Coronary Drug Research Project Group. (Reprinted by permission of the publisher from "Practical aspects of decision making in clinical trials" by the Coronary Drug Project Research Group, Controlled Clinical Trials, 1, 323. Copyright 1981 by Elsevier Science Publishing Co., Inc.)
FIG. 2  z values for clofibrate-placebo differences in proportion of deaths by calendar month since beginning of study (Month 0 = March 1966, Month 100 = July 1974). (Reprinted by permission of the publisher from "Practical aspects of decision making in clinical trials" by the Coronary Drug Project Research Group, Controlled Clinical Trials, 1, 372. Copyright 1981 by Elsevier Science Publishing Co., Inc.)
Coronary Drug Project

![Graph showing Z value over months of study]
FIG. 3 Cumulative mortality curves for 30 months of follow-up, comparing propranolol and placebo treated patients. (Reprinted by permission of the publisher from the "Beta-Blocker Heart Attack Trial" by the Beta-Blocker Heart Attack Study Group. J. Amer. Med. Assoc., 246(18):2073, Nov. 6, 1981.)
Key words: clinical trial; early stopping; data monitoring; decision rules