1971

GROUP SEQUENTIAL METHODS FOR CLINICAL TRIALS APPLIED TO THE LOGISTIC REGRESSION MODEL

Joy Duci Mele

School of Basic Sciences, Medical College of Virginia

Viriginia Commonwealth University

This is to certify that the thesis prepared by Joy Duci Mele entitled "Group Sequential Methods for Clinical Trials Applied to the Logistic Regression Model" has been approved by her committee as satisfactory completion of the thesis requirement for the degree of Master of Science.

[redacted]

Director of Thesis

[redacted]

Committee Member

[redacted]

Committee Member

[redacted]

Committee Member

[redacted]

Committee Member

Committee Member

[redacted]

Department Chairman

[redacted]

Dean, School of Basic Sciences

August 9, 1984

Date

GROUP SEQUENTIAL METHODS FOR CLINICAL TRIALS

APPLIED TO THE LOGISTIC REGRESSION MODEL

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

By

Joy Duci Mele

B.A., Emmanuel College, 1971

Director: Dr. Vernon Chinchilli
Assistant Professor
Department of Biostatistics

Virginia Commonwealth University
Richmond, Virginia
August, 1984

ACKNOWLEDGEMENTS

I would like to thank, above all, my husband, Paul whose unselfishness and love have given me the moral support and time needed to fulfill the réquiremants for this degree. Also, a special thanks to my son Tristan who could make me laugh and play even during the most difficult times of my studies.

I would like to thank my committee members for their time and suggestions. I would especially like to thank my advisor, Dr. Chinchilli for his help in all phases of the preparation of my thesis and Dr. Breen for his help with the simulation study. Thanks also to Bruce Dornseif and my husband , Paul, for carefully editting the first two chapters.

Table of Contents

CHAPTER I.

INTRODUCTION

1.1 Thesis objectives.

The objectives of this thesis are

1. to provide a comprehensive guide to using Pocock's group sequential method for clinical trials and

2. to show by computer simulations that the group sequential method is appropriate when using the logistic regression model.

In section 1.2, clinical trials are defined with an emphasis on Phase III clinical trials. The primary intent of sections 1.2 and 1.3 is to describe clinical trial characteristics which suggest that interim analyses are desirable. In section 1.4, it is shown that interim analyses should not consist of performing the usual significance tests. Chapter 1, then, describes the basis for the development of the group sequential method.

Chapter 2 describes the group sequential method in detail (objective #1) and chapter 3 includes the results of the Monte Carlo study (objective #2).

1.2 Clinical trials.

In order for a new treatment to gain acceptance by clinicians, its safety and efficacy must be proven. Clinical trials serve this purpose. The new treatment is systematically evaluated in four phases. The initial investigation, a Phase I study, provides data collected by observing the effect of the treatment on healthy volunteers or on patients who are not candidates for conventional therapy. The objective of a Phase I study is generally to assess the safety and tolerability of the treatment. For example, in studying a new drug, a Phase I study may be used to investigate acceptable dose levels. Additionally, pharmacological studies to determine absorption and bioavailability of the drug should be implemented (Lesser, 1983). The sample size of a Phase I study is small in order to reduce human exposure to a new treatment.

In Phase II studies, subjects are patients with symptoms of the disease for which the treatment is intended. The purpose of these studies is to collect further data on treatment efficacy and safety in diseased subjects (Neiss and Boyd, 1984). Sufficient data should be collected to determine if the treatment has a therapeutic effect which warrants further study.

Once the safety of the treatment has been determined, a Phase III study is initiated to compare the effectiveness of the new treatment to a control treatment (eg. placebo or

well-accepted treatment). Treatment comparisons may be of the following types:

1. a new drug or therapy vs. a standard drug or therapy;
2. a surgical treatment vs. a medical treatment;
3. two or more accepted treatments;
4. a drug or therapy vs. placebo or no treatment and
5. different forms of the same treatment.

If the new treatment is found to be more beneficial than the control treatment, a Phase IV study would be conducted to measure the effects of chronic usage and to determine new uses for the treatment (Smith, 1976).

The largest body of the scientific literature on clinical trials focuses on Phase III studies; we will do likewise in this paper. A Phase III study can be defined as a randomized, controlled clinical trial. It consists of four essential steps; patient selection, randomization to treatment, treatment period, and statistical analysis (Juhl, 1982). Trial entry criteria are specified and only patients satisfying these criteria are selected. If the criteria are comprehensive, the treatment differences may be small due to heterogeneity in the study population but results may be more widely applicable. If criteria are restrictive, the contrary may be true; that is, a statistical difference is more likely to be found due to homogeneity in the study population, but the results may not be broadly generalized to the population

at large. Therefore, when considering the patient population to be studied, one should consider the magnitude of the treatment differences expected as well as limitations on generalizations of results. After a patient is selected for a clinical trial, randomization is performed to remove systematic bias in the allocation of patients to treatment groups and, based on sampling theory, to validate statistical tests and estimation procedures. Pocock (1979) describes randomization schemes appropriate for clinical trials. It is essential that one consider the experimental design of the trial and the prognostic variables to be controlled when choosing a randomization scheme. Once randomized to treatment, the treatment period begins during which the responses of interest are observed. Armitage (1960) classifies response variables into four types; those reported by the patient, those observed by clinical examiners or by use of some technical apparatus, those involving a change in medical care and life or death. Response data is collected and a statistical analysis is conducted to determine if the treatments differ.

One feature that distinguishes clinical trials from other experimental designs is that subjects do not ordinarily enter the trial all at once. Typically, patients enter sequentially and, consequently, the length of patient entry may be considerable, particularly if the incidence of the disease being studied is low or if the trial is limited to a

single medical center. Responses are then observed serially and investigators may be tempted to do an interim analysis on the data as it accumulates.

1.3 Reasons for doing interim analyses

Pocock (1981) has suggested five possible reasons for doing interim analyses on accumulating data:

1. to insure that the trial protocol is being followed;
2. to monitor for adverse effects to treatment;
3. to check for problems in data collection;
4. to maintain enthusiasm for the trial and
5. to look for a significant difference between treatments which may warrant stopping the trial.

Proper administration of the trial should solve problems relating to reasons #1 and #3. One would hope that Phase I and Phase II studies would identify the toxic effects of the treatment and therefore eliminate reason #2. However, since the patient entry criteria and sample size in a Phase III study may differ markedly from Phase I and Phase II studies of the treatment, additional toxic effects may be observed. For example, the patient population in a Phase II study may be more heterogenous, and therefore, toxic effects may show up in subgroups of the population which were not represented in the earlier studies. Reason #2 then may be sufficient reason for doing an interim analysis with the intent of discontinuing treatment or changing the treatment protocol to

remove the toxic effects. Reason #4 would be particularly relevant for long-term trials or multi-center trials where one might be tempted to satisfy the curiousity of investigators by providing information on the progress of the trial. Supplying data on treatment comparisons, however, would most certainly bias patient selection which in turn lends uncertainty to the validity of the statistical analysis. Additionally, knowledge of unexpected results may diminish enthusiasm for the trial. However, information unrelated to response variables, such as the number of patients entered in specified strata, may be provided which may placate trial participants without sacrificing blindness.

The fifth reason is valid from an ethical as well as an economical viewpoint, the former being of greater importance in a clinical trial. Ethically, it is desirable to stop the use of a less beneficial treatment in favor of a superior treatment so as to decrease the number of patients exposed to the inferior treatment. Moreover, stopping a trial early decreases the overall length of the study and the sample size; these two factors should be economically attractive to supporters of clinical trials.

The reason for doing an interim analysis should be clear to the trial organizers so that a stopping rule may be formulated. A stopping rule is some criteria which when satisfied dictates stopping the clinical trial. For example, achievement of a statistically significant difference between

treatments appears to be the correct criterion for stopping a trial. Yet, in practice, due to the complexity of clinical trials, other factors besides statistical significance must be considered. Secondary response variables may be an important factor, and therefore, it may be decided to continue the trial in order to collect adequate data for measuring the effects of these additional variables. Hence, the decision to stop would not be made on the basis of one response variable.

Another factor to consider would be the effectiveness of the treatments on subgroups of the patients in the study. An interesting example of a clinical trial stopped early due to treatment differences in subgroups is a study done by the Coronary Drug Project Group (1972) to measure the effects of dextrothyroxine on the mortality rate of coronary patients. The dextrothyroxine-placebo difference in overall mortality was not significant, but dextrothyroxine generally manifested significantly higher mortality rates for subgroups of the data. In remaining subgroups, favorable (but not statistically significant) results were observed early in the study but the trend was towards less favorable results as the study progressed. The sample size for these subgroups was considered to be too small to warrent continuing the trial. Therefore, the study was discontinued. Additional factors may be considered in formulating a stopping rule depending on the nature of the clinical trial. Whatever criteria are used,

statistical results are an integral part of the stopping rule and should serve as an objective guide for making decisions (Pocock, 1981).

The complexity of formulating a stopping rule should not deter clinical trial organizers from outlining, a priori, the specific criteria that warrant stopping the trial. However, as Pocock (1978) and McPherson (1974) point out, interim analyses are commonly done with no well defined stopping rules. A survey of 40 clinical trials (Pocock, 1978) revealed that 33 assessed interim results. All of these trials had been in progress at least two years at the time of the survey. In this two year period, 23 trial investigators had looked at their data at least 4 times, implying that interim analyses are done frequently. Yet 22 of the 33 investigators had devised no formal or informal stopping rules. This may suggest that the majority of investigators were not looking at their data with the intent of stopping the trial even if one treatment demonstrated superiority. Their reason for doing an interim analysis, therefore, is of questionable validity. Six trial investigators used repeated significance testing as their stopping rule. As stated above, a statistically significant treatment difference should not constitute a stopping rule but is important in formulating the stopping rule. Furthermore, there are problems with using the results of repeated significance tests. This will be discussed in the following section.

1.4 The problem with repeated significance testing.

Repeated significance testing is defined as the periodic performance of a statistical analysis on accumulating data to determine if a significant difference between treatments exists. Intuitively one might think that as the number of significance tests performed increases, one would expect to find a significant difference purely by chance. In fact, if one performs enough significance tests, any level of nominal significance can be achieved eventually.

This phenomenon is referred to as "sampling to a foregone conclusion" (Cornfield, 1966). Mathematically it can be shown to be a consequence of the law of the iterated logarithm (Anscombe, 1954).

For example, given N independent and normally distributed random variables, $X_1, X_2, \ldots, X_N$ with unit variance and mean μ, the critical region for testing $H_o: \mu=0$ vs. $H_a: \mu\neq 0$ is

$$\left|\sum_{i=1}^{N} X_i\right| \geq \sqrt{N}\, z_{1-\alpha/2},$$

where $z_{1-\alpha/2}$ is the upper $1-\alpha/2$ critical point for the standard normal distribution. The level of significance ( ) for this fixed sample size test is equal to

$$P\left[\left|\sum_{i=1}^{N} X_i\right| \geq \sqrt{N}\, z_{1-\alpha/2} \mid \mu=0\right].$$

If this test is done repeatedly, the level of significance increases to a limit of one by the law of the

iterated logarithm which states for any K (Rao, 1973) that

$$P[|\sum_{i=1}^{n} X_i| \geq K\sqrt{n} \text{ infinitely often}] = 1$$

Since K can be set to any $z_{1-\alpha/2}$, any level of significance may be reached with probability one if testing is done extensively.

Armitage, McPherson and Rowe (1969) have computed the exact probabilities of finding a significant difference when no difference exists using a method of numerical integration. Given that $X_1, X_2, \ldots, X_N$ are random variables independently distributed $N(\mu,1)$, the test statistic for testing $H_o: \mu=0$ vs. $H_a: \mu\neq 0$ is $S_N = |\sum_{i=1}^{N} X_i|$. The density function of S_N is defined recursively by

$$f_N(s_N) = \begin{cases} \int_{-y_{N-1}}^{y_{N-1}} f_{N-1}(u)\,(2\pi)^{-1}\exp[-(s_N-u)^2/2]\,du, \\ \qquad -y_N \leq s_N \leq y_N \\ 0, \text{ otherwise} \end{cases}$$

where $y_N = z_{1-\alpha/2}\sqrt{N}$ with alpha chosen as for a fixed sample size test. The probability of finding a significant difference when none exists at or before N observations is P_N where

$$P_N = 1 - \int_{-y_N}^{y_{N-1}} f_N(u)\,du.$$

The results of these computations are shown in Table 1, which is an abridged version of Table 2 from Armitage et al. (1969). The probabilites in Table 1 clearly support the previously stated results of the law of the iterated

Table 1. Probability of detecting a significant difference using repeated tests (N) at an α significance level when the null hypothesis is true and when sampling from a normal distribution with known variance.

α	.05	.02	.01
z	1.960	2.336	2.576
N			
1	0.050	0.020	0.010
2	0.083	0.035	0.018
3	0.107	0.046	0.024
4	0.126	0.055	0.029
5	0.142	0.062	0.033
10	0.193	0.088	0.047
25	0.266	0.126	0.070
50	0.320	0.156	0.088
100	0.374	0.187	0.107
500	0.487	0.259	0.152
1000	0.530	0.288	0.172

logarithm, which is that repeated significance testing inflates the alpha level and increases the chance of making a type I error. For example, if a statistical test is repeated four times (which is not uncommon in clinical trials) at a level of significance of .05, the chance of making a type I error would be about 2.5 times that level (0.126). It is interesting to note that 10 tests repeated at a nominal significance level of .01 maintains a true significance level of approximately .05.

Probabilities for the binomial case were lower than the normal case for equivalent N, which Armitage et al. (1969) explain as being due to the conservativeness of the binomial test. For the exponential case, probabilities were very close to those for the normal case differing only in the third decimal place.

In a subsequent paper, McPherson and Armitage (1971) looked at the effect of repeated testing when the null hypothesis is not true. The sampling distribution is as described for the null case. The probabilities in Table 2 refer to the probability of crossing an upper boundary which is equivalent to stopping when $S_N \geq y_N$. The results show that repeated testing increases power, which is what we might expect due to the relationship between the probability of a type I error (α) and the probability of a type II error (β). That is, as α increases, β decreases and therefore the power

Table 2. Probability of detecting a significant difference using repeated tests (N) at an α significance level when the null hypothesis is not true and when sampling from a normal distribution with unit variance.

(upper boundary only)

		N		
α	μ	10	25	50
.05	0	0.097	0.132	0.160
	0.5	0.497	0.815	0.962
	1.0	0.925	0.998	0.998
	1.5	0.998	1.000	1.000
.02	0	0.044	0.063	0.078
	0.5	0.341	0.695	0.933
	1.0	0.854	0.998	1.000
	1.5	0.995	1.000	1.000
.01	0	0.024	0.035	0.044
	0.5	0.247	0.594	0.893
	1.0	0.784	0.995	1.000
	1.5	0.990	1.000	1.000

$(1-\beta)$ increases. So, even though the power of detecting a difference is increased by repeated testing, it is done at the expense of increasing the probability of incurring a type I error.

Decreasing the risk of making a type I error is especially important in clinical trials. For example, when comparing a new treatment to a standard treatment, falsely rejecting H_o would lead to the acceptance of an inferior treatment by the medical community and to the rejection of a treatment once considered acceptable. However, falsely accepting H_o would result in no change in acceptable medical practice. Obviously, the former error is less desirable. A statistical method for maintaining alpha at an acceptable level is needed.

Table 1 and Table 2 clearly demonstrate that p-values resulting from repeated classical significance test cannot be interpreted in the usual manner. Instead, a stopping rule must be formulated which considers the number of interim analyses to be performed and the desired overall significance level. The design of the clinical trial, therefore, will be determined by the stopping rule (Armitage, 1960).

CHAPTER II.

GROUP SEQUENTIAL METHOD

2.1 Introduction.

Whereas interim analyses are justified, particularly for ethical reasons, a statistical method which allows repeated testing while maintaining an acceptable overall significance level is desirable.

In the 1940's, Wald (1947) developed the sequential test whose characteristic feature is that the decision to continue testing is based on the outcome of observations as they are made. That is, after each observation one of the following three decisions is made:

1. accept H_o;
2. reject H_o or
3. continue the trial.

The sample size, therefore, is not predetermined and, on the average, is substantially smaller than the sample size for a fixed sample size design.

Armitage (1960) further developed the work of Wald to apply sequential methods to clinical trials. This methodology

requires that patients enter sequentially, in pairs, and are randomly allocated to treatment. The response time should be short relative to the expected length of patient accrual. Consequently, as the response time increases, the maximum number of pairs available for analysis decreases while the number of pairs in the trial increases. Therefore, the objective of decreasing the number of patients exposed to an inferior treatment would not be served when the time to response is prolonged.

Analysis is done after every pair of observations, thereby requiring a constant vigil over the accumulating data. This last requirement is necessary so that immediate action may be taken when a sequential boundary is crossed, where crossing a boundary is equivalent to rejecting or accepting the null hypothesis.

There are some obvious problems associated with applying the classical (also referred to as fully or continuous) sequential method to clinical trials. In practice, it is difficult to continually keep data in the proper form for analysis. This is particularly difficult in a multicenter trial where administration of the trial is already complicated by the need for extensive data monitoring for quality assurance (Bendush et al., 1984). An additional problem is that patients usually do not enter in pairs and, if they did, it is likely that they would not be well-matched on important prognostic variables (Friedman et al., 1981).

This problem has been addressed by the development of sequential tests which do not require that patients enter in pairs (Hoel et al, 1976). Also, patient responses are often prolonged. Armitage (1960) states that a sequential procedure should not be considered if the time to response is more than half the maximum expected length of patient entry. A classical sequential design, then, would be inappropriate in studies of treatments for chronic diseases (e.g. cancer). Since the sample size is dependent on the outcomes of the observations, even the most generous trial sponsor may feel uneasy about the indeterminant length of the trial. Another major disadvantage of the classical sequential method is that it is not well-received by clinicians whose depth of statistical knowledge does not extend beyond the use and interpretation of the usual significance tests (Pocock, 1981).

An approach designed to address these problems associated with the classical sequential method is called the group sequential method. Instead of analyzing the data after each pair of patients, significance tests are performed on data accumulating from groups of patients (2n, n patients on each treatment). The maximum number of tests (N) to be done is predetermined, therefore the maximum length of the trial is known. An overall significance level (α) is maintained by using a more stringent nominal significance level (α') for each of the repeated significance tests.

2.2 Group sequential design.

2.2.1 General description.

For a group sequential design, patients are expected to enter sequentially and be randomly allocated, using a permuted block design, to one of two treatments (A or B), such that after 2n patients have entered the trial, there are n patients on each treatment (Pocock, 1977). The response variable is assumed to be normally distributed with mean, μ_A and μ_B, and common variance, σ^2. For testing the two-sided hypothesis; $H_o: \mu_A - \mu_B = 0$ vs. $H_a: \mu_A - \mu_B = \delta$, the test statistic at the i th test is (Pocock,1977)

$$\bar{D}_i = \sum_{j=1}^{i} (\bar{X}_{Aj} - \bar{X}_{Bj})/\, i,\ i = 1,\ldots,N$$

$\bar{D}_i$ is normally distributed with mean μ_D and variance σ_D^2 where

$$\begin{aligned} u_D = E[\bar{D}_i] &= 1/i\ E[\sum_{j=1}^{i} (\bar{X}_{Aj} - \bar{X}_{Bj})] \\ &= 1/i\ (i*\mu_A - i*\mu_B) \\ &= \mu_A - \mu_B \end{aligned}$$

and

$$\begin{aligned} \sigma_D^2 = \mathrm{Var}[\bar{D}_i] &= \mathrm{Var}\,[1/i \sum_{2j=1}^{i} (\bar{X}_{Aj} - \bar{X}_{2Bj})] \\ &= 1/i\ [i(\sigma^2/n) + i\ (\sigma^2/n)] \\ &= 2\,\sigma^2/in. \end{aligned}$$

The critical region of the likelihood ratio test may be expressed as

$$\bar{D}_i / [2\sigma^2/(in)] \geq z_c,\ i = 1,\ldots,N.$$

Letting $Z_i = \bar{D}_i / [2\sigma^2/(in)]$, the test may be expressed as

$$| Z_i | \geq z_c , \quad i = 1,\ldots,N,$$

where z_c is the critical value corresponding to the appropriate significance level (α'). When the above relationship is true, reject H_o, claiming a treatment difference, and consider stopping the clinical trial. When the relationship is false, do not reject H_o and continue the trial. If $Z_N < z_c$ at the Nth test, the trial terminates and a claim of no treatment difference is made.

2.2.2 Level of significance.

From section 1.4, it follows that a nominal level of significance (α') must be computed so that, at the N^{th} test, the predetermined overall level of significance (α) is maintained. To do so, the maximum number of tests (N) and the desired overall level of significance must be specified. By inverse interpolation, z_c and α' may be obtained from Table 1. For example, for N = 10, the entries of interest (to 5 decimal places) from Table 1 are

α	.05	.02	.01
z_c	1.9600	2.3263	2.5758
	0.19336	0.08776	0.04738

To maintain an α level of .05 , we must test at an α' level between .01 and .02.

To find α', do the following interpolation:

$$\frac{.08776 - .04738}{2.3263 - 2.5758} = \frac{.08776 - .05}{2.3263 - z_c},$$

which yields that $z_c = 2.5596$ which is the standard normal deviate corresponding to $\alpha' = .0106$. Results are given in Table 3 (Pocock, 1977, Table 1) for N in the range 2 to 20 and $\alpha = .05$ or .01. Note that as the number of interim analyses increases, the nominal significance level (α') decreases.

For example, suppose an investigator decides to perform a maximum of 5 statistical tests with an overall α level of .05. The appropriate z_c is 2.413 corresponding to α' of .0158. Then, H_o is rejected at the i^{th} test, for i = 1,...,5, if $Z_i > 2.413$. Each test is repeated at the same nominal significance level.

The nominal significance level is dependent on the maximum number of tests (N) and the overall significance level, not on the number of patients in each group (2n). The latter follows from section 1.4, where it was demonstrated that the probability of finding a significant difference when none exists (P_N) depends on the number of observations (N). Likewise, for the group sequential method, the overall level of significance achieved depends on the number of tests (N) performed at an α' level. To control α, therefore, the choice of α' depends on N.

Table 3. Nominal significance level (α') and corresponding normal deviate z_c required when performing repeated two-sided significance tests (N).

	Overall Significance Level α = 0.05		α = 0.01	
N	α'	z_c	α'	z_c
2	0.0294	2.178	0.0056	2.772
3	0.0221	2.289	0.0041	2.873
4	0.0182	2.361	0.0033	2.939
5	0.0158	2.413	0.0028	2.986
6	0.0142	2.453	0.0025	3.023
7	0.0130	2.485	0.0023	3.053
8	0.0120	2.512	0.0021	3.078
9	0.0112	2.535	0.0019	3.099
10	0.0106	2.555	0.0018	3.117
11	0.0101	2.572	0.0017	3.133
12	0.0097	2.585	0.0016	3.147
15	0.0086	2.626	0.0015	3.182
20	0.0075	2.672	0.0013	3.224

When repeatedly testing at a constant α' level, it is possible that the results of the N^{th} test show no significant difference at the α' level but the final p-value is less than α. Some statisticians (O'Brien and Fleming, 1979, Canner, 1976 and Peto,1981) have suggested varying the significance level so that more stringent levels are used for the initial tests while the significance level for the Nth test would be approximately equal to α. For example, O'Brien and Fleming (1979) found the following nominal significance levels appropriate for N-group design with $\alpha = .05$ (α' approximated from O'Brien and Fleming, 1979, Table 1);

N	5	4	3
α'_1	.0001	.0004	.0009
α'_2	.001	.004	.017
α'_3	.008	.020	.048
α'_4	.023	.043	
α'_5	.041		

Pocock (1981) has shown that, when H_a is true and power is greater than .80, these designs require a larger sample size than a design with constant nominal significance level. For example, the 5-group varying significance level design shown above, with $1-\beta=.90$, would require approximately 10% more patients than the 5-group constant significance level design. However, if all five tests are performed (i.e. $Z_i < z_c$, for i = 1,...,4), the sample size for the former design will be 15% less than the sample size for the latter

design. Compared to a fixed sample size design, the sample size for the varying significance level design is 3% larger. In fact, O'Brien and Fleming (1979) have shown that their method is "nearly identical" to a fixed sample size method, unless the observed treatment difference is very large, in which case early termination is possible. So for small treatment differences, there appears to be little advantage to these designs over a fixed sample size design. Nevertheless, if the treatment difference is small, O'Brien and Fleming's method may show significant results at the Nth test while Pocock's method may not. An example given by O'Brien and Fleming (1979) illustrates this point. However, it should be noted, that the example is somewhat inadequate for comparing the two methods. Pocock's method requires group sizes larger than 10 for a binary response and approximately equal numbers of patients on each treatment arm to maintain prespecified α and β levels (Pocock, 1977). In the example, though, the group size for one treatment arm was 7 and the group size for the other treatment arm was 14. If the treatment arms were of equivalent size (i.e. both equal to 14), Pocock's method would also terminate showing significant results. So, O'Brien and Fleming's method may be useful for finding important differences at the N^{th} test when sample size is small and power inadequate for use of Pocock's method (Pocock, 1982).

Furthermore, Pocock (1982) has shown, using a

numerical integration minimization procedure, that an optimum design (i.e. a design with monotonically increasing significance levels and minimum sample size) is not considerably different from a constant significance level design with reasonable power ($1-\beta$ = .9 or .95). For example, a 5-group varying significance level design, with α=.05 and $1-\beta$=.95, α'_5= .0165 while for the constant level design α' = .0158. Subsequently, the sample size for the two designs are nearly equivalent. So if power is adequate for detecting an important difference, there appears to be no advantage to varying the significance levels.

2.2.3 Power and sample size.

During the design phase of a clinical trial, it is essential to do power calculations to ascertain the required sample size for detecting the smallest important clinical difference (Weiss et al., 1983, Armitage,1979, Freiman et al. 1978). Sample size requirements from power calculations should be compared to the expected rate of patient accrual and the length of the trial to determine the feasibility of the trial. If an investigator finds that the sample size needed for detecting a prespecified treatment difference cannot be achieved, the trial should be either abandoned or expanded into a multicenter trial.

A survey of the published reports of 71 "negative" randomized controlled clinical trials (i.e. trials where the

null hypothesis of no treatment difference was not rejected) showed that the majority of the trials lacked sufficient power for detecting an important clinical difference (Freiman et al., 1978). An important clinical difference (δ) was defined as a 25% or 50% decrease in response rate from the control response rate (binomial response). For δ = 25%, 80% of the trials had power less than .50 and only 7% had power greater than or equal to .80. For δ = 50%, the results were somewhat better: 44% of the trials had power less than .50 and 31% had power greater than or equal to .80. It is impossible to determine if power was inadequate because sample size requirements were not fulfilled since, as Freiman et al. (1978) point out, few details of prior planning are provided in published reports. Interestingly, only 1 trial report stated the α and β levels that were considered before the start of the trial and 14 mentioned that a larger sample size was needed.

In a survey (Pocock, 1978), already referred to in section 1.3, 34 of the 40 trial investigators revealed that the required sample size was specified before beginning the trial. Only 18 investigators used power calculations, 8 of which based their calculations on a difference of 100% (not the <u>smallest</u> important difference). However, the patient accrual rates were not adequate for achieving the required sample sizes resulting in excessively protracted trials or inadequate power. Pocock (1978) stated that "until all trial

organizers obtain a truly realistic assessment of the potential patient accrual ... the problem of poor accrual will continue to ruin a large proportion of clinical trials." Therefore, the importance of considering the results of power calculations in light of the practical aspects of the trial (i.e. the expected number of patients and length of the trial) should not be underestimated by trial investigators.

To do power calculations, it is necessary to specify an overall level of significance (α), the smallest clinically important treatment difference (δ), and the probability ($1-\beta$) of detecting δ. For a fixed sample size test (N=1), the sample size (2n) is computed as follows assuming σ^2 is known:

$$2n = 2 * (z_{\alpha} + z_{\beta}) * 2\sigma^2 / \delta^2$$

for $H_a : \mu_A - \mu_B = \delta$. Similarly, for a group sequential design, the size of the groups (2n) is dependent on the number of interim analyses (N), α, β and $2\sigma^2 / \quad^2$. From section 1.4, we saw that repeated testing increased power as α increased. However, if the nominal significance level is made more stringent with more testing, the power decreases (McPherson, 1982). It is necessary to find 2n such that the power of the trial is maintained. Pocock (1977 and 1981) uses numerical integration methods to find the required values of 2n. These values are given in Table 4 for α = .05 or .01, β = .75, .90 or .95 and N from 1 to 10.

Table 4. Average number of two-sided tests ($\bar{M}$) until trial termination and number of patients per group (2n) required under Ha: $\mu_A - \mu_B = \delta$, using various group sequential designs for responses normally distributed with known variance σ^2.

$\alpha = 0.05$

	$1-\beta = 0.75$		$1-\beta = 0.90$		$1-\beta = 0.95$	
N	$\bar{M}$	2n *	$\bar{M}$	2n *	$\bar{M}$	2n *
1	1	27.75	1	42.04	1	51.98
2	1.58	15.48	1.41	23.12	1.31	28.39
3	2.19	10.85	1.88	16.11	1.71	19.73
4	2.80	8.40	2.36	12.43	2.12	15.19
5	3.41	6.87	2.84	10.14	2.53	12.38
6	4.02	5.83	3.32	8.57	2.94	10.46
7	4.63	5.06	3.80	7.44	3.35	9.07
8	5.24	4.48	4.28	6.57	3.77	8.01
9	5.85	4.02	4.76	5.90	4.18	7.17
10	6.46	3.65	5.24	5.35	4.59	6.50

$\alpha = 0.01$

	$1-\beta = 0.75$		$1-\beta = 0.90$		$1-\beta = 0.95$	
N	$\bar{M}$	2n *	$\bar{M}$	2n *	$\bar{M}$	2n *
1	1	42.25	1	59.54	1	71.27
2	1.64	23.14	1.47	32.24	1.37	38.42
3	2.30	16.10	2.00	22.32	1.83	26.52
4	2.96	12.39	2.53	17.14	2.29	20.34
5	3.62	10.09	3.06	13.93	2.75	16.52
6	4.27	8.53	3.59	11.75	3.22	13.93
7	4.93	7.39	4.12	10.18	3.68	12.04
8	5.58	6.52	4.65	8.98	4.14	10.63
9	6.23	5.85	5.18	8.03	4.61	9.50
10	6.89	5.29	5.71	7.27	5.07	8.60

* Multipy each entry by σ^2 / δ^2.

From Table 4, the maximum sample size may be determined by 2n * N; this is the number of patients needed if one fails to reject H_o by the N-1 test. For example, for N = 5, α = .05 and $1-\beta$ = .90, the additional number of patient responses required prior to each test would be approximately $10\ \sigma^2/\delta^2$. So at the first test, the sample size should be approximately $10\ \sigma^2/\delta^2$ and at the second test, the sample size should be approximately $20\ \sigma^2/\delta^2$. Until trial termination, continue accumulating data so at the fifth test the maximum sample size of $50\ \sigma^2/\delta^2$ is reached.

An operating characteristic curve is useful for measuring the feasibility of a trial when an investigator is uncertain about the expected treatment difference but is fairly certain of the expected maximum sample size. In fact, Freiman et al. (1978) recommend constructing operating characteristic curves before any trial is undertaken.

Table 5 (McPherson, 1982, Table I) gives operating characteristics for a fixed sample size design and for 5 group sequential designs. The table may be generalized for sample sizes larger than 80 by computing K where K = 2nN/80. Each treatment difference (scaled by $\sqrt{2}\sigma$) should be divided by $\sqrt{K}$ so that the powers in the table still hold. For a given sample size, power may be determined for a range of treatment differences and for a range of N. As expected, power increases with decreasing N and with larger treatment differences.

Table 5. Operating characteristics of 6 group sequential designs with overall level of significance of 0.05 and maximum sample size of 80.

N	1	2	4	5	8	10
2n	80	40	20	16	10	8
α'	0.050	0.030	0.018	0.016	0.012	0.011
$\delta/(\sqrt{2}\sigma)$			$1-\beta$			
0.1	0.09	0.08	0.08	0.07	0.07	0.07
0.2	0.24	0.22	0.19	0.19	0.18	0.17
0.3	0.48	0.43	0.39	0.38	0.36	0.35
0.4	0.72	0.67	0.62	0.61	0.60	0.58
0.5	0.89	0.86	0.82	0.81	0.80	0.79
0.6	0.97	0.95	0.94	0.93	0.93	0.92
0.8	1.00	1.00	1.00	1.00	1.00	1.00
1.0	1.00	1.00	1.00	1.00	1.00	1.00

Since the most important feature of a group sequential design is the extent to which it enables a trial to stop early (Pocock, 1981), the average sample size is of greater interest than the maximum sample size. The average sample size is equal to the product of the average number of tests required when H_a is true and the group size (2n). Recall that, under H_a, the sample size is dependent on the observations observed. Therefore, the number of tests performed (M) is a random variable. Given the distribution of M, the E[M] may be computed. The E[M] is the expected number of tests required under H_a ($\bar{M}$). In other words, the $\bar{M}$ is the number of significance tests the investigator will probably need to do in order to find a significant difference between treatments. Values for $\bar{M}$ are given in Table 4 (Pocock, 1978). The average number of patients required for a trial is equal to $2n\bar{M}$. For example, for a 5-group design with α=.01 and $1-\beta$ =.95, the group size should be $16.52*\sigma^2/\delta^2$ and $\bar{M}$ would be 2.75. So, on the average, a significant result may be expected at the third test.

As N increases, $2n\bar{M}$ decreases for any given power and level of significance. Therefore, it seems that frequent testing results in exposing fewer patients to an inferior treatment. This point will be explained further in the next section.

2.2.4 Choice of N.

In the preceding sections, we have assumed that N is known. Since α' and $1-\beta$ depend on the maximum number of tests to be performed (N), the choice of N is important.

From sections 2.2.2 and 2.2.3, it may be seen that as N increases, more stringent levels of α' are required and, subsequently, to maintain a specified power, a larger maximum sample size is required. An advantage, though, to frequent testing is a savings in the average number of patients needed. So if H_a is true, the sample size will decrease with increasing N; conversely, if H_o is true, the sample size will increase with increasing N. For example, compared to the number of patients required in a fixed sample size design, the maximum sample size for N = 2 is 10% larger while the average sample size is 22% smaller. Further comparisons to a fixed sample size are given below.

N	% increase in $2n_N$	% decrease in $2n_{\bar{M}}$
3	15	28
5	21	31
7	24	33
9	26	33
20	33	33

For N > 7, the maximum sample size continues to increase while there is no savings in the average sample size. So there appears to be no advantage in doing more than 7 significance tests if reduction in total sample size is the primary motivation for using a group sequential method.

The limit for N, then, may be set at 7. The task of

choosing the "best" N still remains. In choosing N, an investigator should consider the following factors:

1. the maximum length of time of the trial;
2. the expected rate of patient entry;
3. the time lag between patient entry and response;
4. practical arrangements for trial meetings and
5. uncertainty associated with the treatment difference estimate.

The relevance of the first 4 factors to the choice of N is obvious. The length of the trial may be predetermined by the availability of funds or by the investigators. Interim analyses then may be timed to be equally spaced within the limits of the length of the trial. In order to maintain a given power, adequate numbers of patients must be accumulated between tests. With knowledge of the rate of patient entry, the group sizes (2n) in Table 4 serve as a guide for spacing the interim analyses. Also, the length of time to response must be considered, since it is the number of patient results available for analysis that dictates the actual group size. The fourth factor is particularly relevant for multicenter clinical trials where it is necessary to schedule trial meetings well in advance. It may be appropriate to schedule interim analyses to coincide with trial meetings.

The fifth factor, uncertainty in the estimate of the treatment difference, leads to problems in accurately determining the maximum sample size required for a specified

power to be attained. As a consequence, N cannot be carefully chosen. If the treatment difference is estimated as the smallest clinically important difference and power is between .90 and .99, the problem is not critical since the sample size will be adequate for large differences. The problem is critical, though, when investigators consider the largest plausible difference (McPherson, 1982). Uncertainty associated with the latter, therefore, may lead to an underestimation of sample size, more frequent testing and increased risk of not rejecting H_o when H_o is false.

McPherson (1982) takes a Bayesian viewpoint to describe the effect of the uncertainty associated with the treatment difference on the choice of N. To illustrate this effect, the influence of the prior distribution of the treatment difference estimate on maximum sample size and average sample size is measured. If there is a great deal of uncertainty associated with the estimate (a disperse prior distribution), but a large treatment effect is plausible, frequent testing (N from 5 to 10) was found to be optimum. If there is a strong biological basis for expecting a treatment difference, but smaller trials have shown no treatment difference, a fixed sample size test was shown to be optimum. If there is a high degree of certainty associated with the estimate and large effects are not expected, 2 to 4 tests were found to be optimum. These results then may serve as further guidelines for choosing N in conjunction with the

first four factors mentioned.

2.3 Group sequential analysis.

2.3.1 P-values

Commonly, a statistical test results in the reporting of a p-value which is the probability of finding a treatment difference as great or greater than the one observed if H_o is true. For the group sequential method, at each test, a p-value corresponding to Z_i is compared to the nominal significance level. If significance is found at the first test, analysis of the results may be treated as if from a fixed sample size test. That is, an investigator may report an exact p-value corresponding to the Z_i computed from the data. However, for subsequent tests, it is necessary to compute an overall p-value. In other words, the p-value corresponding to Z_i for i from 2 to N should not be reported. Fairbanks and Madsen (1982) give tables of overall p-values for $\alpha = .05$ or .01 and N from 1 to 5 , computed by numerical integration. These values, for $\alpha = .05$, are shown in Table 6 (Fairbanks and Madsen, 1982, Tables 1 and 3 abridged).

For example, suppose an investigator decided to do a maximum of 5 tests. At the fourth test, say Z_i equaled 2.6. The treatment difference, then, may be reported as significant at the .04 level.

Table 6. P-values for a 2-sided significance test with $\alpha = 0.05$

A. When H_o is rejected

N	z_c	i	Final observed Z_i = 2.2	2.4	2.6	2.8	3.0	3.2	3.
1	1.960	1	0.0278	0.0164	0.0094	0.0052	0.0026	0.0014	0.00
2	2.178	1	0.0278	0.0164	0.0094	0.0052	0.0026	0.0014	0.00
		2	0.0486	0.0397	0.0345	0.0318	0.0304	0.0298	0.02
3	2.289	1		0.0164	0.0094	0.0052	0.0026	0.0014	0.00
		2		0.0332	0.0278	0.0248	0.0233	0.0225	0.02
		3		0.0460	0.0416	0.0395	0.0385	0.0381	0.03
4	2.361	1		0.0641	0.0094	0.0052	0.0026	0.0014	0.00
		2		0.0299	0.0242	0.0211	0.0195	0.0187	0.01
		3		0.0404	0.0357	0.0333	0.0322	0.0317	0.03
		4		0.0488	0.0447	0.0429	0.0421	0.0418	0.04
5	2.413	1			0.0094	0.0052	0.0026	0.0014	0.00
		2			0.0221	0.0189	0.0172	0.0163	0.01
		3			0.0319	0.0294	0.0282	0.0277	0.02
		4			0.0398	0.0378	0.0369	0.0366	0.03
		5			0.0464	0.0447	0.0441	0.0438	0.04

B. When H_o is accepted.

N	Z_N = 1.0	1.4	1.6	1.8	2.0	2.2	2.
1	0.3174	0.1616	0.1096	0.0718			
2	0.3214	0.1703	0.1213	0.0867	0.0634		
3	0.3236	0.1732	0.1246	0.0906	0.0681	0.0541	
4	0.3251	0.1749	0.1265	0.0926	0.0702	0.0565	
5	0.3262	0.1762	0.1277	0.0938	0.0715	0.0579	0.05(

2.3.2 Confidence intervals.

The observed treatment difference is an estimate of the true treatment difference. It is useful, therefore, to construct a confidence interval to obtain an indication of the precision of the estimate. The probability that the confidence interval contains the true treatment difference is specified as $1-\alpha$. Restated, if sampling were performed continuously, $(1-\alpha)100\%$ of the intervals computed would contain the true treatment difference.

The limits for the confidence interval at the i^{th} test may be computed as follows (Jennison and Turnbull, 1984):

$$\text{lower limit} = \bar{D}_i - [\sqrt{2}\sigma/\sqrt{(in)}]\, z_c$$

$$\text{upper limit} = \bar{D}_i + [\sqrt{2}\sigma/\sqrt{(in)}]\, z_c$$

where z_c is the critical value corresponding to α' for a N-group sequential analysis. Refer to Table 3 to find the correct z_c for a 95% (α=.05) or 99% (α=.01) confidence interval. For example, for a 5-group sequential analysis, the appropriate z_c is 2.413 for a 95% confidence interval or 2.986 for a 99% confidence interval. For a 90% confidence interval, use z_c from the following table (Jennison and Turnbull, 1984):

N	z_c
2	1.876
3	1.993
4	2.068
5	2.122
6	2.164
7	2.198
8	2.226
9	2.245
10	2.270.

2.4 One-sided group sequential method.

If a directional difference is expected, a one-sided test may be appropriate, since it is more powerful than a 2-sided test. However, power is less than α for a difference in the unexpected direction; that is, one is penalized for asking the wrong question (Hays, 1963). Since clinical trials may reveal results which are the reverse of what was expected, perhaps Armitage (1960) is correct in stating that "one-sided significant tests are inappropriate, in most, if not all, medical trials." Ethics, however, may preclude doing a two-sided test when comparing a standard treatment to a new treatment (Demets and Ware, 1980). Continuing a trial to find a new treatment inferior to a standard treatment may unnecessarily expose patients to the inferior treatment. A one-sided test would allow trial termination when the treatments were found to be equivalent or the new treatment was found to be better than the standard treatment. Additionally, if an investigator is certain of a directional difference and is only interested in testing the significance of this difference, the choice of a one-sided test is correct. But, if the new treatment has some unattractive qualities (e.g. expensive, possible toxic effects, difficult to administer) and the standard treatment is reasonably effective, a conservative test (i.e. two-sided test) may be more appropriate. The choice of a one-sided or a two-sided test, therefore, depends on the investigator's depth of

knowledge of the treatments under study.

Demets and Ware (1980) modified Pocock's two-sided group sequential method to a one-sided method for testing H_o: $\mu_A > \mu_B$. At the ith test, reject H_o if $Z_i > z_c$ and accept H_o if $Z_i < -z_c$. The critical values, computed following the methods of Armitage et al. (1969), are given for N from 2 to 10 in Table 7 (Demets and Ware, 1980). As expected, the nominal significance levels for the one-sided test (Table 7) are less stringent than the corresponding nominal significance levels for the two-sided test (Table 3).

With a one-sided test, a trial may be terminated with acceptance of H_o. However, from Table 8 (Demets and Ware, 1980), $\bar{M}$ under H_o is approximately equal to N; therefore, the savings in patients exposed to the inferior treatment is minimal. For example, with N=4, $1-\beta=.90$ and $\alpha=.05$, the average sample size under H_o is $3.81*10.29\ \sigma^2/\delta^2$ and the maximum sample size is $4*10.29\ \sigma^2/\delta^2$, a difference of about $2\ \sigma^2/\delta^2$ patients.

Compared to the two-sided test (Table 4), the one-sided test affords an appreciable savings in maximum sample size. For a design with N=7, $1-\beta=.90$ and $\alpha=.05$, the maximum sample size (2nN) for the two-sided test is $52.08\ \sigma^2/\delta^2$ vs. $43.33\ \sigma^2/\delta^2$ for the one-sided test. The average sample sizes, for the designs above, are $28.27\ \sigma^2/\delta^2$ and $22.41\ \sigma^2/\delta^2$, respectively, also important savings in sample size. The

Table 7. Nominal significance level α' and corresponding normal deviate z_c required when performing repeated one-sided significance tests (N).

	$\alpha' = 0.05$		$\alpha' = 0.01$	
N	α'	z_c	α'	z_c
2	0.0306	1.876	0.0057	2.532
3	0.0232	1.993	0.0042	2.637
4	0.0193	2.068	0.0034	2.705
5	0.0169	2.122	0.0030	2.754
6	0.0152	2.164	0.0026	2.793
7	0.0140	2.198	0.0024	2.823
8	0.0130	2.226	0.0022	2.848
9	0.0124	2.245	0.0021	2.870
10	0.0116	2.270	0.0020	2.889

Table 8. Average number of one-sided tests ($\bar{M}$) until trial termination and number of patients per group (2n) using various group sequential designs for responses normally distributed with known variance σ^2.

$\alpha = 0.05$

	Under H_0	Under H_a					
		$1-\beta = 0.80$		$1-\beta = 0.90$		$1-\beta = 0.95$	
N	$\bar{M}$	$\bar{M}$	2n *	$\bar{M}$	2n *	$\bar{M}$	2n *
2	1.94	1.51	13.87	1.38	19.01	1.29	23.85
3	2.88	2.05	9.76	1.82	13.31	1.65	16.65
4	3.81	2.59	7.57	2.27	10.29	2.03	12.85
5	4.75	3.14	6.20	2.72	8.42	2.41	10.48
6	5.68	3.68	5.26	3.17	7.13	2.79	8.88
7	6.61	4.22	4.58	3.62	6.19	3.18	7.71
8	7.54	4.77	4.06	4.07	5.48	3.56	6.81
9	8.47	5.31	3.64	4.52	4.92	3.94	6.11
10	9.40	5.85	3.31	4.97	4.46	4.33	5.54

$\alpha = 0.01$

	Under H_0	Under H_a					
		$1-\beta = 0.80$		$1-\beta = 0.90$		$1-\beta = 0.95$	
N	$\bar{M}$	$\bar{M}$	2n *	$\bar{M}$	2n *	$\bar{M}$	2n *
2	1.99	1.57	22.07	1.45	28.39	1.35	34.20
3	2.98	2.17	15.37	1.95	19.71	1.78	23.68
4	3.96	2.78	11.85	2.46	15.16	2.22	18.18
5	4.95	3.38	9.66	2.97	12.33	2.66	14.79
6	5.94	3.98	8.17	3.48	10.42	3.11	12.48
7	6.93	4.58	7.09	4.00	9.02	3.55	10.80
8	7.91	5.18	6.25	4.50	7.96	4.00	9.52
9	8.90	5.79	5.61	5.01	7.13	4.44	8.54
10	9.89	6.39	5.08	5.53	6.46	4.88	7.73

* multiply each entry by σ^2 / δ^2

maximum and average sample sizes for a one-sided group sequential design are consistently smaller than the comparable two-sided group sequential designs. Therefore, when correctly chosen, a one-sided group sequential method is advantageous if the primary objective is to decrease the number of patients exposed to an inferior treatment.

2.5 Generalization of the group sequential method.

Pocock (1977) has shown that the group sequential method is appropriate for clinical trials with design characteristics varying from those described here. For trials with more than two treatments, a separate group sequential design for each treatment comparison or a global significance test is recommended.

If data analysis is scheduled at equally spaced time intervals, the size of the group is likely to vary. A simulation study showed that the latter had essentially no effect on estimates of α and β when using a group sequential method. Similar results were shown when stratifying, though some loss of power was evident when there were unequal treatment numbers within strata.

Pocock (1977) has also shown that the same nominal significance levels hold for a wide variety of response variables. That is, by simulation, estimates of α and β were found to be very close to required levels.

For a normal response with unknown variance, the trial

terminates at the i th test if $t_i > t_{2(in-1),\alpha'}$. Based on 5000 simulations, a minimal loss of power was observed when applying group sequential designs for t-tests (Pocock, 1977).

To compare two exponential means, λ_A and λ_B, a group sequential F test with α' as in Table 3 was found to be appropriate (Pocock, 1977). Simulations showed that estimates of the true overall significance were close to the required α. Also, power was maintained when using sample sizes computed using Table 4 and multiplying each entry by $[\ln(\lambda_A/\lambda_B)]^{-2}$.

A normal approximation of the Wilcoxon test was also found to be appropriate for a group sequential design (Pocock, 1977). Consider trial termination at the ith test if

$$\frac{\{|R_A - [in(2in+1)]/2\}}{in\sqrt{[(in/6)+(1/12)]}} > z_c ,$$

where R_A is the sum of the ranks for treatment A.

For a binary response, there was no loss of power, for $N \geq 5$ and $n \geq 10$. The test statistic, at the ith test, is the following:

$$U_i = \frac{\sqrt{2}\,[R_A(in-R_A)-R_B(in-R_B)]}{\sqrt{[in(R_A+R_B)(2in-R_A-R_B)]}} ,$$

where R_A is the number of responses on treatment A and R_B is the number of responses on treatment B. U_i has a chi distribution with 1 df. Trial termination should be considered when $U_i > z_c$. Use of the Yates continuity correction or the Fisher-Irwin Exact test is found to be less accurate for the group sequential analysis than the test

based on U_i (Pocock, 1977).

Pasternack and Shore (1980) have shown that Pocock's adaptation for binary data is appropriate for epidemiological data. Results of simulations (Pasternack and Shore, 1981) have shown that the group sequential method for 2 X 2 tables is robust for moderate variations in group size, stratified analysis using the Mantel-Haenszel statistic (without the continuity correction) and unequal sample sizes. Table 4 may be used to compute group sizes by making the following substitutions for σ^2/δ^2:

1. for a prospective study, use
$$\bar{p}\,(1-\bar{p})/(p_2-p_1)^2$$
where p_1 is the estimated proportion of patients with the attribute and the disease, p_2 is the estimated proportion of patients without the attribute and with the disease and $\bar{p} = (p_1+p_2)/2$;
2. for a retrospective study, use the same as above except p_1 is the estimated prevalence of the attribute in the diseased population and p_2 is equal to $(1/2)\, p_1 \{1+\sigma/[1+p_1(\sigma-1)]\}$ where σ is the the estimated odds ratio.

Gail et al. (1981) have shown, for the log rank test, that the results of simulations are consistent with theoretical values proposed by Pocock for the normal case when tests are performed at intervals of equal numbers of deaths (d). The following test statistic is computed at times

$t_d, t_{2d}, \ldots, t_{Nd}$:

$$Z_i = \left[\sum_{k=1}^{id} (U_k - P_k)\right] \Big/ \left[\sum_{k=1}^{id} P_k (1-P_k)\right]^{-1/2}, \quad i=1,..,N.$$

where $U_k = \begin{cases} 0, & \text{treatment A} \\ 1, & \text{treatment B} \end{cases}$. P_k is the proportion of patients on treatment A in the trial at time t_{id} and known to have survived for time t_k or longer. Once again, Table IV may be used to compute the additional number of deaths required before each analysis by multiplying entries by Θ^{-2} where, under H_a, the hazard ratio is $\exp(\Theta)$.

It is expected from these results that the group sequential method may be generalized for a wider variety of responses. In the following section, we will examine how well the Pocock design behaves when using the logistic regression model.

CHAPTER III.

GROUP SEQUENTIAL METHOD APPLIED TO THE LOGISTIC REGRESSION MODEL

3.1 Description of the logistic regression model.

The logistic regression model is useful if the treatment response is dichotomous and if covariates must be taken into consideration when measuring treatment effects (Schoenfeld, 1982 and Kleinbaum et al., 1982). Generally, a dichotomous response may be defined as success (Y=1) or failure (Y=0). The probability of a success given K independent variables is modelled by the logistic function as

$$P(Y = 1) = [1 + \exp(-\underset{\sim}{\beta}\underline{X})]^{-1}$$

where $\underline{X}' = (1, X_1, \ldots, X_K)$ is the vector of independent variables and $\underset{\sim}{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)$ is the vector of model parameters. The logistic function is continuous and monotonically increasing from 0 to 1. The shape of the logistic curve is sigmoid with asymptotes at P=0 and P=1.

These properties suggest a wide range of applications in the biological and chemical sciences (Ashton, 1972). Early papers (Reed and Berkson, 1929 and Berkson, 1951) showed the appropriateness of the logistic function for modeling chemical processes. Further applications have been described

in the biomedical sciences, primarily in epidemiological studies (Kleinbaum et al., 1982) and in dose-response analyses (Cox, 1970).

An iterated transformation on P reduces the nonlinear function to a function linear in the β's as follows;

$$P = [1 + \exp(-\underset{\sim}{\beta}\underset{\sim}{X})]^{-1}$$

$$P^{-1} = 1 - \exp(-\underset{\sim}{\beta}\underset{\sim}{X})$$

$$P/1-P = \exp(\underset{\sim}{\beta}\underset{\sim}{X})$$

$$\ln(P/1-P) = \underset{\sim}{\beta}\underset{\sim}{X}$$

The function ln(P/1-P) is commonly referred to as the logit(P) (an abbreviation for logistic unit) and is seen to be an expression for the log odds since P/1-P is the odds of obtaining a success. Two sets of independent variables , say $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$, may be compared by the log odds ratio (ln OR):

$$\ln OR = \text{logit}(P_1) - \text{logit}(P_2) = \underset{\sim}{\beta}(\underset{\sim}{X}_1 - \underset{\sim}{X}_2).$$

Given n observations, $y_1,\ldots,y_n$, where Y has a Bernoulli distribution given $\underset{\sim}{X}$, the probability of success [P(Y=1)] for the ith response is

$$p_i = [1 + \exp(-\underset{\sim}{\beta}\underset{\sim}{x}_i)]^{-1} = [\exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]\,[1 + \exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]^{-1}$$

and the probability of a failure [P(Y=0)] for the ith response is

$$1 - p_i = [1 + \exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]^{-1}.$$

In order to make statistical inferences about the effects of the independent variables on the probability of a success, it is first necessary to estimate the parameters

$\beta_0, \beta_1, \ldots, \beta_K$. Three procedures available are the weighted least squares procedure (Grizzle et al., 1969), minimum chi-square procedure (Ashton, 1972) and the maximum likelihood procedure. The latter procedure is preferred since maximum likelihood estimators are based on sufficient statistics (Cox, 1970) and are asymptotically efficient (Rao, 1973). Also for the case we consider in the following section, the weighted least squares and minimum chi-square procedures are not applicable since the model contains a continuous covariate and the sample sizes are relatively small.

To find maximum likelihood estimators, the likelihood function is required. The likelihood function for a sequence of Bernoulli variables is

$$L(P) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}, \text{ for } p=p_1,\ldots,p_n, \text{ so}$$

$$L(\underset{\sim}{\beta}) = \prod_{i=1}^{n} \{[\exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]\,[1+\exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]\}^{y_i}\,[1+\exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]^{y_i-1}$$

$$= \prod_{i=1}^{n} \exp(y_i \underset{\sim}{\beta}\underset{\sim}{x}_i)\,[1+\exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]^{-1}$$

$$= \exp\left(\sum_{i=1}^{n} y_i \underset{\sim}{\beta}\underset{\sim}{x}_i\right) \prod_{i=1}^{n} [1+\exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]^{-1}$$

The log likelihood (LL) is

$$LL(\beta) = \sum_{i=1}^{n} y_i \underset{\sim}{\beta}\underset{\sim}{x}_i - \sum_{i=1}^{n} \log[1+\exp(\underset{\sim}{\beta}\underset{\sim}{x}_i)]$$

$$= \sum_{j=0}^{K} \beta_j \left(\sum_{i=1}^{n} x_{ij} y_i\right) - \sum_{i=1}^{n} \log[1+\exp(\sum_{j=0}^{K} \beta_j x_{ij})]$$

The maximum likelihood estimators of $\beta_0,\ldots,\beta_K$ may be found by solving the following K+1 likelihood equations simultaneously;

$$\delta LL(\underset{\sim}{\beta})/\delta\underset{\sim}{\beta} = 0$$

A closed form solution does not exist so a Newton-Raphson procedure is used to solve the equations iteratively.

To use the Newton-Raphson procedure, the column vector of first derivatives of LL with respect to $\underset{\sim}{\beta}$ (the score vector) and the matrix of second derivatives of LL with respect to must be computed. The score vector is defined as $U(\underset{\sim}{\beta})$ where the l^{th} entry is

$$\delta LL(\underset{\sim}{\beta})/\delta\beta_l = \sum_{i=1}^{n} x_{il} y_i - \sum_{i=1}^{n} x_{il} \exp\sum_{j=0}^{k}\beta_j x_{ij} \left[1+\exp\sum_{j=0}^{k}\beta_j x_{ij}\right]^{-1}$$

The matrix of second derivatives is defined as $-I(\ )$ where the (l,m) entry is

$$\delta^2 LL(\underset{\sim}{\beta})/\delta\beta_l\beta_m = -\sum_{i=1}^{n} x_{il} x_{im} \exp\sum_{j=0}^{k}\beta_j x_{ij} \left[1+\exp\sum_{j=0}^{k}\beta_j x_{ij}\right]^{-2}$$

$I(\underset{\sim}{\beta})$ is Fisher's information matrix; therefore, the variance-covariance matrix $V(\hat{\underset{\sim}{\beta}})$ is equal to $I^{-1}(\hat{\underset{\sim}{\beta}})$.

To begin the iterative procedure, an initial estimate of $\underset{\sim}{\beta}$, say $\hat{\underset{\sim}{\beta}}_0$, is required. The first iteration, then is given by

$$\hat{\underset{\sim}{\beta}}_1 = \hat{\underset{\sim}{\beta}}_0 + V(\hat{\underset{\sim}{\beta}}_0)\, U(\hat{\underset{\sim}{\beta}}_0).$$

The i th iteration is given by

$$\hat{\underset{\sim}{\beta}}_i = \hat{\underset{\sim}{\beta}}_{i-1} + V(\hat{\underset{\sim}{\beta}}_{i-1})\, U(\hat{\underset{\sim}{\beta}}_{i-1}).$$

The iterative procedure stops when $|\hat{\underset{\sim}{\beta}}_{i-1} - \hat{\underset{\sim}{\beta}}_i| < \underset{\sim}{E}$, where $\underset{\sim}{E}$ is a vector of small positive numbers (e.g., 0.000001). Then

$\hat{\beta}_i$ is the maximum likelihood estimator of .

Once the maximum likelihood estimators have been obtained, hypothesis tests on the 's may be performed. Suppose the independent variable, X_1, is an indicator variable such that $x_{il} = 0$ indicates the ith patient is on treatment A and $x_{il} = 1$ indicates the ith patient is on treatment B. To measure a treatment effect, then, test $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$.

Three asymptotic hypothesis tests available are

1. the likelihood ratio test,
2. the efficient scores test and
3. Wald's test.

For the likelihood ratio test, the test statistic is

$$-2[\, LL(\hat{\beta}^0) - LL(\hat{\beta}) \,]$$

where $\hat{\beta}^0$ is the maximum likelihood estimator under the null hypothesis (i.e. for the restricted model). Hence, assuming no treatment effect ($\beta_1 = 0$), $\hat{\beta}^0$ estimates $\beta_0, \beta_2, \ldots, \beta_K$.

For the efficient scores test, it is only necessary to compute the maximum likelihood estimators under the null hypothesis ($\hat{\beta}^0$). The score statistic (again for testing $\beta_1 = 0$) is given by

$$U(\hat{\beta}^0)' \, V(\hat{\beta}^0) \, U(\hat{\beta}^0)$$

where U and V are the score vector and the variance-covariance matrix, respectively, under the alternative hypothesis (i.e. for the full model).

Wald's statistic for testing a treatment effect is

$$\frac{\hat{\beta}_1^2}{\mathrm{Var}(\hat{\beta}_1)}$$

where $\mathrm{Var}(\hat{\beta}_1)$ is the (2,2) entry of $V(\hat{\underset{\sim}{\beta}})$.

All three statistics have an asymptotic chi-square distribution with one degree of freedom. Therefore each statistic may be compared to the $1 - \alpha$ critical point for the chi-square distribution. The square root of a chi-square statistic with 1 df, say Z_i, is distributed N(0,1) with critical point at $z_{1-\alpha/2}$.

For the group sequential method using the logistic model, at the ith test, it is necessary to determine if $Z_i > z_c$ where z_c is the upper $1 - \alpha'/2$ critical point as given in Table 3.

3.2 Description of the Monte Carlo study.

A Monte Carlo study is designed to derive a numerical solution, considered to be otherwise intractable, from a given model. The model simulates a real system (Rubinstein, 1981). In this paper, the real system is a clinical trial, where the data are defined by the logistic function. Several models are defined to determine the amount of variation in the numerical solutions obtained from the simulations. Each simulation represents a single clinical trial. Repeated independent simulations constitute a Monte Carlo study.

For this Monte Carlo study, the model of interest is

the following logistic regression model:

$$\ln[p/(1-p)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

where X_1 is equal to 0 if the patient is on treatment A and is equal to 1 if the patient is on treatment B, and X_2 is a continuous covariate. A few examples of a continuous covariate are age, weight and blood pressure. The covariate is defined as continous since the logistic regression model is equivalent to a Mantel-Haenzel test if the independent variables are all discrete. Pasternack and Shore (1981) have already shown that Pocock's method is appropriate for the Mantel-Haenzel test. An interaction term is not included in the model since testing $H_0: \beta_1 = 0$ would no longer be meaningful.

A range of values for the true parameters β_0, β_1 and β_2 were computed. The intercept, β_0, is the ln odds when $X_1 = 0$ and $X_2 = 0$. So

$$\beta_0 = \ln[p_A/(1-p_A)]$$

where p_A is the probability of success for an individual on treatment A with covariate value of 0. β_0 was computed for p = .4, .6 and .8. The ln odds ratio for comparing treatment A to treatment B is β_1 regardless of the values of the covariate. This follows since

$$\ln[P_A/(1-P_A)] = \beta_0 + \beta_2 X_2 \quad \text{and} \quad \ln[P_B/(1-P_B)] = \beta_0 + \beta_1 + \beta_2 X_2,$$

and therefore, $$\ln[P_A(1-P_B)/P_B(1-P_A)] = \beta_1,$$

where p_B is the probability of success on treatment B. β_1 was computed for values of p_A and p_B given in Table 9. Note p_A is

chosen to be greater than p_B with no loss of generality.

An acceptable range for β_2 is obtained by assuming that X_2 affects the probablity of a success [P(Y=1)] and then computing β_2 for various pairs of values (x_{i2} and p_i), with β_0 and β_1 as specified above.

A total of 36 models are defined for the Monte Carlo studies and are given in Table 10. Of the 36 models, 24 represent the nonnull case, $\beta_1 \neq 0$ (models 1-24) and 12 represent the null case, $\beta_1 = 0$ (models 25-36).

The purpose of this study is to determine if Pocock's boundaries are appropriate when performing hypothesis tests for the logistic regression model. Therefore, as required for a group sequential design, a maximum number of tests (N), an overall level of significance (α) and power ($1-\beta$) must be predetermined. For this study, N = 5 and α = .05 which correspond to a critical value, z_c, of 2.413 (Table 3). The three test statistics, (the likelihood ratio test statistic, the score statistic and Wald's statistic) are then compared to 2.413 for the 2-sided test of $H_o: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$. The group size (2n), for power $1-\beta = .90$, is determined by multiplying 10.14 (Table 4) by $\bar{p}(1-\bar{p})/(p_A - p_B)^2$, where $\bar{p}=(p_A + p_B)/2$. See Table 9 for values of p_A and p_B and their corresponding group sizes.

Table 9. Group sizes for $1-\beta = .90$ and for various p_A and p_B.

MODELS	p_A	p_B	GROUP SIZE
1-4	0.4	0.3	230
5-8	0.4	0.2	54
9-12	0.6	0.5	250
13-16	0.6	0.4	64
17-20	0.8	0.7	190
21-24	0.8	0.6	54
25-28	0.4	0.4	22
29-32	0.6	0.6	28
33-36	0.8	0.8	26

Table 10. Monte Carlo study models.

Non-null case				Null case (β_1=0)		
Model	β_0	β_1	β_2	Model	β_0	β_2
1	-.4055	.4418	-4	25	-.4055	-4
2	"	"	-2	26	"	-2
3	"	"	2	27	"	2
4	"	"	4	28	"	4
5	-.4055	.9808	-4	29	.4055	-4
6	"	"	-2	30	"	-2
7	"	"	2	31	"	2
8	"	"	4	32	"	4
9	.4055	.4055	-4	33	1.3863	-4
10	"	"	-2	34	"	-2
11	"	"	2	35	"	2
12	"	"	4	36	"	4
13	.4055	.8109	-4			
14	"	"	-2			
15	"	"	2			
16	"	"	4			
17	1.3863	.5390	-4			
18	"	"	-2			
19	"	"	2			
20	"	"	4			
21	1.3863	.9808	-4			
22	"	"	-2			
23	"	"	2			
24	"	"	4			

For each model, 2n observations are generated for each test up to a maximum of 5 tests. There are an equal number of observations on each treatment so $x_{il}=0$ for $i = 1,\ldots,n$ and $x_{il}=1$ for $i = n+1,\ldots,2n$. The continous covariate, x_{i2}, is generated as a normal random variate with mean 0 and variance 1. With respect to the mean and variance, there is no loss of generality because it is analagous to scaling a given covariate, say x_2', so that

$$x_2 = \frac{x_2' - \bar{x}_2'}{SD(x_2')}$$

Given x_{il}, x_{i2} and the parameters $\beta_0, \beta_1, \beta_2$, the probablity of success (p_i) is computed as

$$p_i = \frac{\exp \underset{\sim}{\beta} \underset{\sim}{x}_i}{1 + \exp \underset{\sim}{\beta} \underset{\sim}{x}_i}$$

To generate the y_i's, 2n uniform variates (u_i) are generated and compared to p_i. If $p_i > u_i$ then $y_i=1$ and, likewise, if $p_i < u_i$ then $y_i=0$.

Empirical estimates of α for the null case (models 25-36) and empirical estimates of $1-\beta$ for the non-null case (models 1-24) are computed based on 500 independent simulations for each model. The $\bar{M}$ is measured for the non-null case.

The program for carrying out the simulations is written in PROC MATRIX (SAS) and may be found in Appendix 1.

3.3 Monte Carlo study results and discussion.

3.3.1 Parameter estimates.

As described in section 3.2, the maximum likelihood estimators (MLE) of the model parameters are computed for each trial simulation. The MLE's of β_0 consistently underestimate the true parameters when $\beta_0 = \pm.4055$ by approximately 8%. The absolute differences between the estimates and the true parameters ranged from .010 to .116. For $\beta_0 = 1.3863$, the MLE's overestimated the true parameter by approximately 3% with the absolute difference ranging from .008 to .214.

The MLE's for β_1 overestimate the true parameter values by approximately 11% for a treatment difference of .1 and by approximately 20% for a treatment difference of .2. The absolute differences range from .015 to .223.

The absolute values of the MLE's for β_2 are greater then the true parameters by an average of 6%. The absolute differences between the true and estimated parameters range from .022 to .646.

3.3.2 Non-null case results.

To determine if the sample sizes specified by Pocock (Table 4) are appropriate when using the logistic regression model, power is estimated for the non-null models (1-24). For all models, the estimated power is less then the expected theoretical power of .90 (see Table 11). From Table 11, the maximum estimated power is .814 for model 6 ($\beta_2 = -2$) when

performing the likelihood ratio test. The minimum estimated power is .445 for model 13 (β_2=-4). The effect of the absolute magnitude of β_2 is readily apparent. It is legitimate to consider β_2 in terms of its absolute value since it is the absolute value which reflects the importance of β_2 to the model. For example, β_2 = -4 and +4 both indicate that the covariate strongly influences outcome but in opposite directions. The average estimated powers for models with $|\beta_2|$ = 4 and for models with $|\beta_2|$ = 2 are respectively .521 and .700. It appears, then, that as the influence of the covariate on outcome increases, the power decreases appreciably.

To confirm that the loss of power is due to the addition of the covariate to the model, several models without β_2 are simulated. A power of .90 would be expected based on the work of Pasternack and Shore (1981) and Pocock (1977). The average estimated powers for models with p_A of .4, .6 and .8 are .93, .84 and .72, respectively. For non-null case models when performing LRT, the average estimated powers are .77 (models 2,3,6 and 7), .69 (models 10.11.14 and 15) and .70 (models 18,19,22 and 23). For models with p_A =.4 and .6, there is a considerable loss of power (.16 and .15, respectively) when including $|\beta_2|$=2 in the model. It is interesting to note that the power when p_A =.8 is 18% less than expected. The simulation studies by Pasternack and Shore (1981) ans Pocock (1977) do not contradict these results

since their p_A's ranged from .3 to .55.

A comparison of the results for the three test statistics reveals that the estimated power for the likelihood ratio test (LRT) is always greater than the efficient scores test (EST) and, in turn, the EST is always more powerful than Wald's test (WT) (Table 11). For models with $|\beta_2| = 4$, WT is, on the average, 2.4% less powerful than the EST and 3.3% less powerful than the LRT. For models with $|\beta_2| = 2$, WT is, on the average, 1% less powerful than the EST and 2% less powerful than the LRT. Therefore, it seems that WT becomes more conservative compared to the LRT or EST as $|\beta_2|$ increases.

Given N=5, power=.90 and α=.05, $\bar{M}$, for Pocock's method, is 2.84. However, since the estimated power is considerably less than .90 for all models, it is expected that the number of tests to trial termination would be greater than 2.84. This is, indeed, the case. For $|\beta_2| = 4$, the average number of tests to termination ranges from 3.83 to 4.15 (Table 12, LRT). This is consistent with the theoretical value for the M which is 4.04 (Pocock, 1977 Table 3). Likewise for $|\beta_2| = 2$, the average number of tests to termination ranges from 3.15 to 3.65 (Table 12, LRT) corresponding to the theoretical value of 3.41 for power of .75 (Table 4).

For all models, on the average, WT requires the maximum number of significance tests to trial termination.

However, considering the standard deviations, the three tests do not differ significantly. All three statistical tests averaged approximately 4 significance tests for trial termination given $|\beta_2|=4$ and approximately 3.5 given $|\beta_2|=2$.

3.3.3 Null case results.

To determine if the Pocock boundaries (z_c's in Table 3) are appropriate for the logistic regression model, the size is measured for the null models (25-36). From Table 13, it appears that size is consistent with the expected level of significance of .05. Size ranges from .028 to .067. Averaged over all three statistical tests, the size is .045.

The conservativeness of WT is once again apparent. For all models, the size for WT is less than the size for LRT and EST by approximately .016 and .009, respectively.

The average number of tests to termination for the null case models is approximately 5 ($4.90\pm.51$) as expected.

3.4 Conclusions and suggestions for future work.

For the models simulated, Pocock's boundary when performing a maximum of 5 tests at an overall significance level of .05 ($z_c=2.413$) is appropriate. The risk of making a type I error is minimally changed by the addition of a covariate to the logistic model.

The sample sizes given in Table 4 are inadequate for maintaining a power of .90. When computing sample size for a given power and maximum number of tests, the influence of a

covariate on outcome must be considered. Further simulation studies may be designed to determine the appropriate sample sizes for a range of β_2 to maintain a given power. It would also be interesting to measure the effect on power when modelling more than one covariate.

In conclusion, it appears that Pocock's group sequential method cannot be generalized for the logistic regression model by simply using the values given in Tables 3 and 4 without considerable loss of power.

Table 11. Estimated power for the non-null case when performing the likelihood ratio test (LRT), the efficient scores test (EST) and Wald's test (WT) based on 500 simulations for each model.

$\lvert\beta_2\rvert=2$				$\lvert\beta_2\rvert=4$			
Model	LRT	EST	WT	Model	LRT	EST	WT
2	.742	.742	.742	1	.478	.474	.468
3	.728	.728	.720	4	.524	.522	.512
6	.814	.808	.792	5	.583	.565	.519
7	.812	.800	.782	8	.628	.610	.563
10	.704	.702	.692	9	.468	.464	.458
11	.654	.652	.648	12	.488	.480	.470
14	.716	.706	.698	13	.497	.487	.445
15	.672	.666	.658	16	.496	.482	.456
18	.704	.704	.700	17	.568	.576	.572
19	.756	.752	.742	20	.596	.588	.582
22	.659	.659	.638	21	.558	.548	.514
23	.679	.657	.651	24	.529	.515	.467
MEAN	.720	.715	.705		.535	.526	.502

Table 12. The average number of tests until trial termination and standard deviation* based on 500 simulations for each model.

$|\beta_2| = 2$

Model	LRT	EST	WT
2	3.40 (1.44)	3.40 (1.44)	3.43 (1.42)
3	3.49 (1.45)	3.51 (1.43)	3.55 (1.42)
6	3.15 (1.50)	3.20 (1.49)	3.36 (1.42)
7	3.28 (1.42)	3.31 (1.42)	3.42 (1.37)
10	3.46 (1.47)	3.47 (1.47)	3.50 (1.46)
11	3.65 (1.45)	3.66 (1.45)	3.69 (1.44)
14	3.55 (1.47)	3.59 (1.45)	3.69 (1.40)
15	3.58 (1.45)	3.63 (1.42)	3.71 (1.38)
18	3.51 (1.49)	3.52 (1.49)	3.56 (1.48)
19	3.35 (1.45)	3.36 (1.44)	3.42 (1.43)
22	3.66 (1.46)	3.72 (1.42)	3.89 (1.30)
23	3.54 (1.32)	3.63 (1.29)	3.75 (1.29)

$|\beta_2| = 4$

Model	LRT	EST	WT
1	4.14 (1.31)	4.16 (1.30)	4.20 (1.27)
4	3.90 (1.45)	3.92 (1.43)	3.96 (1.41)
5	3.90 (1.38)	4.02 (1.30)	4.17 (1.16)
8	3.64 (1.47)	3.77 (1.41)	4.04 (1.19)
9	4.15 (1.33)	4.18 (1.31)	4.21 (1.27)
12	4.12 (1.33)	4.15 (1.31)	4.16 (1.30)
13	4.02 (1.40)	4.05 (1.38)	4.26 (1.16)
16	3.95 (1.41)	4.02 (1.37)	4.19 (1.22)
17	3.97 (1.36)	3.89 (1.39)	3.92 (1.38)
20	3.83 (1.42)	3.86 (1.40)	3.90 (1.37)
21	3.90 (1.42)	4.03 (1.31)	4.24 (1.11)
24	3.93 (1.36)	3.99 (1.33)	4.24 (1.11)

* Standard deviation in parentheses.

Table 13. Estimated size for the null case when performing the likelihood ratio test (LRT), the efficient scores test (EST) and Wald's test (WT) based on 500 simulations for each model.

Model	LRT	EST	WT
27	.067	.059	.042
28*	.039	.041	.028
29	.040	.031	.030
30	.051	.045	.035
31	.065	.056	.039
32*	.037	.024	.039
33	.058	.049	.031
34	.056	.043	.035
35	.071	.065	.053
36	.048	.048	.048
MEAN	.053	.046	.037

* Based on 350 simulations.

LITERATURE CITED

LITERATURE CITED

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

Armitage, P. (1960). Sequential Medical Trials. Oxford: Blackwell.

Armitage, P. (1979). The design of clinical trials. Australian Journal of Statistics 21, 266-281.

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

Ashton, W.D. (1972). The Logit Transformation. London: Charles Griffen and Company Limited.

Berkson, J. (1951). Why I prefer logits to probits. Biometrics 7, 327-339.

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

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

Coronary Drug Project Research Group (1972). The Coronary Drug Project: Findings leading to the further modifications of its protocol with respect to

dextrothyroxine. Journal of the American Medical Association 220, 996-1008.

Cox, D.R. (1970). Analysis of Binary Data. London: Methuen and Company Limited.

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

Fairbanks, K. and Madsen, R. (1982). P-values for tests using a repeated significance test design. Biometrika 69, 69-74.

Freiman, J.A., Chalmers, T.C., Smith,Jr.,H. and Kuebler, R.R. (1978).The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. The New England Journal of Medicine 299, 690-694.

Friedman, L.M.,Furberg, C.D. and Demets, D.L. (1981). Fundamentals of Clinical Trials. Boston: John Wright PSG Inc.

Gail, M.H., Demets, D.L.and Slud, E.V. (1982). Simulation studies on increments of the 2-sample logrank score test for survival time data with application to group sequential boundaries. Survival Analysis, J. Crowley and R.A. Johnson (Eds.), 287-301. Hayword, California: Institute of Mathematical Statistics.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. Biometrics 25,

489-504.

Hays, W.L. (1963). Statistics. New York: Holt, Rinehart and Winston, Inc.

Hoel, D.G., Weiss, G.H. and Simon, R. (1976). Sequential tests for composite hypotheses with two binomial populations. Journal of the Royal Statistical Society 38, 302-308.

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

Juhl, E. (1982). Clinical elements of the randomized clinical trial. In The Randomized Clinical Trial and Therapeutic Decisions, N. Tygstrup, J.M. Lachin and E. Juhl (Eds.), 33-42. Boston: John Wright Inc.

Kleinbaum, D.G., Kupper, L.L. and Chambliss, L.E. (1982). Logistic regression analysis of epidemiologic data: theory and practice. Communications in Statistics. 11, 485-547.

Lesser, M.L. (1982). Design and implementation of clinical trials. In Statistics in Medical Research, V. Mike and K. E. Stanley (Eds.), 225-253. New York: John Wiley and Sons.

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

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

McPherson, K. (1982). On choosing the number of interim analyses in clinical trials. Statistics in Medicine 1, 25-36.

Neiss, E. S. and Boyd, T. (1984). Pharmogeneology: the industrial new drug development process. In The Clinical Research Process in the Pharmacy Industry, G.M. Matoren (Ed.), 200. New York: Marcel Dekker Inc.

Novello, F.J. and Bendush, C.L. (1984). The monitoring process. In The Clinical Research Process of Clinical Trials, Good, C.S. (Ed.), 239-272. New York: Marcel Dekker Inc.

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

Pasternack, B.S.and Shore, R.E. (1980). Group sequential methods for cohort and case-control studies. Journal of Chronic Diseases 33, 365-373.

Pasternack, B.S.and Shore, R.E. (1981). Sample sizes for group sequential cohort and case-control study designs. American Journal of Epidemiology 112, 182-191.

Peto, R. Cited in S.J. Pocock (1981). Interim analyses and stopping rules for clinical trials. In Perspectives in Medical Statistics, J.F. Bithell and R. Coppi (Eds.).

London: Academic Press.

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

Pocock,S.J. (1978). The size of cancer clinical trials and stopping rules. British Journal of Cancer 38, 757-766.

Pocock, S.J. (1979). Allocation of patients to treatment in clinical trials. Biometrics 35, 183-197.

Pocock, S.J. (1981). Interim analyses and stopping rules for clinical trials In Perspectives in Medical Statistics, J.F. Bithell and R. Coppi (Eds.). London: Academic Press.

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

Rao, C.R. (1973). Linear Statistical Inference and Its Applications. New York: John Wiley and Sons.

Reed, L.J. and Berkson, J. (1929). The application of the logistic function to experimental data. Journal of Physical Chemistry 33, 760-769.

Schoenfeld, D.A. (1982). Analysis of categorical data: logistic models. In Statistics in Medical Research, V. Mike K.E. Stanley (Eds.), 432-454. New York: John Wiley and Sons.

Smith, R.B. (1976). Clinical trials phases II, III, and IV. In The Principles and Practice of Clinical Trials, C.S. Good, (Ed.),23-37. Edinburgh: Churchill Livingston.

Wald, A. (1947). Sequential Analysis. New York: Wiley.

Weiss, D.G., Williford,W.O., Collins, J.F. and Bingham, S.F. (1983).Planning multicenter clinical trials: a biostatisticians perspective. Controlled Clinical Trials **4**, 53-62.

APPENDIX 1.

Computer Program for Monte Carlo Study

```
//HB09JAM JOB (7958,HB09,30,5),JOY.MELE,CLASS=E
/*SETUP        EXPECTED EXECUTION TIME IS ONE HOUR
/*ROUTE PRINT RMT41
// EXEC SAS,REGION=1000K
***THIS IS PROGRAM IS STORED IN WYLBUR FILE #THESIS;
*****************************************************
*         DESCRIPTION OF THE SIMULATION             *
* THE OBJECTIVE IS TO TEST THE APPROPRIATENESS OF   *
* POCOCK'S GROUP SEQUENTIAL BOUNDARIES AND SAMPLE   *
*      SIZES FOR THE LOGISTIC REGRESSION MODEL.     *
*              THE MODEL IS                         *
*        LN(P/(1-P) = B0 + B1*X1 + B2*X2            *
* WHERE X1 = 0 INDICATES TREATMENT A AND X2 = 1     *
* INDICATES TREATMENT B AND X2 IS A CONTINUOUS      *
* COVARIATE. THE EXPECTED ALPHA LEVEL IS .05 AND    *
* THE EXPECTED BETA LEVEL IS .90. THE MAXIMUM       *
* NUMBER OF TESTS (N) PER TRIAL IS 5. ESTIMATES OF  *
* ALPHA OR BETA WILL BE COMPUTED FOR EACH TEST      *
*                STATISTIC.                         *
*****************************************************;

TITLE OUTPUT FOR MODEL 40;

PROC MATRIX ;
 N = 270 ;                        ** INPUT SAMPLE SIZE;
 G = N#/10;                       ** GROUP SIZE;
 B =1.3863 .9808    4   ;         ** INPUT PARAMETERS;

 *** INITIAL VALUES;
     S = 0;
     C    = 0;
     SUMBETA = 0/0/0;
     T_WALD = 0;    P_WALD = 0;  TW_SS  = 0;
     T_LR   = 0;    P_LR   = 0;  TLR_SS = 0;
     T_SS   = 0;    P_SS   = 0;  TSS_SS = 0;

ZZ_PM = 1 0 -1/1  0 0/1  0 1/1  1 -1/1  1  0/ 1  1  1;
ZZ_PM = 1 -1/1  0/ 1  1;



 *** INPUT VALUES FOR X0, X1 AND X2:
   X0 = J(N,1,1);          ** CREATE A COLUMN VECTOR OF ONES;
   X1_0 = J(G,1,1);        ** COLUMN VECTOR OF G ONES;
   X1_1 = J(G,1,0);        ** COLUMN VECTOR OF G ZEROES;
   X1A = X1_0//X1_1;
   X1 =  X1A//X1A//X1A//X1A//X1A;
   FREE X1_0 X1_1 X1A;
```

```
BEGIN:                          ** SIMULATION RUN STARTS HERE;

   X2 = NORMAL(J(N,1,0)); ** COLUMN VECTOR OF NORMAL RANDOM
                             VARIATES WITH MEAN 0 AND SD 1;

*** FORM THE DESIGN MATRIX X;
   X = X0||X1||X2;

*** FIND THE PROBABILITY OF SUCCESS (P) ;
   NUM = EXP(X*B');
   P = NUM #/ (1+NUM);

*** ASSIGN A RESPONSE TO EACH P;
   Y = P - UNIFORM(J(N,1,0));
   Y = (SIGN(Y) + X0)#/2;

TEST = 0;
FLAG_W = 0;
FLAG_LR = 0;
FLAG_SS = 0;
BETASUM = 0/0/0;



START:                          ** "CLINICAL TRIAL" STARTS HERE;
TEST = TEST + 1;
T = TEST#2#6;

*** WEIGHTED LEAST SQUARES ESTIMATES FOR INITIAL VALUES;
     TT = 1:T; P = X2(TT,);


     P = DESIGN(SIGN(INT(P#/.44))+J(T,1,2)+(3*X1(TT,)))';
     SSIZES = P(,+); II = LOC(SSIZES); SSIZES = SSIZES(II,);
     P = P*Y(TT,); P = P(II,);
     P = P#/SSIZES;  NP = NROW(P);
          DO I = 1 TO NP;
        IF P(I,1)<=0 THEN P(I,1)=.5#/(MAX(SSIZES(I,1)//5));
        IF P(I,1)>=1 THEN P(I,1)=1-(.5#/(MAX(SSIZES(I,1)//5)));
         END;

      F = LOG(P#/(1-P));
      SS =(SSIZES#P#(1-P)); Z_FM = ZZ_FM(II,);
      NLSE = INV(Z_FM'*DIAG(SS)*Z_FM)*Z_FM'*DIAG(SS)*F;
```

```
*** NEWTON-RAPHSON ITERATION PROCEDURE:
 BETA = NLSE ;
 XPY = X(1:T,)'*Y(1:T,);
 FLAG = 0;
 ITERATE = 0;
 EPS1 = 0.001;
 EPS2 = 0.01*EPS1;
   DO WHILE (FLAG=0);
      ITERATE = ITERATE + 1;
          IF MAX(ABS(-X(1:T,)*BETA)) > 174.673 THEN DO;
                 C = C + 1;
                 S = S+1;
                  IF S=500 THEN DO;
                     CHECK = 3;BETASUM = 0/0/0;
                     S = S - 1;
                     GOTO FINISH;
                  END;

                 GO TO BEGIN;
              END;

      I = (J(T,1,1) + EXP(-X(1:T,)*BETA))##-1;
      U = XPY - (X(1:T,)'*I);
      I = I#(J(T,1,1)-I);
    I = X(1:T,)'*(I@X(1:T,));
    DELTA = SOLVE((I#/T),(U#/T));
IF SSQ(ABS(DELTA)#((ABS(BETA)+J(3,1,EPS1))##-1))<EPS2 THEN FLAG=1;
      BETA = BETA + DELTA;
   END;

  BETASUM = BETA + BETASUM;
  V = INV(I);


*** WEIGHTED LEAST SQUARES ESTIMATES FOR INITIAL VALUES:
     TT = 1:T; P = X2(TT,);

     P = DESIGN(SIGN(INT(P#/.44))+J(T,1,2))';
     SSIZES = P(,+):II = LOC(SSIZES); SSIZES = SSIZES(II,);
     P = P*Y(TT,);P = P(II,);
     P = P#/SSIZES;  NP = NROW(P);
        DO I = 1 TO NP;
          IF P(I,1)<=0 THEN P(I,1)=.5#/(MAX(SSIZES(I,1)//5));
          IF P(I,1)>=1 THEN P(I,1)=1-(.5#/(MAX(SSIZES(I,1)//5))):5))
       END;

       F = LOG(P#/(1-P));
       SS =(SSIZES#P#(1-P)); Z_RM = ZZ_RM(II,);

       NLSERM = INV(Z_RM'*DIAG(SS)*Z_RM)*Z_RM'*DIAG(SS)*F;
```

```
*** NEWTON-RAPHSON ITERATION PROCEDURE FOR THE REDUCED MODEL:
 BETA_RM = NLSERM:
 X_RM = X0||X2:
 XPY_RM = X_RM(1:T,)'*Y(1:T,):
 FLAG = 0:
 ITERATE2 = 0:
 EPS1 = 0.001:
 EPS2 = 0.01*EPS1:
   DO WHILE (FLAG=0):
      ITERATE2 = ITERATE2 + 1:
          IF MAX(ABS(-X_RM(1:T,)*BETA_RM)) > 174.673 THEN DO:
                  C = C + 1:
                  S = S+1:
                   IF S=500 THEN DO:
                      CHECK = 3:BETASUM = 0/0/0:
                      S = S - 1:
                      GOTO FINISH:
                   END:

                  GO TO BEGIN:
               END:
      I_RM = (J(T,1,1) + EXP(-X_RM(1:T,)*BETA_RM))##-1:
      U_RM = XPY_RM - (X_RM(1:T,)'*I_RM):
      I_RM = I_RM#(J(T,1,1)-I_RM):
      I_RM = X_RM(1:T,)'*(I_RM@X_RM(1:T,)):
      DELTA = SOLVE((I_RM#/T),(U_RM#/T)):
IF SSQ(ABS(DELTA)#((ABS(BETA_RM)+J(2,1,EPS1))##-1))<EPS2 THEN FLAG=1
     BETA_RM = BETA_RM + DELTA:
 END:


      IF FLAG_W =0 THEN DO:
***WALD'S STATISTIC:
          WALD = (BETA(2,1)##2) # (1#/V(2,2)):
          Z_WALD = SQRT(WALD):
              IF Z_WALD > 2.413 THEN DO:
                      T_WALD = TEST + T_WALD:
                      TW_SS = TW_SS + (TEST##2):
                      FLAG_W = 1:
                      END:
                 ELSE IF TEST = 5 THEN P_WALD = P_WALD + 1:
           END:
        IF FLAG_LR = 0 THEN DO:
```

```
***LIKELIHOOD RATIO TEST:
    B_LL = BETA:      XX=X:      LINK LOGLIKE:  LL_FM = LL:
    B_LL = BETA_RM:   XX=X_RM:   LINK LOGLIKE:  LL_RM = LL:
     LR_STAT = -2 # (LL_RM - LL_FM):
        IF LR_STAT <= 0 THEN LR_STAT = .1:
     Z_LR = SQRT(LR_STAT):
        IF Z_LR > 2.413 THEN DO:
              T_LR = TEST + T_LR:
              TLR_SS = TLR_SS + (TEST##2):
              FLAG_LR = 1:
            END:
       ELSE IF TEST = 5 THEN P_LR = P_LR + 1:
        END:
      IF FLAG_SS = 0 THEN DO:

***SCORE STATISTIC:
      B_S = BETA_RM(1,1)//0//BETA_RM(2,1):
      I_S = (J(T,1,1) + EXP(-X(1:T,)*B_S))##-1:
      U_S = XPY - (X(1:T,)'*I_S):
      I_S = I_S#(J(T,1,1)-I_S):
      I_S = X(1:T,)'*(I_S@X(1:T,)):
         SCORE_ST = U_S' * INV(I_S) * U_S:
         Z_SS = SQRT(SCORE_ST):
             IF Z_SS > 2.413 THEN DO:
                    T_SS = TEST + T_SS:
                    TSS_SS = TSS_SS + (TEST##2):
                    FLAG_SS = 1:
                    END:
             ELSE IF TEST = 5  THEN P_SS = P_SS + 1:
  END:

 CHECK = FLAG_W + FLAG_LR + FLAG_SS:
FINISH: IF (TEST = 5) OR (CHECK = 3) THEN DO:
        S = S+1:
        AVG_BETA = BETASUM #/ TEST:
        SUMBETA  = SUMBETA + AVG_BETA:
         IF S=500 THEN DO:

 NOTE RESULTS FOR WALD STATISTIC:
M_WALD = (T_WALD + (5 * P_WALD)) #/ (500-C):
SDW=SQRT(((TW_SS+(25#P_WALD))-((500-C)#(ASN_WALD##2)))#/(499-C)):
M2_W = (T_WALD) #/ (500-P_WALD-C):
SDW2 =SQRT((TW_SS-((500-P_WALD-C)#(ASN2_W##2)))#/(499-P_WALD-C)):

       POWER_W  = 1 - (P_WALD #/(500-C)):
          PRINT M_WALD SDW M2_W SDW2 POWER_W:
```

```
NOTE RESULTS FOR LIKELIHOOD RATIO STATISTIC:
      M_LR    = (T_LR + (5 * P_LR))#/ (500-C);
    SDL= SQRT(((TLR_SS+(25#P_LR))-((500-C)#(ASN_LR##2)))#/( 499-C));
    M2_LR= (T_LR) #/ (500-P_LR-C);
    SDL2 = SQRT((TLR_SS-((500-P_LR-C)#(ASN2_LR##2)))#/(499-P_LR-C));
    POWER_LR = 1 - (P_LR #/(500-C));
      PRINT M_LR SDL M2_LR SDL2 POWER_LR;




 NOTE RESULTS FOR SCORE STATISTIC:
     M_SS    = (T_SS + (5 * P_SS)) #/(500-C);
     SDS= SQRT(((TSS_SS+(25#P_SS))-((500-C)#(ASN_SS##2)))#/(499-C));
     M2_SS = (T_SS) #/ (500-P_SS-C);
     SDS2 = SQRT((TSS_SS-((500-P_SS-C)#(ASN2_SS##2)))#/(499-P_SS-C));
     POWER_SS = 1 - (P_SS #/(500-C));
       PRINT M_SS SDS M2_SS SDS2 POWER_SS;

       LASTBETA = SUMBETA #/ (500-C);
       DIF_BETA = B' - LASTBETA;
NOTE COMPARE MODEL PARAMETERS TO ESTIMATES:
    PRINT B LASTBETA DIF_BETA;



 NOTE NUMBER OF ITERATIVE PROCEDURES THAT DID NOT CONVERGE:
     PRINT C;

       STOP;
       END;
      ELSE GOTO  BEGIN;
     END;
   ELSE GOTO START;



*** SUBROUTINE TO COMPUTE THE LIKELIHOOD;
LOGLIKE:
 LL = Y(1:T,)' * (YX(1:T,)*B_LL) - SUM(LOG(J(T,1,1)+ EXP(YX(1:T,)*B_LL)));
RETURN;
```

VITA