INTERIM MONITORING AND CONDITIONAL POWER IN CLINICAL TRIALS

by

Yanjie Ren

BS, Southeast University, Nanjing, China, 2013

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Yanjie Ren

It was defended on

April 23, 2015

and approved by

Gong Tang, PhD, Associate Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Hanna Bandos, PhD, Research Assistant Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dianxu Ren, MD,PhD, Associate Professor

School of Nursing, University of Pittsburgh

Thesis Advisor: Gong Tang, PhD, Associate Professor

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Gong Tang, PHD

INTERIM MONITORING AND CONDITIONAL POWER IN CLINICAL TRIALS

Yanjie Ren, M.S.

University of Pittsburgh, 2015

ABSTRACT

Large-scale randomized clinical trials are usually used to compare therapeutic effect of a new treatment to that of a standard treatment. Interim analyses are often performed at several occasions prior to the definitive analysis so that investigators can stop a trial early for ethical or economic reasons if the accumulating data shows enough evidence of superiority or futility of the new treatment. It has been recognized that the boundaries for the test statistics at those interim analyses need to be adjusted so that the overall type I error can be properly controlled. In contrary to adjustment for multiple testing in general practice, the theory on adjustment on boundaries for interim analyses have been well developed in past decades because of their sequential nature. At an interim analysis, one may be interested in estimating the chance for demonstrating the expected efficacy benefit from the new treatment at the definitive analysis. Conditional power provides such assessment based on currently available empirical data. Here we review and compare the operating characteristics of some fundamental methods and extensions in regulating the spending of exit probabilities at interim analyses under the null so that the overall type I error is controlled at the desired nominal level. Then we review the development and calculation of conditional power under a few typical settings. We have applied a few methods on planning of interim analyses and the usage of conditional power to two trials from the National Surgical Adjuvant Breast and

Bowel projects. Well-planned and scientifically-sound early termination of clinical trials save lives, time and expense in the development of better treatments or regimens. Appropriate use of interim analysis design and conditional power, as we promote here, has tremendous public health significance in improving the lifespan and life quality of the population.

Keywords: interim monitoring; conditional power; clinical trials.

TABLE OF CONTENTS

LIST OF TABLES

PREFACE

I would like to offer my sincere gratitude to Dr. Gong Tang, my advisor. In the past two years, he has supported me in research with his knowledge, enthusiasm, and commitment. Without his guidance and inspiration, the my progress would not have been possible. Besides research support, he has also shared with me his wisdom and experience in life.

I thank my parents, for giving me the opportunity of life, and their unconditional and consistent love.

1.0 INTRODUCTION TO CLINICAL TRIALS

A new medical therapy may go through several stages of testing on human subjects after showing promise in animal experiments. Phase I clinical trials mainly concern about human safety. The aim is to find a suitable dose. Normally, they involve 20 to 80 participants who may benefit from the treatment under investigation. Once the dose is determined, the next step is to assess the efficacy of the new treatment. Phase II clinical trials evaluate the biological effect of the treatment and continue to monitor the safety aspect on a larger scale of subjects. They usually performed on between 100 and 300 diseased patients. Phase III clinical trials are randomized trials designed to compare the effectiveness of the new treatment to standard treatment. They usually involve more than 1000 patients and last 3 to 5 years or longer depending on recruitment rate or follow-up time needed. Therefore, they are very expensive and time-consuming. It is the Phase III clinical trial draws the most attention of statistical design and statistical analysis.[Jennison and Turnbull, 1999]

We focus on clinical trials that are aimed to show superiority of a new treatment over a standard treatment or placebo in the thesis, but there are equivalence studies designed to show that the efficiency of a treatment is the same as or no worse than the existing treatment which may be safer, less expensive or easier to use. It should be noted that failure to show the treatment difference maybe due to small sample size and subsequent low power.

In clinical trials, the endpoint often refers to an outcome variable of interest. It can belong to various types. A dichotomous outcome of success or failure is common, especially

in assessment of treatment efficacy. A quantitative outcome is one measured on a continuous scale and is often modeled by a normal distribution after appropriate transformation. A survival outcome is time of study entry to the first treatment failure or patient death. Sometimes, we do not observe the event of interest at the time of analysis. In this case, we say the patient is right censored and record the partial information that the time to event exceeds the follow-up time.

In any experiment or survey in which data accumulate steadily over a period of time, it is natural to monitor results as they occur so that actions such as modifications of study design or early termination can be taken. The many reasons for performing interim analyses can be categorized into three classes: ethical, administrative and economic. In trials involving human subjects, there is an ethical need to monitor results to ensure that individuals are not exposed to an unsafe, inferior or ineffective treatment regimen. Even in negative trials where there appears to be no difference in the performance of two therapies there is ethical imperative to terminate a trial as soon as possible so that resources can be allocated to study the next most promising treatment waiting to be tested. Ethical considerations prescribe that accumulating data be evaluated frequently, not only with regard to internal comparisons of safety and efficacy, but also in the light of new information from outside the trial.

One administrative reason for interim monitoring is to ensure that the experiment is being carried out correctly, that the subjects are from the right population and satisfy eligibility criteria and the test procedures are as prescribed in the protocol. An early examination of interim results helps to reveal the presence of problems which can be corrected before too many resources are used. Another administrative reason is to examine the assumptions made in designing the trial. For dichotomous outcome, the sample size calculations rely on the assumed value of the background incidence rate; for quantitative outcome, the sample size calculation is often set this variable to be normally distributed with a certain variance.

For time to event outcome, the accrual period was determined on estimating the subject accrual rate. An interim analysis result can reveal inappropriate assumptions in time for adjustments to be made.

There are also economic benefits gained from interim monitoring. For a trial with positive result, early stopping means that a new product can be put into market earlier. Even with a negative result, early termination ensures that resources can be allocated to next most promising treatment waiting to be tested before too many are wasted.

2.0 INTERIM MONITORING: STOPPING FOR BENEFICIAL

These reasons call for the performance of interim monitoring. However, this involves performing multiple significance tests at different stages during the accumulation of data. As we know, if we test 20 hypotheses separately at an alpha level of 0.05, the probability that we observe at least one significant result just due to chance will be inflated to 1-$(1-0.05)^{20} = 0.64$. That is significantly higher than the nominal level of 0.05. That means even if all the tests are actually non-significant, there is 64% chance of observing at least one significant result.

Similarly, if we compute the test statistic at each interim look in a clinical trial and declare statistical significance if it ever exceeds some nominal alpha level, say, 1.96, the probability of eventually reaching a significant result will be inflated.

Methods such as Bonferroni corrections, the false discovery rate and the positive false discovery rate were proposed in the genomics field to control the type I error by adjusting alpha level. However, the multiple testing problem in sequential methods is different in nature that the data is accumulating and if at some stage, the null hypothesis is rejected, the experiment will be terminated.

Armitage et al. [1969] proposed a numerical integration approach to calculate the probability of exceeding the critical values in repeated significance tests on accumulating data.

The sitting Armitage proposed is to conduct a trial consisting of a series of experiments

X_1,X_2,...X_n whose response variable is normally distributed with zero mean and unit variance. After each experiment is conducted, the cumulative sum $S_n = \sum_{i=1}^{n} X_i$ is used to test the null hypothesis. The trial stops when, for the first time, the cumulative sum exceeds a predetermined boundary y_n,

Denote f_n the probability density function of S_n in the sequential procedure. It is given by

$$f_n(S_n) = \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u) \frac{1}{\sqrt{(2\pi)}} \exp\{-\frac{1}{2}(S_n - u)^2\}\, du. \tag{2.1}$$

By recursively using the formula, the density function f_n can be defined, with f_1 being the standardized normal density function.

The probability of stopping the trial for superiority before or at the n experiments is

$$P_n = 1 - \int_{-y_n}^{y_n} f_n(u)\, du.$$

$$\begin{aligned}
P_n - P_{n-1} &= 1 - \int_{-y_n}^{y_n} f_n(v)\, dv. - 1 + \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\, du \\
&= \int_{-y_n}^{y_n} \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(v) \frac{1}{\sqrt{(2\pi)}} exp\{-\frac{1}{2}(u-v)^2\}\, dvdu + \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\, du \\
&= \int_{-y_n-1}^{y_n-1} f_{n-1}(u) \int_{-y_n}^{y_n} \frac{1}{\sqrt{(2\pi)}} exp\{-\frac{1}{2}(u-v)^2\}\, dudv + \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\, du \\
&= \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\{1 + \int_{-y_n}^{y_n} \frac{1}{\sqrt{(2\pi)}} exp\{-\frac{1}{2}(u-v)^2\}\, dv\}\, du \\
&= \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\{1 + 1 - 2\Phi(y_n - u)\}\, du \\
&= 2\int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u)\{1 - \Phi(y_n - u)\}\, du
\end{aligned}$$

An alternative form would be:

$$P_n - P_{n-1} = 2\int_{y_n}^{\infty} \int_{-y_{n-1}}^{y_{n-1}} f_{n-1}(u) \frac{1}{\sqrt{(2\pi)}} \exp\{-\frac{1}{2}(u-v)^2\}\, dvdu.$$

Armitage [1975] used Newton-Cotes formula and Simpson's formula to evaluate the right hand side of (2.1).

This numerical quadrature method was later used by other statisticians to approximate the type I error rates and exit probabilities in sequential methods.

To show how bad the naive method could be, we considered performing interim analyses in three different situations. The probabilities of exceeding critical levels in repeated significance tests on accumulating data is given by Table 1.

First, we consider performing interim looks frequently at early stage, say the time points set to be at 1/16, 1/8, 1/4, 1/2 and the end of the trial. 1/2 means the value of the outcome variable is ascertain on half of the subjects. Then we consider randomly selecting time points in the trial, except for the definitive analysis at the end of the trial. Last, we consider situation that the interim analyses are performed at the late stage of the trial, say 1/2, 3/4, 7/8, 15/16 and the end of the trial for a trial with five looks. The results in the table is integrated using algorithm proposed by Schoenfeld [2001].

Table 1: Type 1 error rates by the total number and timing of looks for a two-tailed test at $\alpha = 0.05$ in three situations

N	early monitoring	random monitoring	late monitoring
2	0.0831	0.0816	0.0831
3	0.1135	0.1039	0.0973
4	0.1424	0.1222	0.1039
5	0.1702	0.1369	0.1073
6	0.1971	0.1490	0.1090
7	0.2232	0.1600	0.1100
8	0.2483	0.1704	0.1105
9	0.2727	0.1783	0.1108
10	0.2963	0.1869	0.1110
11	0.3191	0.1932	0.1110
12	0.3412	0.1999	0.1111
13	0.3625	0.2080	0.1111
14	0.3832	0.2120	0.1111
15	0.4032	0.2176	0.1112
16	0.4225	0.2225	0.1112
17	0.4413	0.2263	0.1112
18	0.4594	0.2329	0.1112
19	0.4769	0.2359	0.1112
20	0.4939	0.2397	0.1112

From table 1, we can see that repeated significance tests without adjusting alpha level lead to the inflated type I error rate. The probabilities of exceeding critical levels are significantly higher if the tests are taken frequently at the early stage. For a trial with ten looks, the cumulative type I error can be inflated to 0.2963. That is nearly six times of the predetermined nominal alpha level.

Armitage [1975] later proposed a useful sequential method by adjusting alpha level for each test. The idea is that after every observation, a significance test is carried out and if a nominal significance level α^* is achieved, the trial is terminated. Here α^* and the maximum number of observations N are chosen so that the overall significance level α is set to the required level, say, 0.05.

2.1 THE POCOCK BOUNDARY

However, the method was applied to only a small fraction of clinical trials because the need for continuous surveillance of every observation can be very difficult and extremely inefficient. Proschan et al. [2006] adapted the method into a group sequential design, which divides patient entry into a number of equal-sized groups so that the decision to stop the trial or continue is based on repeated significance tests of the accumulated data after each group is evaluated.

For a clinical trial with two treatments, A and B, a homogeneous sample of patients was randomized into two treatments, so each consecutive group has n patients. The maximum number of interim looks is denoted as N. Let the response of the treatments is a normal random variable with unknown mean μ_A and μ_B and known variance σ^2. Denote $\bar{x}_{Aj}$ and $\bar{x}_{Bj}$ the sample means of treatment response for the jth interim analysis.

$$\bar{d_i} = \sum_{j=1}^{i} (\bar{x}_{Aj} - \bar{x}_{Bj})/i$$

is N($\mu_A - \mu_B$, $2\sigma^2/in$) After ith interim analysis, the p value for a two-sided significance test of the null hypothesis of $\mu_A = \mu_B$ is

$$P_i = 2[1 - \Phi\{\sqrt{(in)}\bar{d_i}/\sqrt{2}\sigma\}]$$

Pocock set the nominal significance level α^* for each test. If $P_i < \alpha^*$, one stops the trial, claiming treatment difference exists. If $P_i > \alpha^*$, for all i=1,2, ..., N, The trial ends, declaring that there is no treatment difference. Here α^* and N is chosen, so the overall significance level of α is controlled. Since $(\bar{x}_{Aj} - \bar{x}_{Bj})$ for j=1,2, ..., N is normally distributed with mean $\mu_A - \mu_B$ and variance $2\sigma^2/n$, given α and N, the nominal α^* does not depend on n. Compared to the standard sequential method proposed by Armitage, Pocock was actually interested in small value of N.

Table 2 gives the Pocock boundaries at alpha level of 0.05 and 0.01. For example, a ten-look trial with overall alpha level 0.05, the nominal alpha is set to be 0.0106 for each significance test.

Table 2: the nominal significance level α^* and corresponding z-score for normal group sequential testing, with known variance, by the number of group N and overall significance level α

	$\alpha = 0.05$		$\alpha = 0.01$	
looks	α^*	z	α^*	z
2	0.0293	2.179	0.0056	2.772
3	0.0220	2.290	0.0041	2.873
4	0.0182	2.362	0.0033	2.939
5	0.0158	2.414	0.0028	2.987
6	0.0141	2.454	0.0025	3.024
7	0.0129	2.486	0.0023	3.053
8	0.0120	2.513	0.0021	3.078
9	0.0112	2.536	0.0019	3.099
10	0.0106	2.556	0.0018	3.117
11	0.0101	2.573	0.0017	3.133
12	0.0097	2.588	0.0016	3.148
13	0.093	2.602	0.0016	3.161
14	0.0089	2.615	0.0015	3.172
15	0.0086	2.627	0.0015	3.183
16	0.0084	2.637	0.0014	3.193
17	0.0081	2.647	0.0014	3.202
18	0.0079	2.656	0.0013	3.210
19	0.0077	2.665	0.0013	3.218
20	0.0075	2.672	0.0013	3.225

2.2 THE O'BRIEN-FLEMING BOUNDARY

The Pococks procedure requires the same level of evidence for early and late looks in a trial. However, people prefer to require a greater level of evidence to stop the trial in early stage when there is large variability. Also, most of the probabilities are dedicated to early looks in Pocock's procedure while we do not want to pay such a high price at the end of the trial.

To solve these problems, O'Brien and Fleming [1979] proposed an alternative method to periodically review the results accumulating from the study, allowing for the early termination of the trial.

The proposed boundaries allow the exit probabilities increase as the information available increases. Returning to Pocock's example, a trial with n looks, for the ith look, O'Brien and Fleming O'Brien and Fleming [1979] proposed to reject the null hypothesis and stop the trial if the Z-score exceeds $P(n, \alpha)/(i/n)^{1/2}$, where $P(n, \alpha)$ is pre-determined so that the total type I error rate is controlled at α. If the null is not rejected at any of the interim analyses and the test statistic at the definitive analysis does not exceed $P(n, \alpha)$, the study is concluded with the result that the null hypothesis cannot be rejected at the α significance level. The calculation of the O'Brien-Fleming boundaries can be done by gradually adjusting and choosing the value of $p(n, \alpha)$ so that the total type I error rate is α.

Table 3 Gives the O'Brien-Fleming boundary $P(n, \alpha)$ for a two-tailed test at $\alpha = 0.05$ and 0.01 and different number of looks. For example, the tabled value for ten looks and $\alpha = 0.05$ is $P(10, 0.05) = 2.087$. The boundaries for Z(1/10), ..., Z(10/10) would be $2.087/(1/10)^{1/2} = 6.600$, $2.087/(2/10)^{1/2} = 4.667$, $2.087/(3/10)^{1/2} = 3.810$, ..., $2.087/(10/10)^{1/2} = 2.087$.

Table 3: Approximate Values of the OBrien-Fleming boundary P(n,α) by the number of looks and overall significance level α

looks	$\alpha = 0.05$	$\alpha = 0.01$
2	1.978	2.580
3	2.005	2.595
4	2.025	2.610
5	2.041	2.622
6	2.053	2.632
7	2.064	2.641
8	2.073	2.648
9	2.080	2.655
10	2.087	2.660
11	2.093	2.666
12	2.098	2.670
13	2.103	2.674
14	2.107	2.678
15	2.111	2.682
16	2.114	2.685
17	2.118	2.688
18	2.121	2.691
19	2.124	2.693
20	2.126	2.696

Table 4 and table 5 give the cumulative type I error rates and exit probabilities given by O'Brien-Fleming boundaries, respectively.

Table 4: Cumulative type I error rates of the OBrien-Fleming boundaries by number of looks

looks										
2	0.0026	0.025								
3	0.0003	0.0071	0.0249							
4	< 0.0001	0.0021	0.0104	0.025						
5	< 0.0001	0.0006	0.0044	0.0128	0.0249					
6	< 0.0001	0.0002	0.0019	0.0066	0.0145	0.025				
7	< 0.0001	0.0001	0.0008	0.0035	0.0085	0.0158	0.025			
8	< 0.0001	< 0.0001	0.0004	0.0018	0.005	0.01	0.0168	0.0249		
9	< 0.0001	< 0.0001	0.0002	0.001	0.003	0.0064	0.0114	0.0176	0.025	
10	< 0.0001	< 0.0001	0.0001	0.0005	0.0018	0.0041	0.0077	0.0125	0.0183	0.025

Table 5: Exit probabilities of the O'Brien-Fleming boundaries by number of looks

looks										
2	0.0026	0.0224								
3	0.0003	0.0069	0.0178							
4	< 0.0001	0.0021	0.0083	0.0145						
5	< 0.0001	0.0006	0.0038	0.0083	0.0122					
6	< 0.0001	0.0002	0.0017	0.0047	0.0079	0.0105				
7	< 0.0001	0.0001	0.0008	0.0026	0.005	0.0073	0.0092			
8	< 0.0001	< 0.0001	0.0003	0.0015	0.0032	0.005	0.0068	0.0082		
9	< 0.0001	< 0.0001	0.0002	0.0008	0.002	0.0035	0.0049	0.0063	0.0074	
10	< 0.0001	< 0.0001	0.0001	0.0004	0.0012	0.0024	0.0036	0.0048	0.0058	0.0067

2.3 COMPARISON OF POCOCK AND O'BRIEN BOUNDARIES

Pocock procedure requires same level of evidence at the early and late looks, making it easy to stop at the early stage. While O'Biren-Fleming procedure makes it much more difficult to stop early and it extracts a smaller price at the end compared to Pocock's procedure.

The Pocock cumulative type I error rate increases dramatically in the early stage but slowly at the end. In contrast, the O'Biren-Fleming cumulative type I error rate increases very slowly at the beginning and increases sharply at the end.

2.4 INFORMATION FRACTION AND BROWNIAN MOTION

The Pocock and O'Brien-Fleming boundaries both require a pre-specified number of equally spaced looks, but people want more flexibility. They may want to monitor the trial more frequently if the statistic in prior test is near the boundary or they may want to postpone a meeting if the statistic is far away from the boundary.

To construct boundaries that do not require pre-specified number or the timing of the looks, the concept of information time has to be briefly introduced here. Details can be found at [Lan and Zucker, 1993] and [Proschan et al., 2006].

Assume X_1, X_2, ..., X_N, are independently and identically distributed (i.i.d) random variables with unknown mean θ and known variance σ^2. To test the null hypothesis H_0 : $\theta = 0$, We use the z statistic

$$Z_N = \frac{S_N}{\sqrt{V_N}}$$

where $S_N = \sum_{i=1}^{N} X_i$ and $V_N = Var\{S_N\} = Nvar\{X_i\}$ For an interim analysis after nth observation is evaluated,

$$Z_N = \frac{S_n + S_N - S_n}{\sqrt{V_N}} = \frac{S_n}{\sqrt{V_N}} + \frac{S_N - S_n}{\sqrt{V_N}} \tag{2.2}$$

S_n and $S_N - S_n$ are independent, so the two components in (2.2) are independent. $t = v_n/v_N$ is called information fraction, which measures how far a trial has processed. In this case, it is the fraction of patients enrolled relative to the planned sample size, n/N. t=0 means the beginning of the trial and t=1 means the end of the trial.

Denote n(t) as the information after results of nth patients available. Define $Z(t) = Z_{n(t)} = \frac{S_n}{\sqrt{v_n}}$, and define

$$B(t) = \frac{S_n}{\sqrt{V_N}} = \sqrt{t}Z(t) \tag{2.3}$$

At the end of the trial $t = 1$, $B(1) = Z(1) = \frac{Y_N}{\sqrt{V_N}}$ So (2.2) becomes $B(1) = B(t)+B(1-t)$, which implies that the change in B score from now to the end of the trial B(1-t) is independent of the current B-score, B(t).

More generally, B(t) has independent increments, meaning that changes in B values over non-overlapping intervals are independent, which is an advantage of using B score instead of z score.

From (2.3), Var(B(t)) = t Var{Z(t)} = t,

For $t_i < t_j$,

$$Cov\{B(t_i), B(t_j)\} = cov\{\frac{S_{ni}}{\sqrt{V_{ni}}}, \frac{S_{nj}}{\sqrt{V_{nj}}}\} = t_i$$

So the distribution of B(t) has these properties,

$$E\{B(t)\} = 0$$

$$Var\{B(t)\} = t$$

$$Cov\{B(t_i), B(t_j)\} = t_i, \quad for \quad t_i < t_j$$

Which correspond to the properties of Z(t),

$$E\{Z(t)\} = 0$$

$$Var\{Z(t)\} = 1$$

$$Cov\{Z(t_i), Z(t_j)\} = \sqrt{\frac{t_i}{t_j}}, \quad for \quad t_i < t_j$$

Note that $B(t)$ is defined only at the discrete information times t=0, 1/N, ..., N/N. It can be extended to the continuous time scale ($\tau \in [0, 1]$) by setting $B(\tau) = 0$ for $\tau < t_1$ and $B(\tau) = B(t_n) = B(n/N)$ for τ in the interval $[t_n, t_{n+1}]$.

As N goes infinity, the set of t at which $B_N(t)$ is non-differentiable becomes more and more dense. In the limit, we get standard Brownian motion, which is defined as a stochastic process $\{B(\tau) : \tau \in [0, 1]\}$ such that for any $\tau_1 < \tau_2 < \cdots < \tau_p \in [0, 1]$, the random vector $\{B(\tau_1), ..., B(\tau_p)\}$ has a multivariate normal distribution with mean zero and covariance given by $cov(B(\tau_q), B(\tau_r)) = min\{\tau_q, \tau_r\}$. Interim monitoring often uses Brownian motion approximation because the properties of Brownian motion is well studied.

Next, consider a clinical trial with binary outcomes, say success or failure. Let $S_A(i)$ and $S_B(i)$ the indicators for success for the ith pair of subjects randomized into treatments A and B, respectively. Denote D_i as the difference: $D_i = S_A(i) - S_B(i)$. Assuming that $S_A(i)$s and $S_B(i)$s are i.i.d. and follow a Bernoulli distribution with success probability p, then D_is are i.i.d variables with mean 0 and variance $2p(1-p)$. The test statistic for the two-group comparison is given by $Z_N = \frac{\sum_1^N Di}{\sqrt{v_N}}$, where $v_N = var(S_N) = 2Np(1-p)$. p is

usually approximated by the sample proportion in calculations. For unpaired trials, the test statistic is often calculated as

$$Z_N = \frac{\sqrt{N}(\bar{p}_A - \bar{p}_B)}{\sqrt{2\hat{p}(1-\hat{p})}}$$

with $\bar{p}_A$ and $\bar{p}_B$ as the group-wise proportion of successes and $\hat{p}$ the pooled proportion of successes. Brownian motion still provides a good approximation for trials with a binary outcome.

Consider the time-to-event response. Still compare the treatment effect of A and B. For simplicity, consider the event that can only happen only once for each patient. Here, denote N as the total number of patients with events at the end of a trial. For the ith event time, let O_i be the indicator that whether the patient with event is in the treatment group A : $O_i = 1$ if the ith event occurs in treatment A, otherwise O_i is 0. Let Y_A and Y_B be the number of patients at risk just prior to the ith event. Then the expectation of O_i is $E_i = Y_A/(Y_A + Y_B)$ under the null hypothesis that there is no difference in the event risk between treatment groups A and B. Conditioning on the marginal values Y_A and Y_B, O_i follows a Bernoulli distribution with success parameter E_i. Let $D_i = O_i - E_i$. Under the null hypothesis, D_i has mean 0 and variance $V_i = E_i(1 - E_i)$. Unconditionally, the D_is are uncorrelated, mean-zero random variables with variance V_i under the null hypothesis. Conditioning on N, $v_N = var(S_N) = \sum_{i=1}^{N} var(D_i) = E(\sum V_i)$. The log-rank statistic is given by $Z_N = \frac{\sum_{i=1}^{N} D_i}{\sqrt{v_N}}$, where $v_N \approx \sum_{i=1}^{N} V_i$.

The information fraction in survival settings is defined in terms as patients with events instead of patients enrolled. If we condition on N and n, the covariance structure of Brownian motion holds. The Brownian motion is again a good approximation to the process B(t). A practical problem is that at the interim analysis, we would not know or approximate v_N. Under the null hypothesis, $E(V_i) = E\{E_i(1-E_i)\} \approx (1/2)(1-1/2) = 1/4$. Without making any assumptions about the form of the survival curve, this simple argument shows that the variance of D_i is approximately $1/4$. It follows that $v_N \approx N/4$. This leads to that the information fraction by the nth event is $t = n/N$.

2.5 SPENDING FUNCTIONS

Spending functions were discussed by Proschan et al. [2006] to construct boundaries that do not need pre-determined number or timing of looks, giving more flexibility to monitoring.

The key to making boundaries more flexible is to consider the cumulative type I error rate used by different information time. We used a trial with five and ten interim looks as an example. The type I error rates at different information time were given in Table 6.

Table 6: Cumulative type 1 error rates used by the O'Brien-Fleming procedure with five and ten looks and one-tailed $\alpha = 0.025$

number of looks	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
5	NA	0.0000	NA	0.0006	NA	0.0045	NA	0.013	NA	0.025
10	0.0000	0.0000	0.0000	0.0005	0.0018	0.0041	0.0077	0.012	0.018	0.025

Table 6 shows that type I error rates used by the information fractions common to five and ten looks, namely t=0.2, 0.4, 0.6, 0.8, 1 are almost the same for 5 and 10 looks. Therefore, doubling the number of looks does not considerably change its value at previously existing support points. Imagine doubling the number of looks to infinity. The O'Brien-Fleming boundary a_k approaches α as k $\to \infty$, where α is such that $Pr(B(s) > \alpha$ for some $s \leq 1) = \alpha$. The probability of crossing the boundary by time t is

$$\alpha_1(t) = Pr(B(s) > \alpha \text{ for some } s \leq 1) = \alpha$$

$\alpha_1(t)$ is an increasing function defined on all of [0,1] with $\alpha_1(0) = 1$ and $\alpha_1(1) = \alpha$. Instead of specifying the number and timing of looks, we can specify a spending function telling how much alpha to use by information time t.

2.5.1 Linear spending function

To illustrate how the spending function works, we use a linear spending function as an example.

let

$$\alpha_2(t) = \alpha t, \quad 0 \leq t \leq 1$$

with $\alpha = 0.025$ and one-tailed testing. Suppose the first look occurs at information fraction $t = 0.2$. We spend $\alpha_2(0.2) = 0.025(0.2) = 0.005$ at the first interim analysis and therefore determine a critical value c_1 such that $Pr\{Z(t_1) > c_1\} = 0.005$. The corresponding boundary is $c_1 = \Phi^{-1}(0.995) = 2.576$. We reject the null hypothesis at the first analysis if $Z(0.20) > 2.576$. Suppose this does not happen, the next interim analysis occurs at information time t=0.4. The cumulative type I error rate by t=0.4 is $\alpha_2(0.4) = 0.025(0.4) = 0.01$. We determine the boundary c_2 such that $Pr[\{Z(0.2) > 2.576\} \cup \{Z(0.4) > c_2\}] = 0.01$. We use numeric integration to get the value of c_2. Thus, we reject the null hypothesis at the second analysis if $Z(0.4) > 2.492$. Suppose this does not happen and the next analysis occurs at t=0.6. The cumulative type I error at t=0.6 is 0.025(0.6)=0.015. We determine the value c_3 such that $Pr[\{Z(t_1) > c_1\} \cup \{Z(0.4) > 2.492\} \cup \{Z(0.6) > c_3\}] = 0.015$, which yields $c_3 = 2.411$. We reject the null hypothesis at the third analysis if $Z(0.6) > 2.411$. Suppose that does not happen either and the final analysis occurs at the end of the trial. The cumulative type 1 error rate by t=1 is $\alpha_2(1) = 0.025(1) = 0.025$,which yields c_4=2.186.

In this example we found c_i iteratively such that

$$Pr[\cup_{i=1}^{j}\{Z(t_i) > c_i\}] = \alpha(t_j), j = 1, ..., k$$

$$Pr[\cap_{i=1}^{j}\{Z(t_i) \leq c_i\} \cup \{Z(t_j) > c_j\} = \alpha(t_j) - \alpha(t_{j-1}), j = 1, ..., k$$

Here, we used linear function as an example, but we could have used other spending functions.

3.0 STATISTICAL POWER

Statistical power is the probability of rejecting the null hypothesis when the alternative hypothesis is true. It is the likelihood of distinguishing an effect of certain size from pure chance. A study might easily detect large treatment difference, but less likely to detect a subtle treatment difference. In clinical trials, we might miss a valuable therapy because of lacking statistical power to detect the treatment effect. In fact, in a sample of randomized clinical trials published between 1975 and 1990 with negative results, Moher et al. [1994] found that 64% of the studies can not detect a 50% relative difference. Even a treatment can reduce the symptoms by 50% compared to other treatment, there is no sufficient data to conclude that is efficient.

3.1 UNCONDITIONAL POWER

Therefore, adequate sample size and power are essential for a well-designed clinical trial. For asymptotically normal statistics with mean θ and unit variance, we equate θ, the expected z-score, to $z_{\alpha/2} + z_{\beta}$ and solve for either the sample size or power.

For power calculation:

$$\theta = z_{\alpha/2} + z_{\beta}$$

$$Power = 1 - \beta = \Phi(\theta - z_{\alpha/2})$$

Consider a trial with dichotomous outcome using a test of proportions. The expected death in the control arm is 0.6. And we want to have 80 percent power to detect a 20 percent reduction in the treatment arm.

The expected z-statistic is $\theta = \frac{\sqrt{N}(p_A - p_B)}{\sqrt{2\hat{p}(1-\hat{p})}}$, where $p_A = 0.6$ and $p_B = 0.4$, $\hat{p} = (0.6 + 0.4)/2 = 0.5$. To achieve 80 percent power, we set $\theta = 1.96 + 0.85 = 2.81$ and solve for N, which yields N=$\frac{2(0.5)(1-0.5)(1.96+0.85)^2}{(0.6-0.4)^2} \approx 99$ patients per arm.

3.2 CONDITIONAL POWER

As we mentioned, power is the probability to detect the treatment difference when it exists. So if the power is low, it is unlikely to find statistically significant results. Experiments with low power should not proceed. However, it becomes more clear as the experiment proceeds.

Conditional power is the probability of correctly rejecting the null hypothesis at the end of the trial, given accumulating data. If we approximate conditional power using Brownian motion, that is at time $t = 1, B(1) > z_{\alpha/2}$ given B(t)=b.

$$B(1) = B(1) - B(t) + B(t)$$

As previously mentioned, B(1)-B(t) is independent of B(t) and have mean and variance

$$E\{B(1) - B(t)\} = \theta(1) - \theta(t) = \theta(1-t)$$

$$Cov\{B(1), B(t)\} = cov\{B(t) + B(1) - B(t), B(t)\} = var\{B(t)\}$$

$$Var\{B(1) - B(t)\} = var\{B(1)\} + var\{B(t)\} - 2cov\{B(1), B(t)\} = 1 + t - 2t = 1 - t$$

given B(t)=b,B(1)=b+B(1)-B(t) is normally distributed with variance 1-t and mean

$$E\{B(1)|B(t)=b\} = b + E[B(1) - B(t)] = b + \theta(1-t)$$

Conditional power is

$$CPower = Pr[|B(1)| \geq Z_{1-\alpha/2}|B(t)=b]$$

$$CPower = 1 - \Phi(\frac{z_{\alpha/2} - E\{B(1)|B(t)=b\}}{\sqrt{1-t}})$$

Conditional power increases as θ increases.

Under the null hypothesis $\theta = 0$,

$$E_0\{B(1)|B(t)=b\} = b$$

The empirical estimate for θ is $\hat{\theta} = B(t)/t = b/t$. Under this current trend hypothesis,

$$E_\theta\{B(1)|B(t)=b\} = b/t = \hat{\theta}$$

We will use an example with dichotomous outcome to illustrate the application of conditional power.

Suppose that at one interim analysis, 36 of 59 patients in the control arm developed events and 28 of 61 patients in the treatment arm have events. The data is summarized in table 7.

Table 7: Interim data for dichotomous response example

	Event		
	Yes	No	
Control	36	23	59
Treatment	28	33	61
	64	56	120

The information fraction $\tau = (1/59 + 1/61)^{-1}/(2/99)^{-1} = 0.61/\ \hat{p}_{treatment} = 28/61 = 0.46$,$\hat{p}_{control} = 28/61 = 0.61$, so $\hat{p} = (0.46+0.61)/2 = 0.535$ selected patient has success. The z-statistic is given by $Z_N = \frac{\sum_1^N Di}{\sqrt{v_N}}$, where $v_N = var(S_N) = 2Np(1-p)$. p is approximated by the sample proportion.

$$Z(0.61) = \frac{\sqrt{N}(\bar{p_A} - \bar{p_B})}{\sqrt{2\hat{p}(1-\hat{p})}} = \frac{\sqrt{99}(0.61 - 0.46)}{\sqrt{2(0.535)(1-0.535)}} = 2.12$$
$$B(0.61) = \sqrt{(0.61)}Z(0.61) = 1.66$$

Under the original event probaility assumptions in the two arms, the drift parameter is $\theta = 1.96 + 0.85 = 2.81$ so the condition mean of B(1) given B(0.61)=1.66 is 1.66+2.81(1-0.61)=2.76. Conditional power under the original assumptions is

$$CPower(0.61) = 1 - \Phi(\frac{1.96 - 2.76}{\sqrt{1-0.61}}) = 1 - \Phi(-1.28) = 0.90$$

Using the empirical estimates of event probabilities in the two arms corresponds to using the empirical drift parameter estimate B(t)/t=1.66/0.61=2.72.the condition mean of B(1) given B(0.61)=1.66 is 1.66+2.72(1-0.61)=2.72.

$$CPower(0.61) = 1 - \Phi(\frac{1.96 - 2.72}{\sqrt{1-0.61}}) = 1 - \Phi(-1.21) = 0.89$$

4.0 APPLICATIONS

Here, we use two phase III clinical trials to demonstrate how decisions can be made upon results of interim monitoring and conditional power.

4.1 APPLICATION IN THE NASBP B-38 TRIAL

A phase III adjuvant therapy trial was conducted to compare the two regimens of chemotherapy: Docetaxel/Doxorubicin/Cyclophosphamide (TAC) and DD AC Followed by DD Paclitaxel Plus Gemcitabine (DD AC→ PG) for women with node-positive breast cancer.

The aim of the study is to determine whether the DD AC→ PG regimen is superior to the TAC regimen in terms of Disease Free Survival (DFS) and Survival (S).

The study was conducted in women with operable, invisible carcinoma of the breast with histologically positive axillary nodes. Patients were stratified by number of positive nodes, hormone receptor status, and type of surgery and planned radiotherapy and then randomized to one of the three chemotherapy regimens. Women with ER-positive and /or PgR-positive tumors received hormonal therapy following completion of chemotherapy. Accrual of 4800 patients were completed in four years.

Five interim analyses were planned when the minimum number of events in both pairs of groups reaches 204, 286, 368, 450 and 532, respectively. A final analysis was planned when the minimum number of events in both pairs reaches 613. At each interim analyses, the hypotheses was tested: whether DD AC→ PG is superior to AC.

Three different alpha spending functions were used to determine the upper boundaries. For the Pocock-like boundaries, the alpha spending function is $\alpha^*(t) = \alpha ln\{1 + (e-1)t\}$. For linear function, the alpha spending function is $\alpha^*(t) = \alpha t$. For O'Brien-Fleming-like function is $\alpha^*(t) = 2\{1 - \Phi(z_{\alpha/2}/\sqrt{t})\}$. Details can be found in [Proschan et al., 2006].

The overall alpha level was controlled to be 0.025. Exceeding upper boundaries will lead to the conclusion that DD AC→ PG is more beneficial than TAC and the trial will be terminated for efficiency. The cumulative type I error rates and boundaries are given in Table 8 and Table 9, respectively.

Table 8: One-sided cumulative type I error rates given by different alpha spending functions and information fraction for NASBP B-38 trial

	t_1	t_2	t_3	t_4	t_5	t_6
information fraction	0.333	0.467	0.600	0.734	0.868	1.000
Pocock-like	0.011	0.015	0.018	0.020	0.023	0.025
linear-like	0.008	0.012	0.015	0.018	0.022	0.025
O'Brien-Fleming-like	<0.001	0.002	0.006	0.011	0.018	0.025

Table 9: Boundaries given by different alpha spending functions and information fraction for NASBP B-38 trial

	t_1	t_2	t_3	t_4	t_5	t_6
information fraction	0.333	0.467	0.600	0.734	0.868	1
Pocock-like	2.280	2.443	2.450	2.445	2.437	2.431
linear-like	2.395	2.489	2.451	2.404	2.359	2.317
O'Brien-Fleming-like	3.398	2.890	2.579	2.368	2.215	2.100

If we want to have a fair chance to stop the trial, we would consider using the linear alpha spending function. If we want to resist stopping too early then we would consider using O'Brien-Fleming alpha spending function. As we can see in the table, the exit probabilities in the early stage is really small.

By September 30, 2011, there were 207 deaths in DD AC $\rightarrow$ PG treatment group and 181 deaths in TAC treatment group. The information is summarized in Table 10.

Table 10: Interim data for NASBP B-38 trial

Stratum	Treatment	Total	Failed	Censored	Percent Censored
1	DD AC $\rightarrow$ PG	1630	207	1423	87.3
2	TAC	1630	181	1449	88.9
Total		3260	388	2872	88.1

We expected 613 deaths by the end of the study. The current information fraction is $t = 388/613 = 0.633$. The log hazard ratio estimate and log-rank statistic are $ln(0.889) = -0.118$ and $Z(0.633) = 1.902$. The B-value is $B(0.633) = \sqrt{0.633}(1.902) = 1.511$.
The empirical estimate of the drift parameter is $B(0.633)/0.633 = 2.389$. The expected value of B(1) given $B(0.633) = 1.511$ is $1.511 + 2.389(1 - 0.633) = 2.389$, so conditional power is

$$CPower(0.633) = 1 - \Phi(\frac{1.96 - 2.389}{\sqrt{1 - 0.633}}) = 1 - \Phi(-0.71) = 0.76$$

Under the originally assumed log hazard ratio, the drift parameter is $\theta = \sqrt{(N/4)}\delta = \sqrt{613/4}ln(1/0.75) = 3.561$, and the expected B-value at the end of the trial is $1.511 + 3.561(1 - 0.633) = 2.818$. Conditional power is

$$CPower(0.633) = 1 - \Phi(\frac{1.96 - 3.561}{\sqrt{1 - 0.633}}) = 1 - \Phi(-2.643) = 0.99$$

The conditional power under both assumptions are relatively high, indicating that there is high chance of getting a significant result at the end of the trial.

4.2 APPLICATION IN THE NASBP B-40 TRIAL

A randomized phase III trial of neoadjuvant therapy was conducted to determine whether the addition of bevacizumab to docetaxel-based regimens followed by AC will increase the pathologic complete response (pCR) rates in patients with palpable and operable HER2-negative breast cancer.

Three statistical analyses were planned for the outcome of pCR in breast. Two interim analyses were planned when the available pathologic response status reaches 400 and 800. The trial would be considered to assign bevacizumab to all subsequently enrolled patients if it has enough evidence that docetaxel based regimen with becacizumab is superior than

without becacizumab. As in the NASBP B-38 study, three alpha spending functions are used to determine the boundaries and the final significance level is adjusted to be 0.05. Exceeding the boundaries will lead to the conclusion that with bevacizumab is more beneficial. Boundaries and corresponding type I error rates are given in Table 11 and Table 12, respectively.

Table 11: Two-sided cumulative type I error rates given by different alpha spending functions and information fraction for NASBP B-40 trial

	t_1	t_2	t_3
infromation fraction	0.333	0.667	1.000
Pocock-like	0.023	0.038	0.050
linear-like	0.017	0.033	0.050
O'Brien-Fleming-like	0.002	0.022	0.051

Table 12: Bundaries given by different alpha spending functions and information fraction for NASBP-40

	t_1	t_2	t_3
infromation fraction	0.333	0.667	1.000
Pocock-like	2.002	2.001	1.975
linear-like	2.128	2.001	1.880
O'Brien-Fleming-like	2.841	2.026	1.728

The sample size for the trial is 1200 patients (600 per arm). By February 23, 2010, pCR status was available from 806 patients with 401 assigned to the arms with Bevacizumab. Table 9 gives the proportion of patients with events in both arms. The proportions of pCR were 30.4%(123/405) and 34.7%(139/401) for the two groups, respectively.

Table 13: Interim data for NASBP B-40 trial

	pCR		
	Yes	No	
Bevacizumab	139	278	401
non-Bevacizumab	123	266	405
	262	544	806

The information fraction $t = (1/401 + 1/405)^{-1}/(2/600)^{-1} = 0.672$.

For the Bevacizumab group, the estimated pCR at that time is $\hat{p}_{Bev} = 139/401 = 0.347$, For the non-Bevacizumab group, the estimated pCR at that time is $\hat{p}_{non} = 123/405 = 0.304$, so $\hat{p} = (0.304 + 0.347)/2 = 0.326$ selected patient has success. The z-statistic is given by $Z_N = \frac{\sum_1^N Di}{\sqrt{v_N}}$, where $v_N = var(S_N) = 2Np(1-p)$. p is approximated by the sample proportion.

$$Z(0.672) = \frac{\sqrt{N}(\bar{p_A} - \bar{p_B})}{\sqrt{2\hat{p}(1-\hat{p})}} = \frac{\sqrt{600}(0.347 - 0.304)}{\sqrt{2(0.326)(1-0.326)}} = 1.123$$

$$B(0.672) = \sqrt{(}0.672)Z(0.672) = 0.921$$

Under the original event probaility assumptions in the two arms, the drift parameter is $\theta = 1.96 + 0.85 = 2.81$, so the condition mean of B(1) given B(0.672)=0.921 is 0.921+2.81(1-

0.672)=1.843. Conditional power under the original assumptions is

$$CPower(0.672) = 1 - \Phi(\frac{1.96 - 1.843}{\sqrt{1 - 0.672}}) = 1 - \Phi(0.21) = 0.417$$

The conditional power under the original assumption is relatively low. There is fairly high probability of failing to reach a statistically significant result at the end of the trial.

Using the empirical estimates of event probabilities in the two arms corresponds to using the empirical drift parameter estimate B(t)/t=0.921/0.672=1.37.the condition mean of B(1) given B(0.671)=0.921 is 0.921+1.37(1-0.672)=1.37.

$$CPower(0.672) = 1 - \Phi(\frac{1.96 - 1.37}{\sqrt{1 - 0.672}}) = 1 - \Phi(1.03) = 0.152$$

Thus, if the empirical trend is true, then we have 15.2 percent chance of a statistically significant benefit at the end of the trial.

When the test statistics exceed the upper boundaries in interim looks, we would consider terminating the trial for superiority. In this case, the new treatment under investigation can be put into market as soon as possible. However, if a test statistic is far away from the boundaries in the interim looks, we will be concerned whether we could obtain a significant result at the end of the trial. In this situation, we would calculate the conditional power. And based on the conditional power, decisions might be made to stop the trial if it is unlikely to claim the superiority of the new treatment in the end so that further resources would not be wasted. However, this kind of decision is difficult when so many efforts have been put into the trial. Therefore, investigators usually also consider other aspects of the trial such as economics, toxicity of the drugs while making a decision on the trial.

BIBLIOGRAPHY

P Armitage. *Sequential medical trials*. Oxford:Blackwell, 1975.

P Armitage, CK McPherson, and BC Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, pages 235–244, 1969.

C Jennison and BW Turnbull. *Group sequential methods with applications to clinical trials*. Boca Raton, FL:CRC Press, 1999.

KK Lan and DM Zucker. Sequential monitoring of clinical trials: the role of information and brownian motion. *Statistics in Medicine*, 12(8):753–765, 1993.

D Moher, CS Dulberg, and GA Wells. Statistical power, sample size, and their reporting in randomized controlled trials. *the Journal of American Medical Association*, 272(2): 122–124, 1994.

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

MA Proschan, KKG Lan, and JT Wittes. *Statistical monitoring of clinical trials: a unified approach*. New York: Springer Science & Business Media, 2006.

DA Schoenfeld. A simple algorithm for designing group sequential clinical trials. *Biometrics*, 57(3):972–974, 2001.