1-1-2009

# Statistical Approaches for Adding or Switching Hypotheses in Multi-armed Clinical Trials

Jordan Jaskwhich Elm
*Medical University of South Carolina*

# Statistical Approaches for Adding or Switching Hypotheses in Multi-armed Clinical Trials

by
Jordan Jaskwhich Elm
A dissertation submitted to the faculty of the Medical University of South Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Graduate Studies.

Department of Biostatistics, Bioinformatics, Epidemiology
2009

Approved by:

Chairman, Advisory Committee

Yuko Y. Palesch, PhD

Barbara C Tilley

Barbara C. Tilley, PhD

Bernard Ravina, MD, MSCE

Wenle Zhao

Wenle Zhao, PhD

Vanessa Hinson, MD

## TABLE OF CONTENTS

# ACKNOWLEDGEMENTS

I would like to thank my dissertation committee for their support and helpful suggestions. I am greatly indebted to Dr. Barbara Tilley for initially encouraging me to complete a PhD in Biostatistics and giving me the opportunity to complete this work. Without her support, this would not have been possible. I am greatly indebted to Dr. Yuko Palesch for giving me the ideas for this work and for her encouragement and support to finish this in a timely manner. I am greatly indebted to my family for their support and appreciation of the value of education.

This work is dedicated to the memory of my Grandfather, Jordan M. Sheperd, and my Mother, Cynthia S. Jaskwhich.

# ABSTRACT

JORDAN JASKWHICH ELM. Statistical Approaches for Adding or Switching Hypotheses in Multi-armed Clinical Trials. (Under the direction of YUKO Y PALESCH, PhD).

**Background:** As treatments become ready for testing at staggered times, it is desirable to have clinical trials that can accommodate different entry and exit times without prematurely discarding potentially efficacious treatments. In a group sequential multi-armed clinical trial, one treatment arm could be found to be superior or inferior at an interim analysis while the remaining arm is inconclusive. Existing methods address dropping an inferior arm from further study[1,2], but do not fully address the handling of an early finding of overwhelming superiority of one arm (*scenario 1*). We consider an approach to transition from a multi-armed superiority trial to a two-armed non-inferiority trial after superiority for a single arm has been determined. Additionally, the literature does not address statistical methods for adding another treatment arm into an ongoing trial (*scenario 2*). **Methods:** For these novel scenarios, potential adaptive and non-adaptive analytical approaches for pairwise comparisons of a difference in means in independent normal populations are proposed with emphasis on controlling the type I error rate strongly. Statistical operating characteristics are compared via Monte Carlo simulation. An example is given using Parkinson's disease data (NET-PD FS1 and FS-TOO). **Results:** For *scenario 1*, all methods performed similarly, but power was highest when using an inverse chi-square adaptive test. For *scenario 2*, in the presence of a cohort effect, when data were pooled from before/after the design change, the type I error rate was inflated and power was reduced. The alternative approaches given were more powerful and controlled the type I error rate. **Conclusions:** When two treatment arms are equally efficacious, it is likely that one, but not the other, will be found efficacious at an interim analysis. When transitioning into a non-inferiority trial, the adaptive methods allow for a reduction in total sample size with increased power for testing non-inferiority (compared to a non-adaptive approach). Both adaptive and non-adaptive analytical methods are possible when a new treatment arm is added mid-study. When these methods are applied to real Parkinson's disease trial data, the conclusions support the primary trial findings.

# CHAPTER 1 INTRODUCTION

## 1.1 Clinical Motivation

Parkinson's disease (PD) is an age-related, neurodegenerative disease affecting 1% of the population older than 60 years of age and is thought to involve a loss of dopamine neurons predominantly in the substantia nigra pars compacta[3,4]. The National Institute of Neurological Disorders and Stroke (NINDS) sponsored a Committee to Identify Neuroprotective Agents in Parkinson's (CINAPS) to prioritize drugs that are worthy of testing for PD[5]. These agents are at different stages of development. Some require first in-human Phase I evaluation; some require additional safety and putative efficacy assessment in Phase II testing, while others are ready for Phase III testing. In 2006, the NINDS Exploratory Trials in Parkinson's Disease (NET-PD) initiative recently completed two Phase II clinical trials of four agents with potential for slowing clinical decline[6,7]. While three of the four drugs were considered futile to move forward into a phase III trial, one drug (creatine) was found to be worthy of further evaluation. The NET-PD group is now enrolling a large, long-term Phase III trial of creatine.

Due to the nature of PD, it is not possible, using current technology, to evaluate a definitive benefit to the PD rate of progression in a short-term study. When patients are receiving symptomatic treatment for PD (which is usually introduced within a year or less after diagnosis), then five or more years of follow-up is necessary to evaluate a treatment benefit using the clinical outcome measures since validated biomarkers or neuroimaging approaches are not available. The clinical measures currently in use are insensitive to short-term treatment effects (less than 1 year) and have considerable variability[8]. Obviously, the need for long-term follow-up and large sample sizes for clinical trials in PD impacts administrative, logistical, and financial resources. The NET-PD investigators continue to pursue Phase II trials of additional agents, as the agents are identified by CINAPS and from other sources, such as pharmaceutical industry sponsored Phase I or Phase II trials. Information on new agents is available in a staggered fashion. Therefore, a Phase III study design that can accommodate adding and dropping arms at different times is desirable for efficient use of limited resources. This dissertation considers the statistical operating characteristics of potential analytical methods when such flexibility of design is necessary.

While there are many symptomatic treatments of PD, there are no existing treatments with a known neuroprotective benefit. As such, clinicians are interested in identifying any drug that slows clinical decline. Therefore, the main interest in multi-arm studies in this context is not to identify the best drug, but rather, to compare each drug to placebo. This distinction has design implications. As clinical trials are conducted, if a drug is identified that slows clinical decline, then it would become the gold standard. Future trials would compare possible agents to this standard/active control in a non-inferiority design, since it would be unethical to withhold an available treatment that can slow clinical progression. Ongoing trials might need to be re-designed.

### 1.2 Introduction to the problem

The PHRMA white papers define an adaptive design as a clinical study design using accumulating data to decide to modify the study mid-course, while maintaining validity and integrity[9]. They describe adaptation as a "design feature" to enhance the trial, not an *ad hoc* decision to fix poor planning[9]. Indeed, the US FDA viewpoint is that for a clinical trial to be conducted with an adaptive design, it should be "prospectively planned", that is, all possible design changes need to be pre-specified in the study protocol[10,11].

Chow and Chang discuss prospective versus concurrent adaptive designs. Certain adaptive designs lend themselves to prospective planning such as adaptive-randomization, interim analyses, and sample size re-estimation. In contrast, during the course of the study it may become obvious that other designs changes are needed such as: changes to the inclusion or exclusion criteria, change in treatment duration, change in hypothesis, or endpoints. These concurrent design changes necessitate revisions to the protocol and the statistical analysis plan made prior to unblinding, given approval by regulatory agencies.[12] In the latter case, statistical approaches to maintain proper control of the type I error rate based on adaptive combination tests are still of value. Thus, while it would be ideal to anticipate and pre-specify any design change to be made mid-course, in reality, that is not always possible. The penalty for making concurrent design changes is the loss of integrity and persuasiveness of the results obtained, in particular after substantial changes are made[13,14].

In this dissertation, we consider two scenarios that involve a problem with the placebo group comparison. In *scenario one*, consider an ongoing 3-armed clinical trial of two active treatment arms (A and B) and one placebo arm where the primary comparison is between each of the active treatment arms against the placebo arm (i.e. many-to-one (MTO) comparisons, and therefore, not all pairwise comparisons are of interest). Suppose one active treatment arm is found to be overwhelmingly efficacious relative to the placebo arm by a pre-specified interim analysis criterion prior to the end of enrollment. The other treatment arm is not. It may be unethical to continue to enroll subjects to placebo. Then, clearly, the remaining active treatment arm cannot be compared against placebo at a later time point, only against the other active treatment arm (now the active control) in a non-inferiority hypothesis. Now the primary hypothesis has changed from one of superiority to one of non-inferiority and the reference group has changed (from placebo to the standard). The issue of blinding will be discussed.

In *scenario two*, a new treatment arm (B) is added to an ongoing clinical trial (of A versus placebo). The rationale for adding a treatment arm to an ongoing clinical trial is because multi-armed clinical trials reduce the number of patients randomized to placebo as compared to two separate trials. However, when introducing a new treatment arm, the placebo group will be different before (stage 1) and after (stage 2) the design change. The subjects randomized to placebo in the first stage receive only the placebo for treatment A, while the subjects randomized to the placebo group during the second stage receive placebo for treatment A and treatment B. The placebo groups from the two stages may not be amenable to aggregation, because of overt differences. For example, there may be a cohort effect or there may be increased enthusiasm about the new treatment arm that could bias the results if the placebos are pooled[15].

This dissertation considers statistical approaches of two particular design changes: 1. dropping the placebo arm from a 3-armed superiority trial when early efficacy is determined for one treatment and transitioning it to a 2-armed non-inferiority trial, and 2. adding an arm to an ongoing clinical trial. The specific aims of this research are:

Aim 1. To propose and compare adaptive and non-adaptive analytical methods that allow for transitioning from a 3-armed superiority trial (e.g., active treatment A, active treatment B, and placebo) to a 2-

armed non-inferiority trial when efficacy of one treatment arm (A or B) relative to the placebo arm is determined at an interim analysis. The power of all methods and the average sample number for the adaptive methods (when sample size is re-estimated) will be compared under Monte Carlo simulation.

Aim 2. To propose and compare adaptive and non-adaptive analytical methods for the scenario of adding a treatment arm (B) to an ongoing clinical trial of 2 arms (A versus placebo). Both the single-stage and group sequential designs will be considered. The Types I and II error probabilities of each analysis method will be compared under Monte Carlo simulation.

Aim 3. To apply the approaches presented in Aim 2 to the analysis of NET-PD FS1 and FS-TOO clinical trials in PD patients. For the illustration, these concurrently conducted trials were re-analyzed as one study, and the testing methods were compared under bootstrapping.

# CHAPTER 2. BACKGROUND

Given the aims listed in Chapter 1, we provide here a comprehensive review of relevant statistical methodology. The following are statistical methods are reviewed in the context of multi-armed clinical trials: multiple comparison procedures, group sequential methods, and methods for design change (adaptive methods).

## 2.1 Multiple Comparison Procedures (MCPs) for Multi-Armed Trials

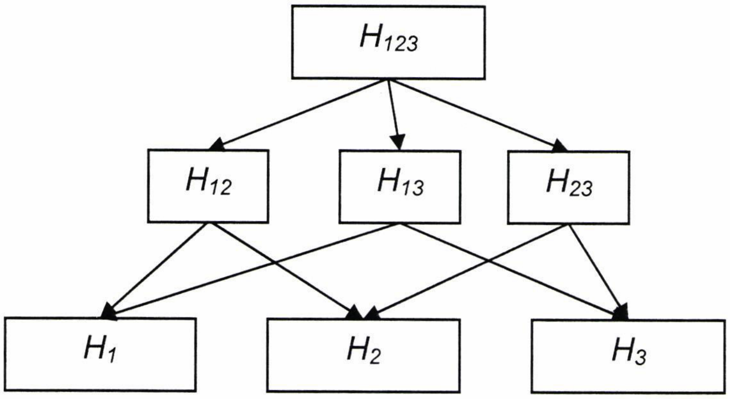
Multiplicity refers to the inflation of the pre-specified type I error rate due to multiple testing. Sources of multiplicity include multiple tests for more than one endpoint, more than one treatment group, and repeated analyses of accumulating data (interim analyses). All three aims of this dissertation involve multi-armed clinical trials, thus, we will review existing methods to control multiplicity for the primary endpoint. The familywise/experimentwise error rate refers to the probability of rejecting any true hypothesis from a family of hypotheses. In most studies, particularly confirmatory clinical trials, investigators wish to constrain the familywise error rate (FWE) at some pre-specified alpha (e.g. 0.05 two sided, 0.025 one sided, or 0.10 for a phase I or II trial). *Weak* control of alpha means that the familywise error rate is no more than the specified overall alpha under the complete null configuration, but not all other configurations. *Strong* control of alpha means that the familywise error rate is no more than the specified overall alpha under all configurations of any component null hypothesis.

Most confirmatory multi-armed clinical trials require *strong* control of alpha, although it has been argued that this is an unfair penalty for undertaking to test more than one drug at the same time[16]. Thus, depending on the situation, sometimes only the per-comparison error rate need be constrained[16].

In these aims, we will consider the case in which it is desirable to control the familywise error rate strongly for the primary outcome measure. There are many approaches to adjusting for multiplicity, or multiple comparison procedures (MCPs). Some approaches adjust the p-value without regard for the distribution from which it came, such as the *single-step* methods (Bonferroni, Šidák[17]), *sequentially rejective/stepwise* methods (Bonferroni-Holms[18], Šidák-Holms, Simes' Modified Bonferroni[19], Hochberg[20], Hommel[21], Rom[22]) and *closed testing* procedures[23]. If independence of the multiple tests is assumed (as in the Bonferroni type methods) when dependence actually exists, these approaches will be conservative (more difficult to reject), and power will be reduced[24]. While single-step MCPs can be used to test hypotheses and accurately compute simultaneous confidence intervals (CIs), sequentially rejective/stepwise methods and closed testing methods can only be used to test hypotheses.

The closed testing procedure  (closure principle) provides a powerful method for strong control of the type I error rate for ordered hypotheses[23]. The closed testing procedure involves forming the set of all possible intersection hypotheses among the $H_i$ treatment versus placebo comparisons (the closure of the set). Each intersection hypothesis is tested using an appropriate alpha-level test, where the appropriate test is chosen based on the type of data and number of groups to be compared. Any hypothesis $H_i$ can be rejected with strong control of the FWE when: (1) the test of $H_i$ is statistically significant; and (2) the test of every intersection hypothesis that includes $H_i$ is statistically significant.  Figure 2-1 gives an example of a closed testing scheme where there are three treatment arms and a control. Denoting the control group mean by $\mu_0$, the three pairwise hypotheses of interest are $H_1 : \mu_1 = \mu_0$, $H_2 : \mu_2 = \mu_0$, and $H_3 : \mu_3 = \mu_0$. We can reject the null hypothesis for $H_1$ at an alpha level test if $H_{123}$, $H_{12}$, and $H_{13}$ can also be rejected.

**Figure 2-1. Closed Testing Procedure for a Trial with Three Treatment versus Placebo Comparisons (Many-To-One)[23]**



MCPs with distributional assumptions are useful for ANOVA and ANCOVA models. Available methods differ for the balanced or unbalanced case. In the case of one-way ANOVA, for dependent estimates, we can compute simultaneous confidence intervals for all pairwise (AP) comparisons with the Tukey method (based on the Studentized range distribution), or compare many-to-one (MTO) treatments to control with Dunnet's method (based on the range distribution)[25]. In general, these methods are slightly less powerful than the closed testing procedure[25]. For comparison purposes, in these aims, we compared analytic methods using Simes' method[19] and the Closed principle for each case, rather than applying MCPs specific for ANOVA, restricting our attention to MTO comparisons.

## 2.2 Group Sequential Methods for Multi-armed Trials

In Aims 1 and 2, the implications for design change in the group sequential scenario are considered. When repeated tests are conducted on accumulating data, the overall type I error rate will be inflated[26]. Inflation from this kind of multiplicity (due to interim looks at the data) can be constrained by applying group sequential test procedures (e.g. O'Brien-Fleming[27] or Pocock[28] bounds) or an alpha spending function[29]. In the *classical group sequential* framework, the type I error rate is corrected for multiple looks, but one cannot change design parameters (e.g. re-estimate the sample size) based on these looks.

Classical group sequential methods assume that the number of looks is pre-specified and are equally spaced. In the O'Brien-Fleming[30] method, it is more difficult to stop early and the critical value for the final analysis is close to the critical value of that needed if no interim looks had been performed. The null hypothesis is rejected when $|Z_k| \leq C_B\sqrt{K/k}$ where *k* is the *k*th interim analysis of *K* planned interim analyses, $Z_k$ is the standardized test statistic using all the data up to the *k*th look, and $C_B$ is a constant based on the desired type I error probability $\alpha$ and *K*. Another popular choice is Pocock's test. These boundaries are constant, where the null hypothesis is rejected when $|Z_k| \leq C_P$ where $C_P$ depends on $\alpha$ and *K.* Pocock's method is more conservative at the final analysis. Alpha spending functions[29] are more flexible in that the number of interim looks need not be pre-specified  or equally spaced.

Methods for multi-armed group sequential clinical trials are still under development. Jennison and Turnbull[31] provide a review of existing methodology. Simply, one could monitor the study by testing the global hypothesis (e.g. F-test, chi-squared test). If the global test is significant, MCPs can be applied to test each arm with strong control of alpha[32,33]. Tang and Geller[33] give an approach for a clinical trial with multiple endpoints and group sequential monitoring for early stopping (for overwhelming efficacy). By use of closed testing methods and a global test, single endpoints are tested in a step-down procedure. They show how this approach can be modified if there are multiple treatment arms and a single endpoint.

When several pairwise hypotheses (not just a global hypothesis) are of interest, a simple approach is to apply Bonferroni adjustments to pre-specified alpha and then to find the corresponding O'Brien-Fleming or Pocock type critical values such that each individual hypothesis can be tested at any stage[1]. Follmann, Proschan, Geller[1] recommend requiring strong control of alpha and equal amount of evidence for each treatment arm. There are existing methods for the case in which an arm is found to be inferior at an early interim look, in which case, it would be dropped from further study. Follman *et al.* derive exact critical values for each interim analysis and show that when one arm is dropped for inferiority[1], then the remaining looks would use a critical value that corresponds to the reduced number of arms. This sequentially rejective procedure is less conservative than the Bonferroni approach, and this result holds for up to 4 arms in a trial.

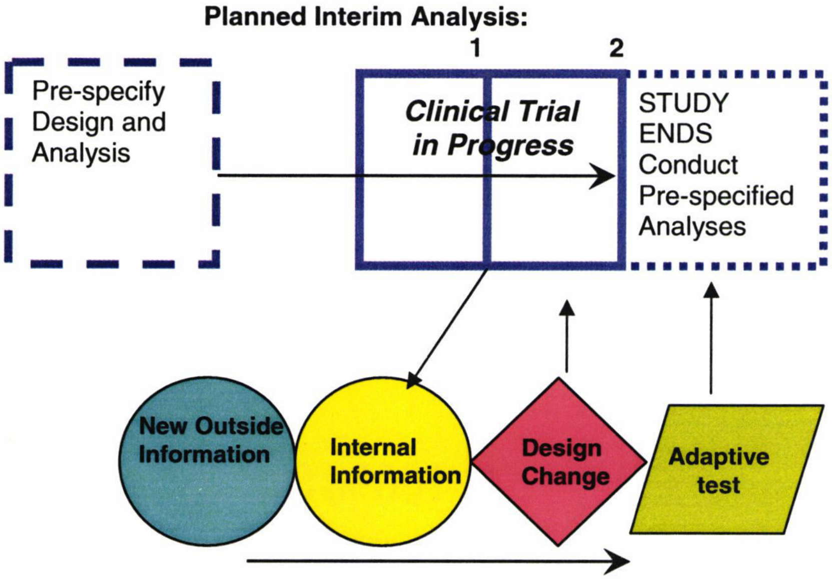
## 2.3 Mid-Course Design Change

In principle, the key features of a clinical trial design (e.g., hypotheses, primary outcome, treatment, target population) should not change once the study has begun. In reality, not all parameters are known when planning a trial, and if the trial is planned with incorrect assumptions about these parameters, the chance of success can be diminished. It is possible that the variance estimate used for sample size calculation is an underestimate of the variance actually observed. There are existing approaches to re-estimate the variance mid-study (without estimating the effect size and without unblinding), and this re-estimated variance can be used to increase the sample size without inflating the alpha[34,35].

Recently much work has been done on adaptive testing methods which, in theory, allow one to modify virtually any design parameter. Adaptive designs began to be developed in the 1990s, but have only recently come into practical use. In 2006, the FDA (US Food and Drug Administration) announced their position on adaptive designs, recognizing it as a useful design tool and their plans to develop a guidance document. In the past decade the number of published adaptive clinical applications has been increasing, primarily from Germany[36], and there have been many adaptive design submissions to the FDA[10]. At this time, the appropriate use of adaptive designs remains a topic of debate. A special issue of *Biometrical Journal* includes an extensive review of the pros and cons of adaptive designs[37].

In general, adaptations driven by internal information are more controversial than those driven by external information. See Figure 2-2 for an overview. The basic idea is that if a mid-course design change is made (based either internal or external data), then the penalty is the use of non-standard test statistics which are combination tests of stage-wise p-values (not sufficient statistics)[36]. There are two commonly used adaptive tests[36]: (1) Inverse Chi-square (Fisher's) combination test[38,39] and (2) Weighted inverse normal combination test[40-42].

**Figure 2-2. How Adaptive Testing Methods can be applied**


Planned Interim Analysis:
1
2
Pre-specify Design and Analysis
Clinical Trial in Progress
STUDY ENDS Conduct Pre-specified Analyses
New Outside Information
Internal Information
Design Change
Adaptive test

*Adaptation based on Internal information*

When examined at an interim look, information within the study (internal information) could prompt a change in the design. For example, when a planned interim look suggests that the effect size is smaller than expected, then clinical investigators may want to increase the sample size to detect this smaller effect size. However, when the design change is made based on observed data, then the type I error for the final (pre-specified) analysis will be inflated. By applying adaptive methods (e.g. a combination test) then, the pre-specified alpha can be constrained.

*Adaptation based on External information*

When external information (independent from the study) prompts a design change for a study that has also a planned interim look, then the integrity of the study may be questioned because it is difficult to prove that internal information did not play a role in the design change. Thus, adaptive methods should also be considered for this situation. An example of an adaptation based on external information (outside the study) is the change in primary endpoint based on the development/validation of a new instrument. The scenario of AIM 2, adding a new treatment arm, is also an example of a design change based on external information.

## 2.4 Basic Adaptive Design Principle

Let $H_0$ be the null hypothesis of interest to be tested in two stages. Then $p_1$ is the p-value from the first stage and $p_2$ is the p-value from the second separate stage. We will use the terminology Stage 1 to refer to the time period prior to the design change and Stage 2 to refer to the time period after the design change. Then the decision for the final analysis is based on a combination function $C(p_1, p_2)$, which is continuous and monotonically increasing. If $C(p_1, p_2) \leq c$ then $H_0$ is rejected. We assume that the p-values $p_1$ and $p_2$ are independent and uniformly distributed on [0, 1]. Any 2-stage combination test can also be defined by a conditional type I error function[43,44], $\widetilde{A}(p_1)$.

### 2.4.1 Conditional Type I Error method

In 1995 Proschan and Hunsberger[43] gave an approach to extend a study beyond the planned termination time based on conditional power when the final test is not statistically significant. The overall type I error is preserved as follows. Let $\tilde{A}(p_1)$ denote the *conditional error function:* the conditional probability of rejecting the null hypothesis, given $p_1$, the p-value observed at stage 1. We reject the null hypothesis for sufficiently low values of $p_2$, given $p_1$. Then $\alpha_{\max} = \int_0^1 \tilde{A}(p_1)dp_1$. The second stage is essentially a second study carried out at a significance level $\tilde{A}(p_1)$.

Proschan and Hunsberger introduce the circular conditional type I error function, $\tilde{A}_{cir}(p_1)$. Since the overall type I error rate is $\alpha$, then

$$\alpha_{\max} = \int_0^1 \tilde{A}(p_1)dp_1 \text{ where}$$

$$\tilde{A}_{cir}(p_1) = \begin{cases} 0 & \text{if } p_1 \ge \alpha_0 \\ 1-\Phi(\sqrt{c_{PH}{}^2 - (\Phi^{-1}(1-p_1))^2} & \text{if } 1-\Phi(c_{PH}) < p_1 < \alpha_0 \\ 1 & \text{if } p_1 \le 1-\Phi(c_{PH}) \end{cases}$$

where $\alpha_0$ is the critical lower limit for early stopping in favor of the null and $\Phi(\bullet)$ denotes the cumulative normal distribution function (cdf). Then we choose $\alpha_0$ and $c_{PH}$ is determined so that $\int_0^1 \tilde{A}_{cir}(p_1)dp_1 = \alpha$.

A *conditional error function* can be used to solve simultaneously for an adjusted critical value and the additional sample size such that the type I error rate is constrained. This method can be used to make adjustments to the sample size (e.g. based on variance estimate or effect size), assuming that the same outcome measure will be used.

Later work shows how this approach can be thought of a combination rule for the p-values (or test statistics) from the 2 stages. As stated before, any 2-stage combination test can also be defined by a conditional type I error function[44], $\tilde{A}(p_1)$. For example, the Prochan and Hunsberger conditional type I error function can be thought of as a type of combination function[45] $C(p_1, p_2) = (\Phi^{-1}(1-p_1))^2 + (\Phi^{-1}(1-p_2))^2$, where $C(p_1, p_2)$ is a function of the p-values from stages 1 and 2. Extending the work of Proschan & Hunsberger[43], Denne[46] uses conditional power to re-estimate the sample size (while maintaining the type I error) in the framework of an error-spending group sequential analysis.

### 2.4.2 Inverse Chi-square (Fisher's) combination test

Bauer and Köhne suggest a two stage combination test that maintains the type I error rate by combining p-values from the interim and final analysis. This test is based on Fisher's criterion which uses the product of the stochastically independent p-values[39]. We reject the null hypothesis if

$$p_1 p_2 \le c_\alpha = \exp[-\frac{1}{2}\chi_4^2(1-\alpha)]$$

where $p_1$ is the p-value from stage 1, $p_2$ is the p-value from stage 2, and $\chi_4^2(1-\alpha)$ is the $(1-\alpha)$-quantile of the central $\chi^2$ distribution with 4 degrees of freedom. This is based on the result that the natural log of the product of *k* random variables distributed uniformly [0,1] form a chi-square distribution with 2**k* degrees of freedom[38]. Then $c_\alpha$ is the critical value for the combination test at the end of the trial.

This method assumes that under the null hypothesis the p-values of the tests from the two stages are from stochastically independent samples and are uniformly distributed [0,1]. Thus, unlike in group sequential methods, data from the second stage are not cumulative (i.e. do not include the data from stage 1).

This approach can be modified to allow the study to stop after stage 1 (with failure to reject the null hypothesis) when little or no effect has been observed[39]. We can allow for early stopping for lack of effect by letting $\alpha_1$ $(c_\alpha \le \alpha_1 \le \alpha)$ be the critical upper limit and $\alpha_0$ $(\alpha \le \alpha_0 \le 1)$ be the critical lower limit. Then, if $p_1 \le \alpha_1$, the interim analysis stops to reject the null hypothesis. Likewise, if $p_1 \ge \alpha_0$ then the trial stops early, with failure to reject the null hypothesis. The corresponding conditional type I error function is

$\tilde{A}(p_1) = \begin{cases} 1, & p_1 \leq \alpha_1 \\ c_\alpha / p_1, & \alpha_0 < p_1 < \alpha_1 \\ 0, & p_1 \geq \alpha_0 \end{cases}$ so in order to constrain the overall alpha to $\alpha$, we solve the following for $c_\alpha$

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \frac{c_\alpha}{p_1} dp_1 = \alpha$$

which is $c_\alpha = (\alpha - \alpha_1) / \ln(\alpha_0 / \alpha_1)$.

One advantage of the Bauer and Kohne[39] adaptive combination method is that it is not restricted to the normal case, but can be applied to any continuous test statistic. In addition, the test statistics used for stage 1 and stage 2 can be different. The loss of power for the adaptive methodology in the normal case has been shown by Bauer and Kohne[39] to be small compared to the classical test statistic, however, the loss of power with other test statistics (e.g., non-normal) has not been explored.

### 2.4.3 Weighted Inverse Normal Method

The weighted inverse normal combination test was developed by Mosteller and Bush[40]. Let $Z_k = \Phi^{-1}(1 - p_k)$ be the normal Z-value at the $k^{th}$ time point. Then the weighted inverse normal combination test rejects the null hypothesis if $w_1 Z_1 + ... + w_k Z_k > \Phi^{-1}(1-\alpha)$ where $\sum_{k=1}^{K} w_k^2 = 1$. For example, the weights can be $1/\sqrt{2}$.

More recently, Cui *et al.*[41], Shen and Fisher[47], and Lehmacher and Wassmer[42] developed adaptive designs based on the weighted inverse normal method. In Cui *et al.*[41] and Lehmacher and Wassmer[42], this is achieved by weighting the test-statistic with the square root of the original sample size, instead of with the square root of the inflated sample size. Specifically, this is accomplished by decomposing the test statistic into the sum of treatment differences up to the time of sample size adjustment and the sum after the adjustment is made up to the $k^{th}$ observation, then weighting these by the square root of the original sample size. If there is no increase in the sample size, this is exactly equivalent to group sequential methods[41]. The classical group sequential stopping boundaries are still applied[41].

This combination test is a function of the sufficient statistic which is a desirable property. This approach was developed for *k* stages and can also be thought of as a combination rule where the p-values from the *k* disjoint stages are combined into a test statistic as

$$C(p_1,\ldots,p_k) = \sum_{k=1}^{K} w_1 \cdot \Phi^{-1}(1-p_k)$$

where $\sum_{k=1}^{K} w_k^2 = 1$ and $w_k = \left(\sum_{k=1}^{K} n_k\right)^{-1/2} \cdot \sqrt{n_k}$ where $n_k$ is the total number of observations at the $k^{th}$ stage. It can be shown that under certain conditions, this method is identical to the approach of Shen and Fisher[47]. The corresponding conditional type I error function can be shown which corresponds to the classical group sequential design.[44]

### 2.4.4 Extensions of Adaptive Combination Tests

Kieser, Bauer, Lehmacher combine the adaptive methodology with the closed testing procedure to control multiplicity from *a priori* ordered hypotheses.[48] Others have shown how adaptive combination methodology can be applied for modifying hypotheses mid-study, in particular, for the case of changing the *a priori* ordering of hypothesis or adding a new hypothesis[49]. The adaptive methodology has been applied to multi-armed studies, in particular for the case of eliminating treatments (compared to control) at stage 1[50].

Others have extended the Bauer and Kohne[39] methodology to allow Lan-Demets[51] alpha-spending functions to be applied in this framework[52,53]. Posch, Bauer, Brannath[54] discuss general adaptive combination tests and optimal sample size re-estimation techniques. As shown above, they showed how inverse normal adaptive tests (e.g. Cui *et al.*[41], Shen Fisher[47]) can be thought of as a combination test of p-values.

Hommel *et al.* consider the impact of correlated p-values (bivariate normal test statistics) from different stages on the type I error rate in an adaptive clinical trial[55]. They show that the Simes' test can be modified to allow for early stopping for futility. This Modified Simes' test (MST) is another type of adaptive 2-stage test where: $\{p_1 \le \alpha_1\} \cup \{\alpha_1 < p_1 \le \alpha_0 \cap p_2 \le \alpha^*\}$ where $0 \le \alpha_1 < \alpha_0 \le 1$ and $0 < \alpha^* < 1$. In comparison with Fisher's combination test, the Modified Simes' test (MST) achieves similar power with smaller maximum sample sizes.

### 2.4.5 Recursive Combination tests

There is a more general method of adaptation called recursive combination tests which are based on the conditional type I error function given by Proschan and Hunsberger[43,56]. Recursive combination tests are a method of evaluating *k*-stages through a series of two-stage adaptive tests without specifying an error spending function[57,p.153]. Two-stage adaptive designs and classical group sequential procedures can be thought of as special cases[56]. Branath, Posch, Bauer show how this methodology defines a p-value function and can accommodate computations of confidence intervals[56]. Müller and Schafer[58] use a Brownian motion model to create an adaptive method within the classical group sequential framework. Their method allows for the sample size to be changed, interim analyses for early stopping to be added, the number of interim analyses to be increased/reduced, change of error spending function, change of test statistic, or change of outcome measure, while protecting the type I error rate.

## 2.5 Change of Hypothesis (Non-Inferiority/Superiority)

Adaptive designs have been touted as advantageous because it is possible to change virtually any feature of the design including the hypothesis and still maintain adequate control of the type I error rate. In reality, changing a hypothesis mid-study is not generally acceptable regardless of the type I error control, as it suggests a fatal flaw in the design. In the US, the regulatory viewpoint is that for a clinical trial to be conducted with an adaptive design, it should be planned prospectively, that is, all possible design changes need to be pre-specified in the study protocol[10,11]. However, there are times when changing to a related hypothesis, e.g. from superiority to non-inferiority, could be an option. Such a scenario is addressed in Aim 1. In the context of non-adaptive designs, Denne and Koch have shown that it is possible to test both non-inferiority and superiority sequentially without adjusting for multiplicity because they are nested hypotheses; closed testing methods are applied[23,46]. Also in the setting of a group sequential active controlled clinical trial, Wang *et al* give an adaptive design and show how the trial can begin with the non-inferiority hypothesis and proceed to test superiority, or vice versa[59]. Koyama describes a flexible two-stage adaptive design that can control both the type I and type II errors for these two simultaneous objectives, non-inferiority and superiority[60]. Kropf, Hommel, and Schmidt give a procedure for testing superiority of two treatments compared to standard/placebo and then test for non-inferiority of the two treatments in the context of a two-stage adaptive clinical trial, but they do not consider the need to stop the placebo arm after the first stage or the ethics involved[61].

### 2.6 Significance of this research

When a multi-armed clinical trial of several treatments to placebo is conducted and interim analyses are planned for the opportunity of early stopping in favor of the alternative for one or more arms, then the possibility arises that one treatment arm will be found to be superior or inferior at an early look while the remaining arms will be inconclusive. As described above, there are methods for the case in which an arm is found to be inferior at an interim look, and it is dropped from further study[1,2]. This is rather straightforward and simplifies the trial by reducing the number of future pairwise comparisons. However, the scenario in which a single treatment arm is found to be superior and the ethical considerations for continuing the trial have not heretofore been addressed. Most existing methods consider switching between non-inferiority and superiority within the context of a two armed trial. Here, we consider this scenario in a multi-armed trial. In Chapter 3, we give an approach to change the hypothesis from a multi-armed superiority trial to a 2-armed non-inferiority trial. Operating characteristics are compared for both adaptive and non-adaptive analysis approaches.

As treatments become ready for testing at staggered times, it is desirable to have clinical trials that can accommodate different entry times and exit times, without prematurely discarding potentially efficacious treatments. Currently the literature does not address statistical methods for adding a new treatment arm into an ongoing study. This scenario is addressed in Chapter 4, and simulation is used to evaluate the operating characteristics of several analytical methods. In Chapter 5, these analytical methods are applied to actual data from PD clinical trials. The NET-PD FS1 and FS-TOO are similar studies, and are re-analyzed as if the arms from the later study (FS-TOO) were introduced into the first study (FS1).

# CHAPTER 3: Paper 1. Transitioning from a 3-armed superiority trial to a 2-armed non-inferiority trial when superiority of a single treatment arm is determined early

Jordan J. Elm, Yuko Y. Palesch, Barbara C. Tilley, Wenle Zhao, Vanessa Hinson, Bernard Ravina, Gary Koch

## Abstract

**Background:** With group sequential multi-armed clinical trials, there is an increased chance of finding a single winning treatment before the planned end of the trial (e.g. before all subjects have completed follow-up) because either of the treatments could show overwhelming efficacy at an interim analysis. This scenario has not yet been addressed in the literature. In this case, it may be necessary to discontinue enrolling patients to the placebo arm for ethical reasons. Investigators may wish to transition from a 3-armed superiority trial (e.g., active treatment A, active treatment B, and placebo) to a 2-armed non-inferiority trial when efficacy of one treatment arm (A) relative to the placebo arm is determined at an early interim analysis. Assuming there is early outcome knowledge, this scenario is most suited to the case in which the interim analysis is conducted while enrollment is still ongoing. **Methods:** Analytical methods are proposed for this novel application to a unique clinical trial case. Monte Carlo Simulation was used to compare the following statistical methods: (1) *t*-test with the data pooled across stages (non-

adaptive); (2) Inverse Chi-square (Fisher's) adaptive combination test; and (3) Weighted-Inverse Normal adaptive combination test. The power and average sample size were compared with and without sample size inflation based on interim information. **Results:** Empirical power was highest when using an inverse chi-square adaptive testing procedure. The other methods performed similarly. When sample size was allowed to be re-estimated based on the observed effect size, the adaptive procedure required a sample size approximately 50-80% of that necessary for a non-inferiority hypothesis based on the original sample estimates.  **Conclusions:** The adaptive methods allow for a reduction in total sample size with increased power for testing non-inferiority (compared to a non-adaptive approach).

### 3.1 Introduction

Suppose a multi-armed clinical trial of more than one treatment or dose arm against a common placebo group is conducted and interim analyses are planned. When a single active arm is found to be superior at an interim analysis, we may wish to continue the study for the remaining treatment arms, since they too could be efficacious. In this case, however, it may be considered unethical to continue the placebo arm (since a superior treatment has been found)[65]. This would particularly be true if this were the second confirmatory trial for the winning drug. At this point, it may be desirable to transition into a two-armed non-inferiority study with the newly identified superior treatment as the standard. This problem can be conceptualized as changing hypotheses mid-study, where the comparison arm has changed from placebo to an active control. This scenario is most applicable to the case in which the interim analysis is conducted while enrollment is still ongoing, i.e. assuming there is early outcome knowledge.

Several authors have addressed the issue of changing from non-inferiority to superiority hypothesis, or vice versa, in the context of a single treatment versus placebo study[60,66,67]. Denne and Koch have shown that it is possible to test non-inferiority and superiority sequentially without adjusting for multiplicity because they are nested hypotheses; closed testing methods are applied[66,23]. Also in the setting of a group sequential active controlled clinical trial, Wang *et al.* illustrate an adaptive design and show how the trial can begin with the non-inferiority hypothesis and go on to test superiority, or vice versa[67]. Koyama describes a flexible two-stage adaptive design that can control both the type I and type II errors for the two simultaneous objectives of non-inferiority and superiority[60]. Kropf, Hommel, and Schmidt give a procedure for testing superiority of two treatments and then non-inferiority in the context of a two-stage adaptive clinical trial, but do not consider the need to stop the placebo arm after the first stage[61]. In this paper we consider the scenario of changing from superiority to non-inferiority in the context of a multi-armed study given a single arm wins at interim; statistical analytical methods are proposed for this novel scenario.

Consider a multi-armed clinical trial with two active treatments and a placebo in which we have three arms treatment A, treatment B, and placebo P, $K > 1$ looks. Treatment A and treatment B may be different doses of the same treatment. The aim of the trial is to determine whether A and/or B are superior to placebo rather than identifying the single best treatment. Suppose the responses are $X_{ij} \sim N(\mu_i, \sigma^2)$, $i$=(A, B, P), $j$=1, 2, ....., $n$, where a larger value of $X_{ij}$ indicates a favorable outcome. Initially, we are testing the following superiority hypotheses $H_0^{AP} : \mu_A - \mu_P = 0$ and $H_0^{BP} : \mu_B - \mu_P = 0$ in a group sequential trial. A maximum of $n$ patients per arm will be enrolled to test for superiority at the desired power. We will consider the case in which strong control of the familywise error rate is desired. Follman, Proschan and Geller recommend requiring equal evidence against any pairwise hypothesis of interest in multi-armed clinical trials[1]. A simple approach to control alpha strongly is to apply Bonferroni adjustments to pre-specified alpha such that the overall alpha is split for each individual pairwise hypothesis of interest, and then to find the corresponding (e.g. O'Brien-Fleming or Pocock type) critical values such that each pairwise hypothesis can be tested at any stage[1].

In a group sequential multi-armed clinical trial setting, the possibility arises that one arm will be found to be superior or inferior at an early interim analysis while the remaining arms will be inconclusive. If at some look $k < K$ at information time $t_k$, we reject $H_0^{AP}$ for one treatment arm, say A, versus placebo for superiority, such that $Z_{AP,k} > c$ (where *c* is a constant based on the desired type I error probability $\alpha$ and *k*) then we have concluded the efficacy trial of drug A with a positive result. However, it may be clinically desirable to continue the trial to test $H_0^{BP} : \mu_B - \mu_P = 0$ until *n* subjects per group are enrolled or $Z_{BP,k} > c$. At this point, however, it may not be considered ethical to continue to enroll patients to the placebo arm given that a superior treatment has now been identified.

If at the outset of the trial, investigators anticipate that one of the arms will show early superiority, and they wish to show non-inferiority for the second arm versus the first treatment arm, they may pre-specify that a non-inferiority test be conducted given $H_0^{AP}$ or $H_0^{BP}$ was reject at an interim analysis. From henceforth, we will refer to the arm that wins first as treatment A. After the $k^{th}$ interim analysis (where $H_0^{AP} : \mu_A - \mu_P = 0$ was rejected), enrollment to the placebo arm is stopped and a test for non-inferiority of B versus the newly established standard, A is performed. Thus, for the interim analyses >*k*, the non-inferiority hypothesis is written as $H_0^{BA}: \mu_B - \mu_A \leq -\delta$ where $\delta$ is some pre-specified margin of non-inferiority[68,69].

Now, both hypotheses $H_0^{BP}$ and $H_0^{BA}$ are of primary interest, so that the overall primary hypothesis is the intersection hypothesis $H: H_0^{BP} \cap H_0^{BA}$. Yet the power to test $H_0^{BP}$ is reduced since no more placebos are to be enrolled. It can be pre-planned to increase the sample size to that needed for a non-inferiority test should one arm be determined superior before the other. Assuming this is all pre-specified at the outset and independent of the interim observed effect size, then multiplicity can be easily controlled. The $H_0^{BA}$ hypothesis is nested within the $H_0^{BP}$ hypothesis, so closed testing methods can control multiplicity where $H_0^{BA}$ is tested as an alpha level test if $H_0^{BP}$ is rejected[66]. However, given that power may be reduced to test $H_0^{BP}$, another appropriate multiple comparison procedure is Simes' method[19] for the overall intersection hypothesis $H: H_0^{BP} \cap H_0^{BA}$, and if this is rejected then both $H_0^{AB}$ and $H_0^{BP}$ can be tested simultaneously (closure methods).

To test the intersection hypothesis $H: H_0^{BP} \cap H_0^{BA}$, we could use a *t*-test along with Simes' method or use an adaptive combination test. For the adaptive combination test of the intersection hypothesis, the $k^{th}$ stage p-value corresponds to the p-value for the test of $H_0^{BP}$ via $Z_{BP,k}$ observed at the $k^{th}$

look, and the $k^{th}$*+1* stage p-value for the non-inferiority test of $H_0^{BA}$ corresponds to the test statistic using the observations from after the $k^{th}$ look only.

Possible adaptive combination tests include Fisher's combination test (inverse Chi-square) or the Cui *et al*/Lehmacher and Wassmer method (weighted inverse normal method)[41,42]. The combination rule must be pre-specified. During an interim analysis, the sample size can be re-estimated based on internal information about the observed effect size. If the inverse Chi-square (Fisher's) combination test is specified as the adaptive combination test, then the overall intersection hypothesis $H: H_0^{BP} \cap H_0^{BA}$ is rejected if

$$C(p_1,...,p_k) = p_1...p_k \le c_\alpha = \exp[-\frac{1}{2}\chi^2_{2k,1-\alpha}]$$

where $p_1 .... p_k$ are the p-values from the first through $k^{th}$ stages of data and $\chi^2_{2k}(1-\alpha)$ is the $(1-\alpha)$-quantile of the central $\chi^2$ distribution with 2*k* degrees of freedom[38,39].

If the weighted inverse normal rule is specified as the adaptive combination test, then this can be written as $C(p_1,...,p_k) = \sum_{k=1}^{K} w_k \cdot \Phi^{-1}(1-p_k)$ where $\sum_{k=1}^{K} w_k^2 = 1$ and $w_k = \left(\sum_{k=1}^{K} n_k\right)^{-1/2} \cdot \sqrt{n_k}$ where $n_k$ is the total number of observations at the $k^{th}$ stage. Then the overall null $H$ is rejected at level alpha if $C(p_1,...,p_k) > c$ where $c$ is the usual group sequential stopping boundary[40].

If the overall intersection hypothesis $H: H_0^{BP} \cap H_0^{BA}$ is rejected via an adaptive combination test, then closed testing procedures can be applied to test $H_0^{BP}$ and $H_0^{BA}$ at level-alpha tests[48,61]. The individual test of $H_0^{BA}$ would use the sample data from stage 2 only, because this hypothesis was not named as one of primary interest at stage 1[61]. Combination testing cannot be used to test $H_0^{BP}$ because no data are available for Placebo in stage 2. The test of $H_0^{BP}$ can only be performed as the usual *t*-test, using all available data for B and Placebo, regardless of stage.

### 3.2 Methodology

The scenario above describes a change in study design from a 3-armed superiority trial to a 2-armed non-inferiority trial when efficacy of one treatment arm relative to the placebo arm is determined at an interim analysis. Monte Carlo simulations were performed to estimate the power and average sample number (ASN) for several methods when arm A is found to be superior to placebo after 50% of subjects have been enrolled and the hypothesis is changed to non-inferiority for arm B versus A (the new standard). If one arm was found to be superior at the first interim analysis, but the other was not, we denote the winning arm as treatment A. Then several methods were compared to test firstly the overall hypothesis $H: H_0^{BP} \cap H_0^{BA}$ and then the individual hypotheses $H_0^{BP}$ and $H_0^{BA}$. Methods compared were 1.) *t*-test with the data pooled across stages (non-adaptive), 2.) Inverse Chi-square (Fisher's) combination test (ICHI), and 3.) weighted inverse normal combination test (CWH). See Figure 3-1.

All simulations were done in STATA. The number of replications was set at 15,000. Samples were drawn from a normal distribution with a common variance $\sigma^2 = 1$. The true configuration of the relationships between the treatment means was specified as $\mu_A = \mu_B > \mu_P$ where '=' denotes equivalency, '>' indicates superiority. A medium-sized standardized effect size of $\theta = \frac{\mu_B - \mu_P}{\sigma} = 0.42$ was arbitrarily selected. To achieve a power of 90%, the maximum sample size needed per group for a two sample superiority *t*-test of $H_0^{AP}$ (or $H_0^{BP}$) is *n*=134 given standardized effect size $\theta = 0.42$, $\alpha = 0.025$ (one-sided). The simulation was conducted for different sizes of the non-inferiority margin $\delta$ assumed to be a fixed constant equal to $\frac{1}{5}\theta,\ \frac{1}{4}\theta,\ \frac{1}{3}\theta$, or $\frac{1}{2}\theta$. For each value of $\delta$ the corresponding maximum sample size was set to *M,* the sample size needed for a non-inferiority test (530, 1191, 2118, 3309, respectively).

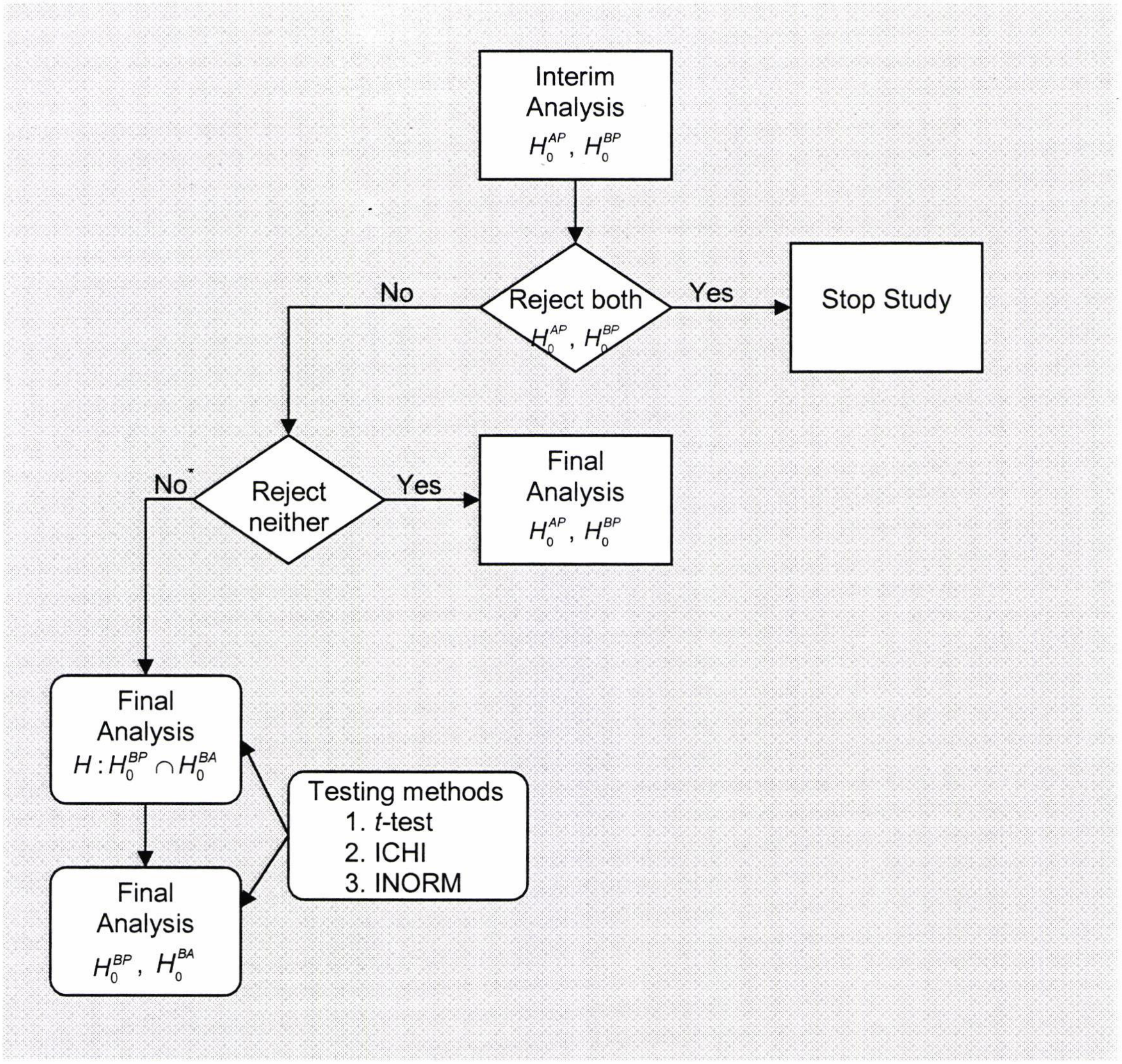
For adaptive analytical methods it is recommended to perform one-sided rather than two-sided tests to avoid contradictory decisions when the stagewise test-statistics go in opposite directions. Thus, for comparison purposes a one-sided test was performed for all methods. In order to constrain the familywise error rate to no more than 0.05, $\alpha = 0.025$ was allocated for each of the two treatment versus placebo comparisons. Then, an alpha spending function with Pocock type stopping boundaries (2.157, 2.201) was applied. An interim analysis was performed using 50% of the sample. To achieve a common feature for all comparisons, Pocock stopping boundaries were chosen (rather than the more commonly used O'brien-Flemming boundaries) because these provided congruous stopping boundaries for the inverse Chi-square (Fisher's) combination test.

For the inverse Chi-square (Fisher's) combination test, $c_\alpha$ the stopping boundary for stage 2 is given by $\alpha = \alpha_1 + c_\alpha \ln\frac{1}{\alpha_1}$ which can be solved for $c_\alpha = 0.00228$ given $\alpha = 0.025$, $\alpha_1 = 0.0155$ (the corresponding Pocock stopping boundary for stage 1), and there is no stopping in favor of the null hypothesis at stage 1.[57] (If O'brien-Flemming stopping boundaries were used then $\alpha_1 < c_\alpha$.) It is necessary to select stopping boundaries where $\alpha_1 > c_\alpha$. Otherwise it is illogical to continue the trial and perform a stage 2 test since $p_1 \cdot p_2$ would always be less than $c_\alpha$ if $p_1 < c_\alpha$.[70]

The sample size inflation rule was based on achieving a target conditional power for rejecting the null hypothesis at stage 2 expressed as $\text{Prob}(Z_2 > 2.201 | Z_{1-p_1}, \delta) = 1 - \beta$ and was bound at *M*. For the weighted inverse normal method (INORM), this expression yields the sample size per group for the second stage as $n_2 = \frac{2\sigma^2}{\delta^2}\left[\frac{1}{w_2}\left(2.201 - w_1 Z_{1-p_1}\right) - \Phi^{-1}(1-\beta)\right]^2$ where $\delta$ is the non-inferiority margin, 2.201 is the Pocock critical value for the final analysis, $w_1$ and $w_2$ are the weights as defined by the inverse normal combination test, and $Z_{1-p_1}$ is the Z-value from stage 1[57]. For the inverse Chi-square (Fisher's) combination test (ICHI) the conditional power expression yields the sample size per group[57] for the second stage as

$n_2 = \frac{2\sigma^2}{\delta^2}\left[Z_{1-\frac{\alpha_2}{p_1}} - \Phi^{-1}(1-\beta)\right]^2$. For both testing methods (ICHI and INORM), the Z-value from stage 1 was replaced with $Z_{AB,1}$ the normal Z-value for the test of the non-inferiority hypothesis at the interim analysis (stage 1) (denoted ICHI1 and INORM1) or $Z_{BP,1}$ the Z-value for the test of treatment B versus Placebo at the interim analysis (denoted ICHI2 and INORM2).

**Figure 3-1.**


Interim Analysis
$H_0^{AP}$, $H_0^{BP}$
Reject both $H_0^{AP}$, $H_0^{BP}$
Yes
Stop Study
No
Reject neither
Yes
Final Analysis $H_0^{AP}$, $H_0^{BP}$
No*
Final Analysis $H: H_0^{BP} \cap H_0^{BA}$
Final Analysis $H_0^{BP}$, $H_0^{BA}$
Testing methods
1. *t*-test
2. ICHI
3. INORM

*Assume reject only $H_0^{AP}$ where A denotes the treatment arm that wins at interim.

### 3.3 Simulation Results

The empirical power to reject $H: H_0^{BP} \cap H_0^{BA}$, $H_0^{BP}$ and $H_0^{BA}$ (given $H_0^{AP}$ was rejected after the first look and enrollment to the placebo arm was stopped) is shown in Table 3-1. Whether the per group sample size (for treatments A and B) was increased to *M* (where *M* is the sample size required to achieve 90% power for a non-inferiority test given the original sample size estimates and $\delta$) or re-estimated based on observed data, all methods achieved more than 90% nominal power for the overall intersection hypothesis $H: H_0^{BP} \cap H_0^{BA}$. For the (cumulative) *t*-test of $H_0^{BP}$ power was greater than 85% regardless of the testing method used for $H: H_0^{BP} \cap H_0^{BA}$ or the final sample size. Of note, the ICHI2 and INORM2 adaptive methods used much smaller sample size than *M* with only a minor reduction of power for the tests of $H: H_0^{BP} \cap H_0^{BA}$ and $H_0^{BP}$.

The methods compared showed different results for the power to reject $H_0^{BA}$. The inverse Chi-square adaptive test (ICHI1) achieved the nominal power to reject $H_0^{BA}$. In general the inverse Chi-square adaptive tests (ICHI1 and ICHI2) achieved moderately more power to reject $H_0^{BA}$ than the weight-inverse normal tests (INORM1 and INORM2), for the respective re-estimation methods. The (cumulative) *t*-test did not achieve the nominal power (83% vs 90%). Of note, the whole simulation was repeated (exception for the ICHI tests) with O'brien-Flemming type stopping bounds and the results were similar to those given in Table 3-1 where Pocock bounds were used (results not shown).

**Table 3-1. Empirical Power (95% CI) to reject $H: H_0^{BP} \cap H_0^{BA}$, $H_0^{BP}$ and $H_0^{BA}$ given $H_0^{AP}$ was rejected after the first look and enrollment to the placebo arm was stopped (12,000 out of 60,000 reps where $\mu_A = \mu_B > \mu_P$)**

| **Method** | **Sample Size for groups A and B** | $H: H_0^{BP} \cap H_0^{BA}$ | $H_0^{BP}$ † | $H_0^{BA}$ |
|---|---|---|---|---|
| *t*-test (Simes' method) | $n_A = n_B = M$ | 0.943<br>(0.940, 0.946) | 0.886<br>(0.882, 0.891) | 0.825<br>(0.820, 0.831) |
| Inverse Chi-square (Fisher's) combination | | | | |
| ICHI1 | Re-estimated (CP given $Z_{AB,1}$) | 0.969<br>(0.966, 0.971) | 0.882<br>(0.878, 0.887) | 0.869<br>(0.864, 0.874) |
| ICHI2 | Re-estimated (CP given $Z_{BP,1}$) | 0.942<br>(0.939, 0.945) | 0.858<br>(0.853, 0.863) | 0.747<br>(0.741, 0.753) |
| Weighted-Inverse Normal combination | | | | |
| INORM1 | Re-estimated (CP given $Z_{AB,1}$) | 0.973<br>(0.971, 0.975) | 0.886<br>(0.882, 0.891) | 0.867<br>(0.862, 0.871) |
| INORM2 | Re-estimated (CP given $Z_{BP,1}$) | 0.938<br>(0.935, 0.942) | 0.855<br>(0.850, 0.860) | 0.690<br>(0.683, 0.696) |

Note: Results based on 12,000 of 60,000 (32%) simulations where one arm (but not the other) was found to be superior to Placebo at the first interim analysis (e.g. $H_0^{AP}: \mu_A - \mu_P = 0$ was rejected at stage 1). Results averaged across the non-inferiority margins $\delta$. The power for $H_0^{BP}: \mu_B - \mu_P = 0$ and $H_0^{BA}: \mu_B - \mu_A \le -\delta$ assumes that $H: H_0^{BP} \cap H_0^{BA}$ was also rejected. (Closed testing methods were applied to control for multiple comparisons).

† $H_0^{BP}$ is performed via *t*-test using all available data regardless of method to test $H: H_0^{BP} \cap H_0^{BA}$. (Combination testing cannot be used to test $H_0^{BP}$ because no data are available for Placebo in stage 2).

Sample size re-estimation was based conditional power (CP) for either Inverse Chi-square (Fisher's) combination test (ICHI) or weighted-inverse combination test (INORM) given $Z_{AB,1}$ or $Z_{BP,1}$ (ICHI1, ICHI2, INORM1, INORM2 respectively).

The adaptive combination tests of $H_0^{BA}$ achieved higher power when sample size re-estimation was based on $Z_{AB,1}$ (ICHI1, INORM1) rather than $Z_{BP,1}$ (ICHI2, INORM2) regardless of $\delta$. For adaptive combination tests INORM1 and ICHI1 where the sample size was re-estimated using conditional power given $Z_{AB,1}$, then the resultant power was equal to setting the sample size at *M*.

The average sample number (ASN) per group after sample size re-estimation is given in Table 3-2. When $Z_{AB,1}$ is used to re-estimate the sample size, then ASN is roughly 80% of *M*. In contrast, when $Z_{BP,1}$ is used to re-estimate the sample size (ICHI2 and INORM2), the ASN is much smaller (approximately 50% of *M*).

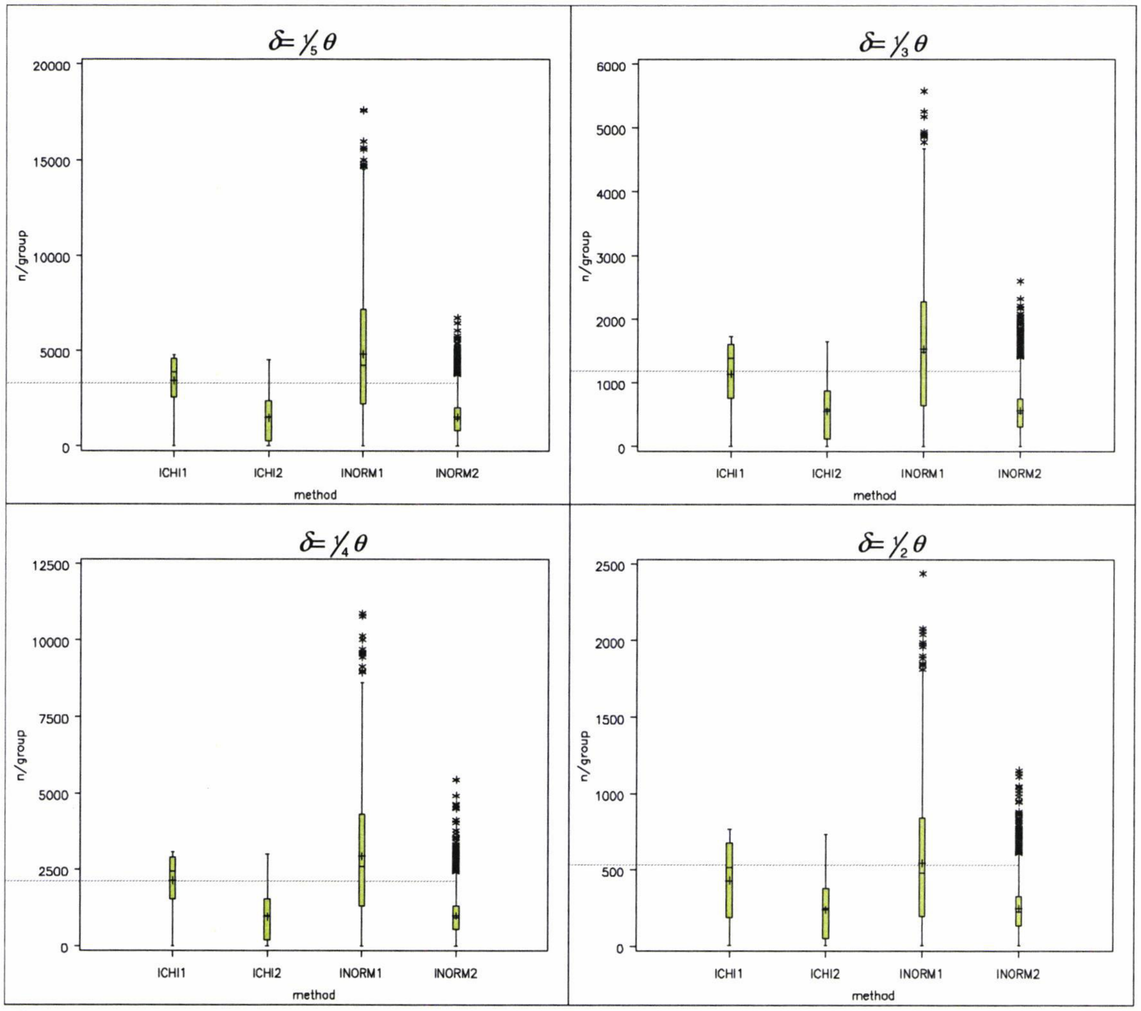
Figure 3-4 shows Box and Whisker plots of the sample size per group after re-estimation based on conditional power for either Inverse Chi-square (Fisher's) combination test or weighted-inverse combination test given $Z_{AB,1}$ or $Z_{BP,1}$ (ICHI1, ICHI2, INORM1, INORM2 respectively). For the ICHI1 method a greater amount of the unbounded distribution fell below the gray horizontal line (representing *M* the maximum sample size per group) versus the INORM1 method. Thus, ICHI1 more often resulted in smaller sample sizes than INORM1, yet achieved higher power for the test of $H_0^{BA}$.

**Table 3-2. Average Sample Number (ASN) and Maximum Sample Size per group for Treatment arms A and B after Sample Size Re-estimation based on Conditional Power (CP)**

| $\delta$ | Sample Size Re-estimation Method | ASN per group | Maximum per group (bound at *M*) |
|---|---|---|---|
| $\frac{1}{5}\theta$ | ICHI1 | 2778 | 3309 |
| | ICHI2 | 1561 | 3309 |
| | INORM1 | 2843 | 3309 |
| | INORM2 | 1534 | 3309 |
| $\frac{1}{4}\theta$ | ICHI1 | 1740 | 2118 |
| | ICHI2 | 1026 | 2118 |
| | INORM1 | 1775 | 2118 |
| | INORM2 | 1012 | 2118 |
| $\frac{1}{3}\theta$ | ICHI1 | 939 | 1191 |
| | ICHI2 | 597 | 1191 |
| | INORM1 | 951 | 1191 |
| | INORM2 | 585 | 1191 |
| $\frac{1}{2}\theta$ | ICHI1 | 401 | 530 |
| | ICHI2 | 303 | 530 |
| | INORM1 | 395 | 530 |
| | INORM2 | 298 | 530 |

$\delta$ is the non-inferiority margin assumed to be a fixed constant equal to $\frac{1}{5}\theta$, $\frac{1}{4}\theta$, $\frac{1}{3}\theta$, or $\frac{1}{2}\theta$, where $\theta = 0.42$ is the standardized effect size. The maximum sample size for the complete study was bound at *M* per group for Treatment A and Treatment B. The Placebo arm is stopped at *n*=67 after stage 1. Sample size re-estimation was based conditional power for either Inverse Chi-square (Fisher's) combination test (ICHI) or weighted-inverse combination test (INORM) given $Z_{AB,1}$ or $Z_{BP,1}$ (ICHI1, ICHI2, INORM1, INORM2 respectively).

**Figure 3-4. Box and Whisker plots of sample size re-estimation (unbounded)**



$\delta$ is the non-inferiority margin assumed to be a fixed constant equal to $\frac{1}{5}\theta$, $\frac{1}{4}\theta$, $\frac{1}{3}\theta$, or $\frac{1}{2}\theta$, where $\theta = 0.42$ is the standardized effect size. Gray horizontal line represents $M$ (maximum sample size per group) for each $\delta$. Sample size re-estimation was based on conditional power for either Inverse Chi-square (Fisher's) combination test or weighted-inverse combination test given $Z_{AB,1}$ or $Z_{BP,1}$ (ICHI1, ICHI2, INORM1, INORM2 respectively).

### 3.4 Discussion

These simulations show that if we allow for an interim look in a multi-armed study using Pocock type stopping boundaries, there is a 32% chance that only one of the arms will be found to be superior to placebo at the first look (given $\mu_A = \mu_B > \mu_P$). Simulations under the same parameters as above except with O'Brien-Flemming (rather than Pocock) stopping bounds indicated that there is a 25% chance that at least one of the arms (but not the other) will be found to be superior to placebo at the interim analysis (given $\mu_A = \mu_B > \mu_P$, complete results not shown). Thus, when multi-armed studies are conducted in a group sequential design, there is a nontrivial chance that we could be discarding a potentially effective treatment just because one treatment was found to be superior to placebo sooner and the trial was stopped early. Hence, it is important to consider the methods that will be used for this scenario.

In this paper, we consider ways to continue the trial to compare B vs Placebo and B vs A (for non-inferiority), when the placebo arm is dropped for ethical reasons (e.g. because treatment A is clearly established as superior to placebo). It is necessary to increase the sample size from the original size needed for $H_0^{BP}$, if $H_0^{BA}$ is to be adequately powered. If we were to continue the study of A and B, stopping the placebo arm but without increasing the original sample size, then the power to reject $H_0^{BP}$ is around 50% and power to reject $H_0^{BA}$ would be less than 10%.

In this scenario, non-inferiority was only tested if drug A was found to be superior to placebo at the interim analysis. However, the interim estimate of drug A must be greatly superior to placebo to win at an interim analysis. For the non-adaptive approach performing a *t*-test for the non-inferiority hypothesis, then the data from patients in Treatment A (selected as extreme) are later used to form the test of $H_0^{BA}$. The selection of the extreme data at stage 1 has an additional effect of making it harder to reject the non-inferiority hypothesis $H_0^{BA}$ at the final analysis.

This can be seen in Table 3-1 where we were unable to achieve the nominal power with the (cumulative) *t*-test of $H_0^{BA}$ (empirical power was 83% rather than 90%) when n=*M*. Thus, an extreme result for drug A will make it somewhat less likely that B will be found non-inferior to A at the final analysis (even when $\mu_A = \mu_B$). For the adaptive testing procedure, the non-inferiority hypothesis is added mid-course when the test statistic for drug A versus placebo is significant, and the stage 1 data are not used to test $H_0^{BA}$ since this hypothesis was not specified as primary at the outset. Thus, the adaptive tests outperformed the non-adaptive testing scheme in these simulations.

It is well known that sequential testing produces bias in the usual (MLE) point estimation[71]. Future research could address applying a bias correction to the test statistic (correcting for the selection of the winning arm in stage 1) and address bias-adjusted point estimation (such as where the estimate is conditioned on the number of stages performed)[72].

For both adaptive methods, the chance of rejecting $H: H_0^{BP} \cap H_0^{BA}$ and $H_0^{BP}$ while failing to reject $H_0^{BA}$, occurs rarely (with ICHI1 and INORM1) or occasionally (with ICHI2 and INORM2). However, the re-estimated sample size is substantially smaller than *M* (particularly when the re-estimation is based on $Z_{BP,1}$). These simulations were conducted across a range of non-inferiority margins $\delta = (\frac{1}{5}\theta,\ \frac{1}{4}\theta,\ \frac{1}{3}\theta,\ \frac{1}{2}\theta)$. While $\frac{1}{2}\theta$ may seem like a large non-inferiority margin, it may be appropriate for this situation[61]. Considering that the original primary aim was to show that treatment B was better than placebo, then treatment B could be considered efficacious compared to placebo if it is more than one-half of that between reference (treatment A) and placebo[66]. Albeit $\theta$ is the clinically important difference, not the actual difference between reference and placebo.

It is worth noting that the choice of the non-inferiority margin should be clinically based and pre-specified prior to the start of trial. Here, we defined the non-inferiority margin as a fixed constant. The non-inferiority margin could have been defined as a percent retention of the (observed) active control effect[68,61]. In this example, the maximum sample size was increased by x25 when the smallest non-inferiority margin was chosen. This substantial increase would argue for keeping the placebo arm in the study (if ethically possible) or choosing a larger non-inferiority margin.

As with all multi-armed studies, multiplicity is an important consideration. In this simulation we split the alpha for the two arms and then applied an error-spending function for the two interim analyses. Exact methods would increase power over the Bonferroni adjustment, but are computationally more complex. The sequentially rejective procedure given by Follman *et al.* was developed for when one arm is dropped for inferiority[1]. With this procedure, exact critical values are derived for each interim analysis such that the remaining looks would use a critical value that corresponds to the reduced number of arms. This sequentially rejective procedure is slightly less conservative than the Bonferroni approach and holds for up to four arms[1]. However, Hellmich (2001) has shown that the sequentially rejective procedure may fail to strongly protect the type I error rate if the arm is dropped for some reason other than statistical inferiority[2]. Thus, in the case where the placebo arm is dropped due to statistical superiority of treatment arm A, the critical value should be based on the original number of arms, not the reduced number of arms.

Adaptive methodology allows for much flexibility of clinical trials, yet the planning of such trials is more involved. Some advise to pre-specify all possible scenarios in the protocol of an adaptive study and to perform simulations at the planning stage[12,65]. Often, adaptive designs are used to control the type I error rate when the study sample size is re-estimated using the observed data. The rule used to re-estimate the sample size (e.g. conditional power) is independent of our ability to control the type I error rate; rather, it is the use of internal data that will inflate alpha if the test statistic is not adjusted[73]. However, if the study sample is increased to $M$ based solely on the rejection of the primary hypothesis for arm A (and without looking at the effect size for arm B), then the type I error rate will not be inflated for the treatment B comparison. This would be the case even if the treatment arms are two doses of the same drug, since it is the use of internal information (rather than a pre-specified $M$) which inflates the alpha.  One limitation of the adaptive testing procedure is the restriction to one-sided testing, as confirmatory clinical trials are frequently conducted as two-sided tests. However, one solution may be to perform doubly one-sided tests developed by Wassmer (publication in German)[74].

The finding of overwhelming efficacy for a single treatment at an interim analysis of a multi-armed clinical trial may not be entirely straightforward. For instance, while treatment A may be efficacious, it may be difficult to tolerate or be very costly for the patient, while treatment B may be inexpensive and/or better tolerated. For such reasons, it may be clinically worthwhile to continue the study of treatment B in the hopes that it may be non-inferior to A[61]. If A is greatly superior to Placebo at interim, then in order for B to have a chance at being non-inferior to A at the end, B would have to be close to the stopping boundary at interim. In that case it makes sense to assess the conditional power for B versus Placebo at the interim analysis to decide whether to continue at all. Incorporating rules such as early stopping for futility (e.g. based on conditional power of 50% for treatment B) would help dictate whether to stop the whole study where A would be the only winner.

From a regulatory perspective, it is necessary to show efficacy of a drug in two independent confirmatory (phase III) trials in order for that drug to be considered for regulatory approval of indication[65]. In the absence of this, the FDA may not view treatment A as a standard after the interim results. A protocol with a plan to transition into a non-inferiority trial upon finding a winning treatment may not meet FDA approval, unless there was a prior successful confirmatory trial.

As with all non-inferiority trials conducted without a concurrent placebo arm, there is the assumption of constancy of the historical control estimate. A cohort effect is an important consideration for any type of design change, particularly when the placebo rate is not well established or is known to shift over time[15]. One would need to account for this in the analysis. Another possible approach not considered in these simulations that would allow one to adjust for the stage or cohort effect is regression based approach. Specifically, a four parameter model including an intercept, an indicator for cohort, an indicator for treatment A and an indicator for treatment B could be fit. If there are covariates that account for a cohort effect, such as differences in inclusion/exclusion criteria or a time effect, these covariates could also be included in the model[15]. Such a model may be more powerful than the pairwise *t*-tests presented here. A comparison of such methods is beyond the scope of this paper.

When an interim analysis is conducted, there is the issue of who becomes unblinded. Because a one-sided test is performed, it is necessary to have a fully unblinded statistician to perform the analysis. It may be possible for the DSMB to discontinue one arm (e.g. the placebo arm) without announcing the results of the interim analysis. This could be achieved if all possible scenarios were pre-specified in the protocol such that a protocol amendment was not required to drop an arm. Logistical issues and ramifications of any major of design change should be carefully considered and are reviewed by Geller and Pocock (1987)[75].

This design is suitable for a clinical trial where the enrollment is relatively slow as compared to the length of follow-up such that there are patients remaining to be enrolled at the time of the interim analysis. If enrollment has been fully completed by the time of the first interim analysis, then transitioning into a non-inferiority trial would necessitate re-starting enrollment. In this case, the ethical concerns of continuing to enroll new patients to the placebo arm are moot. However, handling of some placebo subjects who have yet to complete the protocol-specified treatment will remain an issue.

An example of an ongoing PD clinical trial in which this design would be well-suited is the ongoing NIH-funded trial, "Effects of Coenzyme Q10 in Parkinson Disease-Phase 3 (QE3)". This is a multi-armed trial of several doses of Coenzyme Q10 and placebo where 600 patients are to be follow-up for 16 months. Given the desired sample size, it is likely that the speed of enrollment will be relatively slow compared to the length of follow-up. If a higher dose of CoQ10 is found to be efficacious at an interim analysis, it would still be of interest to consider the non-inferiority of a lower dose arm.

Long-term clinical trials generate more ethical dilemmas. If an interim analysis finds a superior drug (A), yet the trial is to continue for drug B, what is to be done with the subjects already enrolled on placebo who have not yet completed follow-up? Switching placebo subjects to the winning drug (with or without their knowledge) and thereby using their data may or may not be an option. The increase in sample size could outweigh the possible dilution of the treatment effect. However, this is likely to be true only for chronic disease types. Progressive diseases (such as Parkinson's disease) may have irreversible worsening over time such that introducing the winning treatment a long time after baseline (rather than just after diagnosis) would not be efficacious.

One implication of conducting multi-armed group sequential clinical trials is the increased chance of having a single winning treatment arm at interim. If switching to a non-inferiority hypothesis is meaningful, then the adaptive methods are more powerful at a smaller sample size to declare treatment B non-inferior to A.

**Appendix 3-1. Empirical Power to reject $H: H_0^{BP} \cap H_0^{BA}$, $H_0^{BP}$ and $H_0^{BA}$ given $H_0^{AP}$ was rejected after the first look and enrollment to the placebo arm was stopped (15,000 reps)**

| | non-inferiority margin ($\delta$) | $\frac{1}{5}\theta$ | $\frac{1}{4}\theta$ | $\frac{1}{3}\theta$ | $\frac{1}{2}\theta$ |
|---|---|---|---|---|---|
| | ***M*** | **3309** | **2118** | **1191** | **530** |
| ***t*-test** | $H: H_0^{BP} \cap H_0^{BA}$ | 0.968 | 0.958 | 0.942 | 0.902 |
| | $H_0^{BP}$ | 0.919 | 0.909 | 0.884 | 0.832 |
| | $H_0^{BA}$ | 0.855 | 0.837 | 0.818 | 0.791 |
| **Inverse Chi-square (Fisher's) combination ICHI n=*M*** | $H: H_0^{BP} \cap H_0^{BA}$ | 0.974 | 0.973 | 0.964 | 0.964 |
| | $H_0^{BP}$ | 0.91 | 0.903 | 0.875 | 0.841 |
| | $H_0^{BA}$ | 0.888 | 0.877 | 0.861 | 0.848 |
| **ICHI1** | $H: H_0^{BP} \cap H_0^{BA}$ | 0.974 | 0.973 | 0.964 | 0.964 |
| | $H_0^{BP}$ | 0.91 | 0.903 | 0.875 | 0.841 |
| | $H_0^{BA}$ | 0.888 | 0.877 | 0.861 | 0.848 |
| **ICHI2** | $H: H_0^{BP} \cap H_0^{BA}$ | 0.945 | 0.946 | 0.937 | 0.941 |
| | $H_0^{BP}$ | 0.883 | 0.876 | 0.85 | 0.821 |
| | $H_0^{BA}$ | 0.749 | 0.747 | 0.737 | 0.753 |
| **Weighted-Inverse Normal combination<br>INORM n=*M*** | $H: H_0^{BP} \cap H_0^{BA}$ | 0.976 | 0.977 | 0.97 | 0.969 |
| | $H_0^{BP}$ | 0.913 | 0.906 | 0.88 | 0.846 |
| | $H_0^{BA}$ | 0.887 | 0.875 | 0.859 | 0.846 |
| **INORM1** | $H: H_0^{BP} \cap H_0^{BA}$ | 0.976 | 0.977 | 0.97 | 0.969 |
| | $H_0^{BP}$ | 0.913 | 0.906 | 0.88 | 0.846 |
| | $H_0^{BA}$ | 0.887 | 0.875 | 0.859 | 0.846 |
| **INORM2** | $H: H_0^{BP} \cap H_0^{BA}$ | 0.941 | 0.941 | 0.933 | 0.938 |
| | $H_0^{BP}$ | 0.881 | 0.874 | 0.847 | 0.82 |
| | $H_0^{BA}$ | 0.691 | 0.696 | 0.674 | 0.698 |

Note: Results based on simulations where one arm (but not the other) was found to be superior to Placebo at the first interim analysis (e.g. $H_0^{AP}$ was rejected at stage 1). $\delta$ is the non-inferiority margin assumed to be a fixed constant equal to $\frac{1}{5}\theta$, $\frac{1}{4}\theta$, $\frac{1}{3}\theta$, or $\frac{1}{2}\theta$, where $\theta = 0.42$ is the standardized effect size. $H_0^{BP}: \mu_B - \mu_P = 0$ and $H_0^{BA}: \mu_B - \mu_A \leq -\delta$.

# CHAPTER 4: Paper 2. Flexible Analytical methods for Adding a Treatment Arm Mid-study to an Ongoing Clinical Trial

Jordan J. Elm, Yuko Y. Palesch, Barbara C. Tilley, Wenle Zhao, Vanessa Hinson, Bernard Ravina, Gary Koch

Abstract

**Background:** It is not uncommon for clinical trialists to have experimental drugs under different stages of development (Phase I, II, III). For some diseases (such as Parkinson's disease), confirmatory clinical trials require a large number of subjects followed long term in order to identify a clinically meaningful treatment effect. Multi-armed clinical trials maximize efficiency by requiring smaller number of subjects receiving placebo as compared to separate trials. Methods are proposed for use when another treatment arm (B) is to be added mid-study to an ongoing clinical trial (of treatment A vs. Placebo). **Methods:** For this novel scenario, potential analytical approaches considered for pairwise comparisons of a difference in means in independent normal populations include 1.) a linear model adjusting for the design change (stage effect) or 2.) the use of an adaptive combination test. Monte Carlo simulation was used to compare the power and type I error rate for these approaches. **Results:** In the presence of intra-stage correlation, simply pooling the data will result in a loss of power for both treatment group comparisons and will inflate the type I error rate. The linear model approach is more powerful, but the adaptive method allows for more flexibility in terms of re-estimating the sample size. **Conclusions:** The flexibility to add a treatment arm to an ongoing trial will allow for cost savings (in terms of sample size as compared to two separate trials) as treatments that become ready for Phase III testing can be added to ongoing large, long-term studies in PD. Either linear model methods or adaptive combination testing methods may be applied.

## 4.1 Introduction

Statistical analytical methods for adding a new treatment arm to an ongoing clinical trial have not been addressed in the literature. Consider the scenario of a randomized, double-blind parallel arm clinical trial of Treatment A versus Placebo. This study may be large and long-term. At some point after randomization has begun, but prior to the end of enrollment, a new Treatment B showing promise is identified. Investigators and Sponsor desire to add Treatment B to the ongoing study in order to reduce the number of placebo subjects that would be needed if two separate clinical trials were to be conducted. It would be prudent logistically and economically to add Treatment B to the ongoing study. Thus, without re-randomizing previously enrolled subjects, the decision is made to randomize all new patients to one of three arms: A (with Placebo for B), B (with Placebo for A), or Placebo for A and B.

One example where such a scenario may be applied is the NINDS Exploratory Trials in Parkinson's Disease (NET-PD) initiative. This program funds a series of clinical trials[6,7] of potentially disease modifying agents. These agents are at different stages of development; some require Phase II testing, while others are ready for Phase III testing. The NET-PD group is concurrently conducting a large, long-term phase III trial and continues to pursue Phase II trials of additional agents, as new agents are identified. Since information on new agents is available in a staggered fashion, a Phase III study design that can accommodate adding arms at different times is desirable.

While there are many symptomatic treatments of PD, there are no existing treatments that have been shown to slow disease progression. As such, clinicians are interested in identifying any drug that is better than placebo (i.e. any drug that slows clinical decline). Therefore, the main interest in multi-arm studies in this context is not to identify the best drug, but rather, to identify any and all drugs that are better than placebo.

When an arm is added to an ongoing trial there are several statistical considerations. Here, we focus on the family-wise type I error rate, power, sample size, and the choice of statistical analytical methods. It is assumed that it is possible to ensure adequate blinding, that re-randomization of existing subjects cannot and will not be done, and the optimal allocation ratio will be applied. (The allocation scheme would be unequal after the new treatment arm is added and would minimize the time to total enrollment.) In this paper, analytical methods for both single-stage and group sequential designs are addressed for this novel scenario. The power and type I error rate are compared for several analytical methods for a single-stage (fixed sample) design.

## 4.2 Methodology

### *Single-stage design*

The choice of test statistic to be applied depends on the original design of the comparison of A versus P, and the potential for a cohort or study effect.

Let $A_1, A_2, ....$ and $P_1, P_2, ....$ be sequences of independent observations receiving Treatment A and placebo, respectively, in a two armed clinical trial. Restricting our attention to the test of normal means, assume their respective means are $\mu_A$ and $\mu_P$ with variance unknown. The null hypothesis of interest for a test of two normal means with variance unknown is $H_A : \theta_A \leq 0$, where $\mu_A - \mu_P = \theta_A$ and a positive difference represents a treatment benefit. Later, a new Treatment B is added ($B_1, B_2, ....$), where the null hypothesis of interest is $H_B : \theta_B \leq 0$. We will restrict our attention to one-sided testing because it is advisable in the adaptive setting to perform one-sided tests to avoid conflicting decisions based on the test statistics going in different directions at each stage. Now we have the new overall null hypothesis $H_{AB} : H_A \cap H_B$.

There are several methods that could be used to test $H_{AB} : H_A \cap H_B$. We will use the terminology stage 1 to refer to the time period prior to the design change and stage 2 to refer to the time period after the design change. Firstly, one could ignore that the subjects randomized to placebo in stage 1 did not receive the placebo for drug B, and then naively pool the data across stages. Then, two 2-sample *t*-tests would be computed. Alternatively, a linear model adjusting for a stage/cohort effect (as a fixed effect) could be applied. The regression model can be specified as follows:

$$Y_{ik} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_{ik}$$

for $i$=1,2,...,$n$ subjects within $k$=1,2 stages, where $\beta_1$ is the effect for treatment A, $\beta_2$ the effect for treatment B, $\beta_3$ the stage (cohort) effect and $\varepsilon_{ij}$ is the random effect distributed normal (0, $\sigma^2$). A third approach would be to apply an adaptive combination rule for the data from the two stages. By using this method, we are penalizing ourselves for the adaptation and ensuring that the pre-specified type I error rate is not compromised (even though the design change was externally driven).

Given another arm is added mid-study, multiple comparison procedures must be utilized in order to control the family-wise error rate (FWE) at the pre-specified rate. Although it has been argued that those performing multi-armed studies are penalized by having to maintain the FWE rate at alpha while two separate studies could each be performed at level alpha[16], we will restrict our attention to methods for controlling the FWE rate strongly, as this is standard practice for confirmatory clinical trials. Either all pairwise comparisons or many-to-one comparisons could be of interest, but we will restrict our attention to many-to-one comparisons. A simple approach valid for any test is the Bonferroni-adjustment, although other approaches such as Holmes[18], Hochberg[20], or stepwise closed testing methods[23,62] may be less conservative.

*Adaptive Procedure*

When an arm is added to an ongoing trial, a change has been made to the study design. The primary hypothesis changes from a pairwise test of A versus placebo to a global (intersection) hypothesis of two pairwise tests (or to an F- test). An adaptive design may be well suited for this application, as a way of splitting the study into two stages (before and after introduction of the new arm). In this situation, the decision to add another arm is likely based on external information (rather than based on an interim analysis). Adaptive combination tests, such as the two-stage Inverse Chi-square (Fisher's) combination test[39] (ICHI) and weighted inverse normal combination test[40] (INORM), can be used to combine data across the two stages. The combination rule must be pre-specified at or before the time of the design change.

If Inverse Chi-square (Fisher's) combination test (ICHI) is specified as the adaptive combination test, then $H_{AB}$ is rejected if

$$C(p_1, p_2) = p_1 p_2 \leq c_\alpha = \exp[-\frac{1}{2}\chi_4^2(1-\alpha)]$$

where $p_1$ is the p-value from stage 1 and $p_2$ is the p-value from stage 2 and $\chi_4^2(1-\alpha)$ is the $(1-\alpha)$-quantile of the central $\chi^2$ distribution with 4 degrees of freedom[39].

If the weighted inverse normal (INORM) rule is specified as the adaptive combination test, then this can be written as

$$C(p_1, p_2) = w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2) \text{ where } \sum_{k=1}^{2} w_k^2 = 1 \text{ and}$$

$w_k = \left(\sum_{k=1}^{2} n_k\right)^{-1/2} \cdot \sqrt{n_k}$ and $n_k$ is the total number of observations at the $k^{th}$

stage. Then the null intersection hypothesis $H_{AB}$ is rejected at level alpha if $C(p_1, p_2) > \Phi^{-1}(1-\alpha)$[40]. It is advisable to perform one-sided rather than two-sided tests to avoid conflicting decisions when the intermediate test-statistics (for each stage) go in different directions.

The test of $H_{AB}$ is as follows. We denote the intermediate p-values as $p_{k,m}$ for *k*=1, 2 stages and *m*=A, B, AB hypotheses. For stage 1, the $H_A$ null hypothesis would be tested via the usual one-sided, two-sample *t*-test, using all observations from stage 1 to obtain the intermediate p-value $p_{1,A}$. Using the observations from stage 2 only, the intermediate p-value for $H_A$ the test of treatment A versus Placebo is $p_{2,A}$ obtained via a one-sided, two-sample *t*-test. Likewise, the intermediate p-value $p_{2,B}$ for $H_B$ (the test of treatment B versus Placebo) is obtained. The stage 2 p-value for $H_{AB}$ using Simes' method[19,20] is defined as $p_{2,AB} = min[2min(p_{2,A}, p_{2,B}),\ max(p_{2,A}, p_{2,B})]$. Finally, the intermediate p-value from stage 1 and the (multiplicity-adjusted) intermediate p-value from stage 2 form a combination test of $H_{AB} : H_A \cap H_B$.

Several authors have shown the usefulness of closed testing (closure) methods for controlling multiplicity in complex adaptive designs[48,49]. Closed testing methods are applied to test $H_A$ and $H_B$ across stages. By closure methods, any individual hypothesis can be rejected at global level alpha if the set of all possible intersections is rejected at an appropriate alpha-level test. Thus, $H_A$ can be rejected given $H_{AB}$ is also rejected, where $H_{AB}$ and $H_A$ are both rejected via combination tests over stages 1 and 2. Similarly, $H_B$ is rejected at global level alpha if $H_{AB}$ and $H_B$ are both rejected[23], where $H_B$ is tested using just the stage 2 data via $p_{2,B}$ (since there is no data for treatment B in stage 1).

*Adaptive Procedure within Group Sequential setting*

Assume the same design as above, except now the null hypothesis $H_A$ was originally planned to be tested at *k* interim looks using classical group sequential methods (e.g. O'Brien-Fleming, Pocock, or an error-spending function) to allow for early stopping in favor of the alternative hypothesis. Then, to add treatment arm B and test $H_{AB} : H_A \cap H_B$ via an adaptive combination test, one must select an adaptive test that incorporates the existing group sequential framework such as the Cui *et al* weighted inverse normal method[41]. Optionally, the interim conditional power can be computed (either for the next interim look or for the whole study), and the sample size can be increased accordingly based on internal information about the effect size observed at an interim analysis.

*Hypothetical Example of Adaptive Procedure within Group Sequential setting*

Assume a phase III clinical trial of Treatment A versus Placebo is ongoing with an interim analysis planned after 50% of subjects have completed follow-up. Assuming an alpha spending function with O'Brien-

Flemming type stopping boundaries for a one-sided test of alpha=0.05, the planned maximum sample size is 200 per group. After 40 subjects per group have been enrolled (20% of the total sample size), then a new drug B becomes ready for phase III testing and the study sponsor would like to add it to the ongoing trial. The protocol is amended to include Treatment B, and the planned sample size is increased to 600 total (200 per group) assuming the original design parameters for A versus Placebo. The weighted inverse normal adaptive combination rule given by Cui, Hung, Wang (1999) is pre-specified[41]. Multiple comparison procedures will be applied to control the FWE rate strongly at alpha=0.05.

At the first interim analysis, the data are subset into stages 1 and 2 (before and after the design change). Using the stage 1 data, the p-value for the test of $H_A$ is $p_{1,A}$=0.20 and the p-value for the test of $H_B$ is not available. So the p-value for $H_{AB}$ at stage 1 is $p_{1,AB} = p_{1,A}$=0.20.

Using the stage 2 data at the first interim analysis, the p-value for the test of $H_A$ is $p_{2,A}$=0.15 and the p-value for the test of $H_B$ is $p_{2,B}$=0.06. Then using Simes' method the p-value for $H_{AB}$ at stage 2 is

$p_{2,AB} = min[2min(p_{2,A}, p_{2,B}),\ max(p_{2,A}, p_{2,B})]$=0.12.

Then the combination test for $H_{AB}$ at the first interim analysis using weights proportional to the original sample size is

$$C(p_{1,AB}, p_{2,AB}) = \sqrt{0.4} * \Phi^{-1}(1-p_{1,AB}) + \sqrt{0.6} * \Phi^{-1}(1-p_{2,AB}) = 1.442$$

where $1.442 < 2.538$ the O'brien Flemming stopping boundary when the information time is 0.50.

Since the overall null hypothesis $H_{AB}$ is not rejected, we continue the trial. Optionally, the sample size can be re-estimated by computing the conditional power for $H_A$ and $H_B$ and the sample size can be inflated this as needed. Then, using the data from after the first interim analysis only, the p-value for the test of $H_A$ is $p_{3,A}$=0.2 and the p-value for the test of $H_B$ is $p_{3,B}$=0.03. So the p-value for $H_{AB}$ at look 2 is

$p_{3,AB} = min[2min(p_{3,A}, p_{3,B}),\ max(p_{3,A}, p_{3,B})]$=0.06.

Then the combination tests at the final analysis are as follows:

For $H_{AB}$, $C(p_{1,AB}, p_{2,AB}, p_{3,AB}) = \sqrt{0.5} * C(p_{1,AB}, p_{2,AB}) + \sqrt{0.5} * \Phi^{-1}(1 - p_{3,AB}) = 2.119$

where 2.119>1.6621

For $H_A$, $C(p_{1,A}, p_{2,A}, p_{3,A}) = \sqrt{0.5} * C(p_{1,A}, p_{2,A}) + \sqrt{0.5} * \Phi^{-1}(1 - p_{3,A}) = 1.539$

where 1.539<1.6621

For $H_B$, $C(p_{2,B}, p_{3,B}) = \sqrt{0.5} * \Phi^{-1}(1 - p_{2,B}) + \sqrt{0.5} * \Phi^{-1}(1 - p_{3,B}) = 2.429$

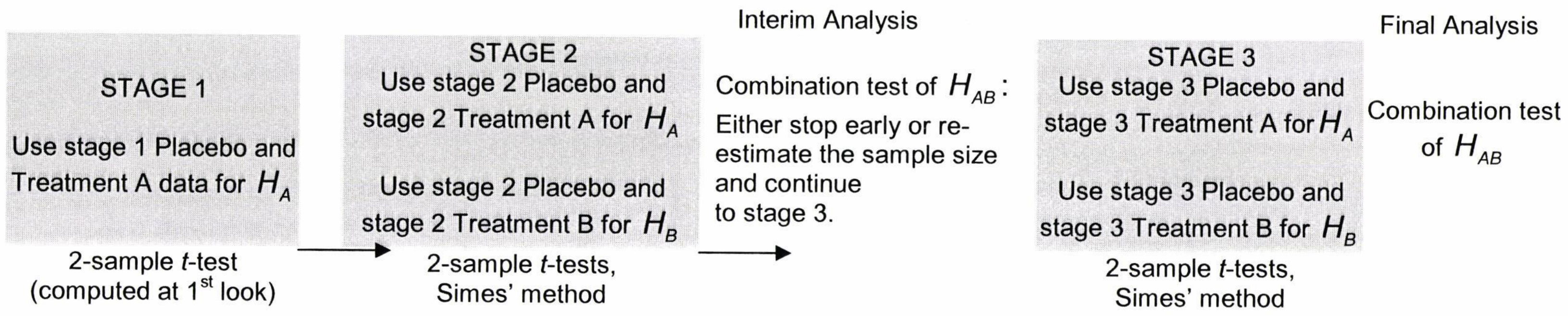
where 2.429>1.6621

Since the overall null hypothesis $H_{AB}$ is rejected, we go on to test $H_A$ and $H_B$, via combination tests. For $H_A : \theta_A \leq 0$, we fail to reject the null hypothesis since the Z-statistic for the combination test, 1.539 is less than

1.6621 (the O'Brien-Flemming type stopping boundary when 100% of the subjects have completed follow-up). For $H_B : \theta_B \leq 0$ we reject the null hypothesis, since 2.429 is greater than 1.6621. In this example, treatment A fails, but treatment B is superior to Placebo. For one arm to stop early, one would have to reject $H_{AB}$ and then $H_A$ or $H_B$. The trial can still continue the other arm. If only $H_{AB}$ is rejected, but not $H_A$ or $H_B$ the trial can still continue.

In order to ensure that the familywise type I error rate is strongly controlled, the following sources of multiplicity must be constrained. In this example, multiplicity due to treatment groups within stage was adjusted via Simes' test. Multiplicity due to the interim looks was controlled via the (O'Brien-Flemming type) alpha-spending function. Multiplicity due to treatment groups across stages was controlled via closed testing methods. Multiplicity due to the sample size readjustment was controlled via the adaptive combination test. See Figure 4-1.

## Figure 4-1. Group Sequential Testing Method when Adding a Treatment Arm Mid-Study


STAGE 1
Use stage 1 Placebo and Treatment A data for $H_A$
2-sample *t*-test (computed at 1st look)
STAGE 2
Use stage 2 Placebo and stage 2 Treatment A for $H_A$
Use stage 2 Placebo and stage 2 Treatment B for $H_B$
2-sample *t*-tests, Simes' method
Interim Analysis
Combination test of $H_{AB}$:
Either stop early or re-estimate the sample size and continue to stage 3.
STAGE 3
Use stage 3 Placebo and stage 3 Treatment A for $H_A$
Use stage 3 Placebo and stage 3 Treatment B for $H_B$
2-sample *t*-tests, Simes' method
Final Analysis
Combination test of $H_{AB}$

### 4.3 Simulations

*Simulation Study*

Monte Carlo simulation was used to compare the power and type I error rate for testing $H_{AB}$, $H_A$, and $H_B$ in a single-stage (fixed sample) clinical trial. The following statistical analysis approaches were compared: 1. two-sample *t*-test with the data pooled across stages; 2. Linear model adjusting for a fixed stage/cohort effect; 3. Inverse Chi-Square (Fisher's) adaptive combination test[38]; 4. Weighted-Inverse Normal adaptive combination test[40]. See Figure 4-2.

Samples were drawn from the multivariate normal distribution assuming that

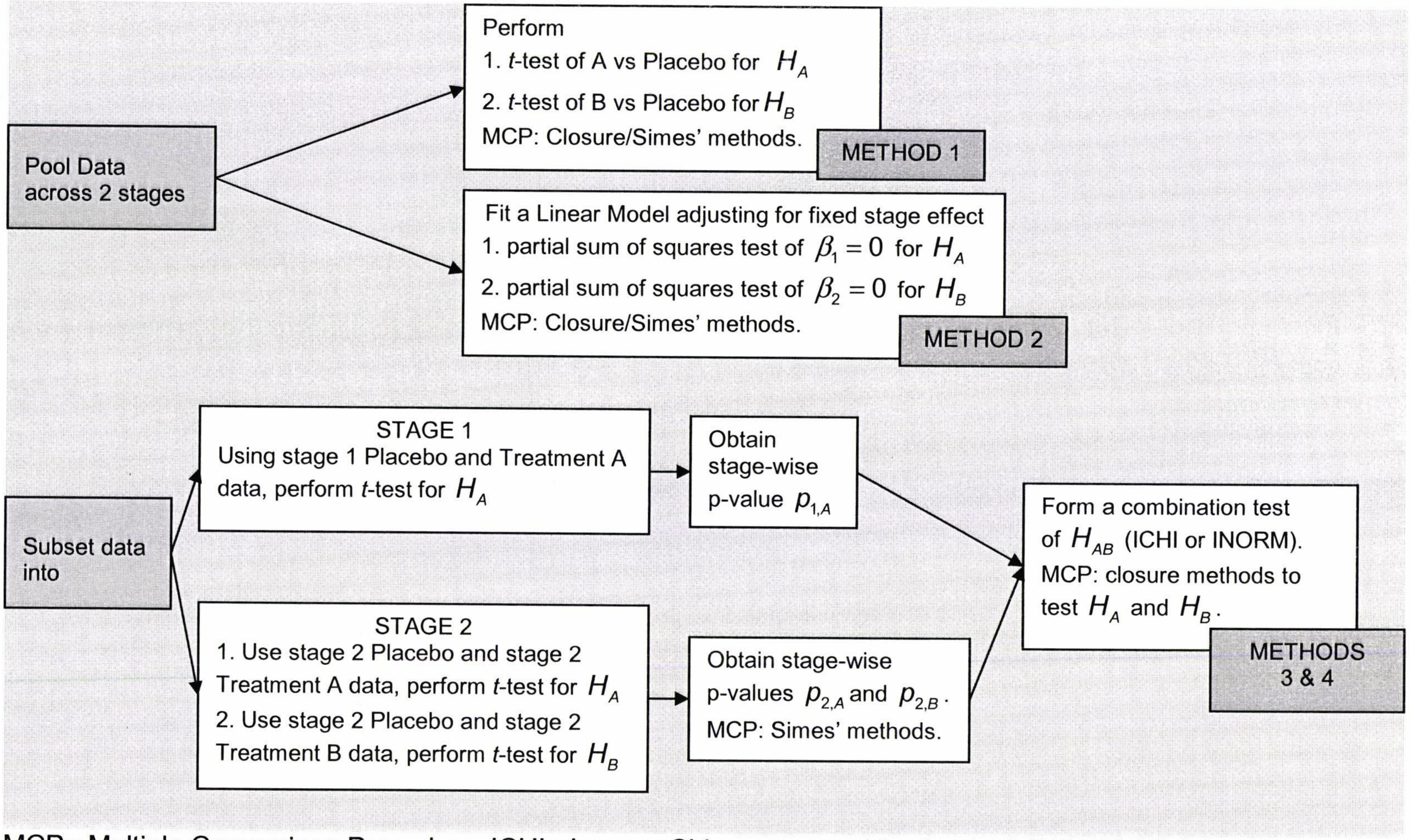
$$\begin{aligned}
&A_i \sim N(\mu_A,\sigma^2),\ i=1,\ldots,m\\
&A_j \sim N(\mu_A,\sigma^2),\ \mathrm{j}=m+1,\ldots,n\\
&B_k \sim N(\mu_B,\sigma^2),\ k=1,\ldots,n\\
&P_i \sim N(\mu_P,\sigma^2)\\
&P_j \sim N(\mu_P,\sigma^2)
\end{aligned}$$

$$\underline{Y} \sim MVN(\underline{\mu},\Sigma),\quad \underline{Y}=\begin{pmatrix} A_i\\ A_j\\ B_k\\ P_i\\ P_k \end{pmatrix}\quad \Sigma=\begin{pmatrix} \sigma^2+\sigma_u^2 & \cdots & \sigma_u^2\\ \vdots & \ddots & \vdots\\ \sigma_u^2 & \cdots & \sigma^2+\sigma_u^2 \end{pmatrix}$$

where $A_i$ and $A_j$ are the observations from treatment group A from stage 1 and stage 2 respectively, $B_k$ are the observations from treatment group B from stage 2, $P_i$ and $\mathrm{P}_j$ are the observations from Placebo group from stage 1 and stage 2, respectively, $n$ is the sample size needed per group for a two sample $t$-test of $H_A$ or $H_B$, and $\sigma_u^2$ is the covariance between observations clustered within the same stage (i.e. cohort effect). The sample size per group was set to $n$=120 for detecting an effect size $\theta = 0.38$, at $\alpha = 0.05$ (one-sided) and power=0.90. The data were simulated assuming different sizes of $\sigma_u^2$ $\{0, \frac{1}{4}\theta, \frac{1}{2}\theta, \theta\}$ and $t_P$ {0.1, 0.3, 0.5}, where $0 < t_P < 1$ is the time that the design change is made in terms of the fraction of the sample size. One analysis is performed at the end of the study (with no interim analyses). For the adaptive combination tests, at the end of the study the data were subset into stages 1 and 2 prior to forming the combination test.

For all approaches closed testing multiple comparisons procedures were applied, as this approach is more powerful than single-step approaches, but do not provide confidence intervals[63]. For consistency across methods, Simes' Modified Bonferroni procedure[19] was used to obtain an adjusted p-value for the global null hypothesis $H_{AB}$. All simulations were done in STATA. The number of replications was set at 6000. With 6000 replications, a two-sided 99% CI around the expected type I error rate of 0.025 will extend ±0.005, and a two-sided 99% CI around for the expected power of 0.90 will extend ±0.01.

**Figure 4-2. Several Methods to test $H_{AB}$, $H_A$, and $H_B$**


Pool Data across 2 stages
Perform
1. $t$-test of A vs Placebo for $H_A$
2. $t$-test of B vs Placebo for $H_B$
MCP: Closure/Simes' methods.
METHOD 1
Fit a Linear Model adjusting for fixed stage effect
1. partial sum of squares test of $\beta_1 = 0$ for $H_A$
2. partial sum of squares test of $\beta_2 = 0$ for $H_B$
MCP: Closure/Simes' methods.
METHOD 2
Subset data into
STAGE 1
Using stage 1 Placebo and Treatment A data, perform $t$-test for $H_A$
STAGE 2
1. Use stage 2 Placebo and stage 2 Treatment A data, perform $t$-test for $H_A$
2. Use stage 2 Placebo and stage 2 Treatment B data, perform $t$-test for $H_B$
Obtain stage-wise p-value $p_{1,A}$
Obtain stage-wise p-values $p_{2,A}$ and $p_{2,B}$.
MCP: Simes' methods.
Form a combination test of $H_{AB}$ (ICHI or INORM).
MCP: closure methods to test $H_A$ and $H_B$.
METHODS 3 & 4

MCP= Multiple Comparison Procedure, ICHI= Inverse Chi-square combination test, INORM=Weighted Inverse Normal combination test.

## 4.4 Results

Table 4-1 shows the results for the type I error rate for the four methods. For all methods the familywise type I error rate is controlled strongly when covariance $\sigma_u^2=0$. However, the *t*-test (with data pooled) results in anticonservative p-values when $\sigma_u^2>0$, as high as 0.24 when the design change is made after 50% of the sample size for arms A and Placebo have already been enrolled. The inflation of the familywise error rate is due to the Treatment B versus placebo comparison, rather than the Treatment A comparison (results not shown).

**Table 4-1. Family-Wise Type I error rates for the 4 Methods (Fixed Analysis) results of 6000 simulations ($\mu_A = \mu_B = \mu_P$)**

| Tp | Covariance | *t*-test (Pool Data) | Linear Model (adjusting for cohort) | Inverse Chi-Square Combination test (ICHI) | Weighted Inverse Norm Comb Test (INORM) |
|---|---|---|---|---|---|
| 0.1 | $\sigma_u^2=0$ | 0.047 | 0.047 | 0.051 | 0.045 |
| | $\sigma_u^2>0$ | 0.06 | 0.048 | 0.046 | 0.047 |
| 0.3 | $\sigma_u^2=0$ | 0.044 | 0.046 | 0.046 | 0.045 |
| | $\sigma_u^2>0$ | 0.155 | 0.047 | 0.046 | 0.046 |
| 0.5 | $\sigma_u^2=0$ | 0.046 | 0.046 | 0.044 | 0.043 |
| | $\sigma_u^2>0$ | 0.242 | 0.046 | 0.044 | 0.044 |

NOTE: FWE= Probability of rejecting any component null hypothesis $H_{AB}$, $H_A$, or $H_B$ given they are all true. $\sigma_u^2>0$ is averaged across $\sigma_u^2=\{0,\frac{1}{4}\theta,\frac{1}{2}\theta,\ \theta\}$

Figures 4-3, 4-4, and 4-5 give the results for the empirical power when both treatments are superior to placebo. In general, when the time at which Arm B is added in terms of the fraction of the sample size is early ($t_P$=0.1), then all methods obtained nearly 90% power (the nominal level). When there is no stage effect ($\sigma_u^2 = 0$), then the pooled method has the highest power for both $H_A$ and $H_B$. The linear model method loses power as $t_P$ increases for the test of $H_B$, but is still superior to the two combination tests. Fisher's combination test (ICHI) performs worst for both $H_A$ and $H_B$. (See Figure 4-3.)

**Figure 4-3. Power for tests of $H_A$, $H_B$ when $\sigma_u^2 = 0$, $\mu_A = \mu_B > \mu_P$, $n_A$=$n_B$=$n_P$**

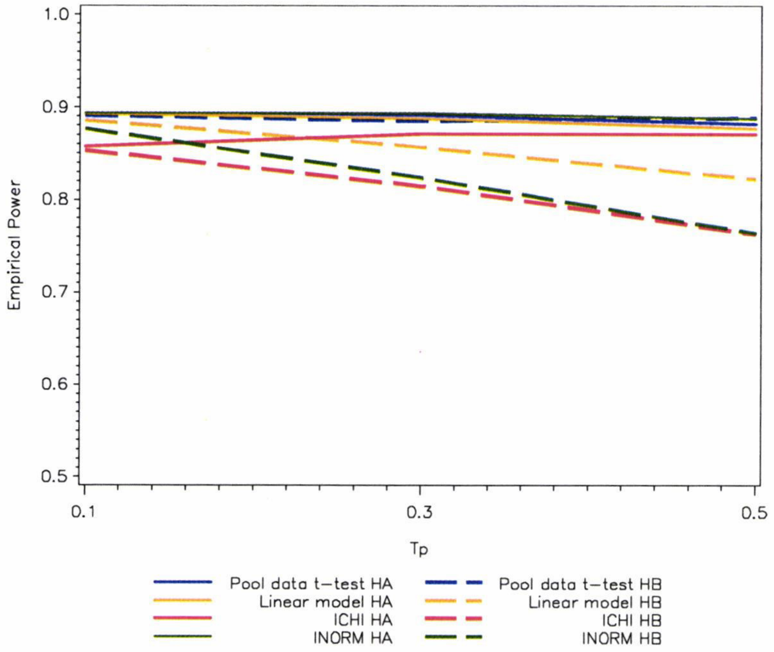
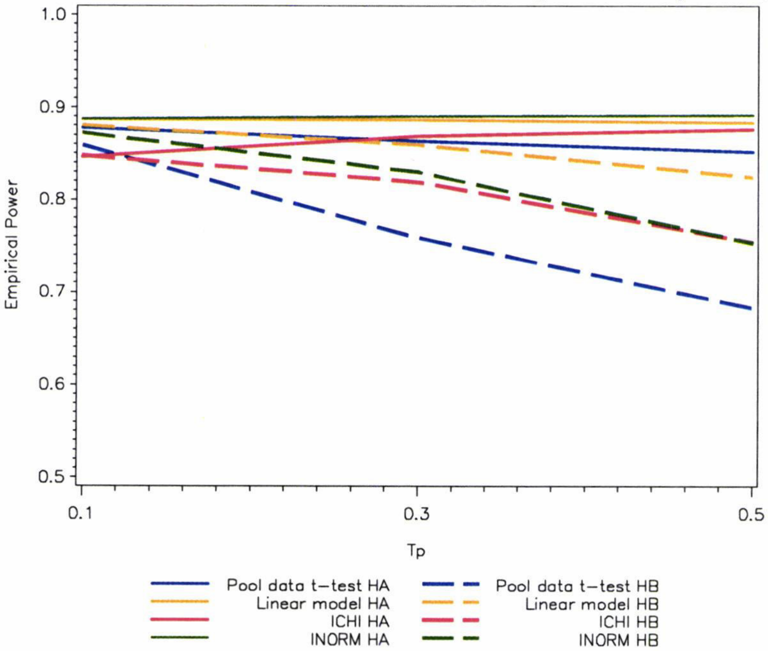
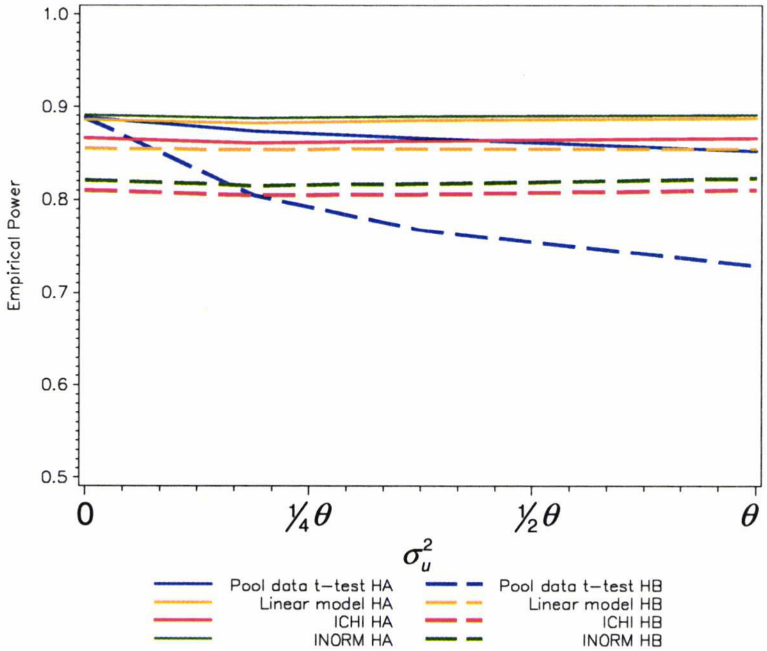


**Figure 4-4. Power for tests of $H_A$, $H_B$ when $\sigma_u^2 > 0$, $\mu_A = \mu_B > \mu_P$, $n_A$=$n_B$=$n_P$**



When the within stage (intracluster) covariance $\sigma_u^2 > 0$, then the pooled method performs worst for both $H_A$ and $H_B$ across all time points ($t_P$), while the linear model approach is superior. (See Figure 4-4.) Averaging across all time points ($t_P$), the amount of covariance does not affect the power for the linear model or the combination tests, but power decreases as covariance increases for the pooled method. (See Figure 4-5.)

**Figure 4-5. Power by Covariance for tests of $H_A$, $H_B$ averaged across different Tp when $\mu_A = \mu_B > \mu_P$, $n_A=n_B=n_P$**



## 4.5 Conclusions

Adding an arm to an ongoing clinical trial is a savings in sample size because fewer patients are randomized to placebo compared to two trials with separate placebo groups (A vs. Placebo and B vs. Placebo). When there is no correlation among the observations within stage, then power is highest when the data are simply pooled (from before and after the addition). When a new treatment arm (B) is added to an ongoing clinical trial (of A versus placebo), the placebo group will be different before and after the design change. The whole Placebo group received Placebo for drug A. However, this is not the case for the B versus Placebo comparison. The placebo group in the first stage receives only the placebo for treatment A, while the placebo group in the second stage receives placebo for treatment A and treatment B. Thus, it is questionable whether it would be appropriate to pool the placebos from before and after the design change, since not all Placebo subjects received the Placebo for B. This may not be a concern when Treatment B is an increased dose of Treatment A (given the placebo looks the same and is taken in the same daily frequency). However, given the situation in which Treatment B looks different than A, there may be ambiguity about the conclusions if the data are simply pooled. There is the potential that the original treatment versus placebo comparison could be perturbed due to enthusiasm over the new drug or a cohort effect[15].

From a statistical standpoint, these simulations show that pooling the data without any adjustment for stage is unadvisable. When the observations within stage are correlated, then pooling the data across stages will result in a loss of power for both treatment group comparisons, but particularly for the treatment B comparison. Furthermore, pooling the data ignores the intra-stage correlation and will inflate the type I error rate, since the estimate of the variance of the mean is underestimated.

When there is no intra-stage correlation, the power for the linear model approach and the weighted-inverse normal adaptive combination tests were only slightly less than the nominal level for the test of Treatment A. Since the adaptive tests are not based on sufficient statistics, this small loss of power was expected.

In all cases Fisher's combination test is outperformed by the weighted-inverse normal combination test. Both the fixed-effect linear model method and the adaptive method control the type I error rate in the presence of a random stage effect. When the design change is made after only a small proportion of subjects have been enrolled, $t_P$ , (e.g. 10%), then the weighted-inverse adaptive method has similar power to the linear model approach. However, both of these methods lose power for testing $H_B$ as $t_P$ increases, especially the adaptive test (INORM). For both the linear model and the adaptive methods, these results show that the sample size for the Treatment B versus placebo comparison would need to be increased in order to achieve the desired power. For the test of $H_B$ , the linear model is more powerful, but adaptive allows for more flexibility to re-estimate the sample size mid-study.

At what point do the cost savings of introducing a new treatment arm into an ongoing trial rather than starting a new trial become negligible? These simulations did not consider the case when more than 50% of patients have already been enrolled. As we have seen, there is a loss of power by as high as 15 percentage points (75% rather than 90% power) for the test of $H_B$ as the proportion of patients already enrolled increases to 50%, when an adaptive combination test is applied (regardless of the covariance).

The reduction in power we observed in these simulations is partially due to the use of non-standard test statistics (adaptive methods), but primarily due to the use of unequal sample sizes since only the stage 2 placebo is used to test $H_B$. One solution to attain the desired power for the linear model or adaptive approaches is to enroll $n$ new patients into the stage 2 placebo group. Further simulations (see Appendix 4-2) indicate that this will mitigate the loss of power for the INORM and linear model approaches as $\mathrm{t}_\mathrm{P}$ increases. We can see that enrolling $\mathrm{t}_\mathrm{P} \times n + n$ subjects into the placebo group is still less than the $2 \times n$ subjects enrolled into placebo across two separate trials. While it is undesirable to have a total number of subjects enrolled into placebo greater than either treatment arm, the probability of being allocated to the placebo group need not be greater than the chance of allocation to a treatment arm at any point in time. Further research is needed to adequately address approaches to increase the sample size, including the potential benefit of re-estimating the sample size. If an adaptive test is to be used, the sample size could be re-estimated using the observed effect size from stage 1 for the treatment A versus Placebo comparison, without any risk of inflating the type I error rate.

There may be other considerations for selecting between methods. Adaptive methods began to be developed in the 1990s, but have only recently come into practical use. In general, mid-course design changes driven by internal information are more controversial than those driven by external information. Internal information refers to information within the study (such as the interim effect size or a sub-group analysis) that prompts a change in the design (change in sample size, change in target population, etc.). However, when the design change is made based on observed data, then the type I error for the final (pre-specified) analysis will be inflated. By applying adaptive methods (e.g. combination rule) then the pre-specified alpha can be constrained.

When external information (independent from the study) prompts a design change for a study, then the integrity of the study may be questioned because it is difficult to prove that internal information did not play a role in the design change. Thus adaptive methods are also suitable for this situation. In this case, adding a new treatment arm to an ongoing study is an example of a design change based on external information. In this context, the type I error probability will not be inflated due to an internal look at the data since the design change is externally-driven (not involving an unplanned look). However, if a new dose, or dose regimen, is added to an ongoing trial, it may be advisable to adjust for the change in design by a combination test, because it is difficult to show that inside knowledge of the treatment effect for the current dose, did not impact the decision to increase the dose, unless perhaps the treatment arm that is added is a lower dose.

Nevertheless, there are situations where a regression approach (adjusting for stage/cohort) is perfectly acceptable. One such scenario is when the arm that is added is an active comparator. For instance a clinical trial may be initiated in the US, and shortly thereafter investigators may realize that in order to meet European Regulatory authority (EMEA) approval, the current study treatment would need to be shown non-inferior to the standard of care in Europe. Then, it may make sense to modify the ongoing trial to add this standard as another arm. In this case the investigators wish to show study treatment superiority versus placebo (for the FDA) and non-inferiority of the study treatment versus the standard (for the EMEA). In this case, since the arm that is added is a standard and the rationale is clearly externally driven, there seems to be no need to adopt an adaptive analysis approach. An approach such as that given by Denne and Koch would be well suited for this scenario since the hypotheses (superiority, non-inferiority) are nested.[64] In the context of non-adaptive designs, Denne and Koch have shown that it is possible to test both non-inferiority and superiority sequentially without adjusting for multiplicity because they are nested hypotheses; closed testing methods are applied.[23,46]

When the decision is made to add an arm to an ongoing clinical trial, the original design considerations must be taken into account. According to Follman *et al.* there should be equal criteria used to evaluate each treatment arm[1]. In order to add a treatment arm, the protocol must be amended. The timing of the protocol re-design may affect the degree to which design changes can be made. It may be more difficult to re-design a trial after an interim analysis has been performed.

When adding another treatment arm, it is important to adjust the randomization allocation ratio to ensure that all three treatment arms complete enrollment at roughly the same time. Firstly and above all, this is necessary to ensure blinding, such that the last patients enrolled are not all receiving treatment B. Secondly, if the randomization allocation ratio is 1:1:1, then, once all patients are enrolled into treatment A, there is a possibility of a observing a third cohort set, who are all enrolled in treatment arm B, having different characteristics than those patients in stage 1 or stage 2.

These results suggest that the linear model method will always outperform the adaptive methods as $t_P$ increases, even in group sequential setting, unless the effect size is smaller than expected for one or both treatment arms or the variance is larger than expected. In this case an adaptive method allows the sample size to be increased due to internal information observed at the first interim analysis, and thus the desired power can be attained.

In Parkinson's disease clinical trials, there is a chance that there will be increased enthusiasm about a new drug. Publicity concerning phase II studies for certain drugs can impact the rate of enrollment into a Phase III study of the same drug, and this is likely to have an impact on subjective self-reported outcome measures. For Parkinson's disease clinical trials the cohort effect may be a legitimate concern. Prior NET-PD studies have shown that changes in clinical practice over time have a major impact on outcome measures[6,7]. Another important point is that $t_P$ and the covariance within stage are likely to increase together. That is, the longer you wait to add the new treatment arm, the higher the $t_P$ will be and the more likely things are to have changed (new standards of care, etc.) introducing cohort effects.

The NET-PD investigators continue to pursue Phase II trials of additional agents for the treatment of Parkinson's disease. If new agents become ready for Phase III testing by the NET-PD group, design modification to the current ongoing Phase III trial is a possibility. These results suggest that it would be inadvisable to simply pool the treatment groups across the stages, since the type I error will be inflated and power will decrease if the intra-stage correlation is greater than zero. In the presence of possible intra-stage correlation, the linear model approach is more powerful, but the adaptive method allows for more flexibility to re-estimate the sample size. Both analysis approaches (regression and adaptive) control the type I error rate when no internal study information is used in the decision to add a new treatment arm mid-study.

**Appendix 4-1. Simulated Power for tests of $H_{AB}$, $H_A$, $H_B$ (given $\mu_B = \mu_A > \mu_P$ is true and $n_A$=$n_B$=$n_P$=120 after 6000 replications)**

| | | Pooled Placebo | | | Linear Model | | | Fisher's Combination test | | | Weighted Inverse Norm Comb Test | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $t_P$ | Covariance $\sigma_u^2$ | $H_{AB}$ | $H_A$ | $H_B$ | $H_{AB}$ | $H_A$ | $H_B$ | $H_{AB}$ | $H_A$ | $H_B$ | $H_{AB}$ | $H_A$ | $H_B$ |
| 0.1 | 0 | 0.942 | 0.893 | 0.891 | 0.943 | 0.893 | 0.886 | 0.917 | 0.857 | 0.853 | 0.945 | 0.894 | 0.877 |
| | ¼$\theta$ | 0.937 | 0.881 | 0.878 | 0.934 | 0.882 | 0.876 | 0.908 | 0.843 | 0.847 | 0.936 | 0.885 | 0.869 |
| | ½$\theta$ | 0.933 | 0.878 | 0.859 | 0.94 | 0.888 | 0.881 | 0.909 | 0.845 | 0.844 | 0.94 | 0.888 | 0.868 |
| | $\theta$ | 0.93 | 0.875 | 0.838 | 0.94 | 0.891 | 0.884 | 0.909 | 0.85 | 0.851 | 0.942 | 0.889 | 0.878 |
| 0.3 | 0 | 0.939 | 0.891 | 0.885 | 0.932 | 0.888 | 0.857 | 0.92 | 0.871 | 0.815 | 0.938 | 0.894 | 0.824 |
| | ¼$\theta$ | 0.925 | 0.872 | 0.8 | 0.929 | 0.882 | 0.86 | 0.917 | 0.863 | 0.818 | 0.934 | 0.886 | 0.829 |
| | ½$\theta$ | 0.92 | 0.866 | 0.758 | 0.936 | 0.889 | 0.861 | 0.924 | 0.87 | 0.819 | 0.941 | 0.892 | 0.83 |
| | $\theta$ | 0.912 | 0.85 | 0.718 | 0.935 | 0.888 | 0.857 | 0.924 | 0.873 | 0.819 | 0.939 | 0.892 | 0.829 |
| 0.5 | 0 | 0.941 | 0.882 | 0.889 | 0.924 | 0.877 | 0.823 | 0.922 | 0.871 | 0.763 | 0.931 | 0.888 | 0.764 |
| | ¼$\theta$ | 0.925 | 0.868 | 0.734 | 0.928 | 0.885 | 0.826 | 0.927 | 0.878 | 0.747 | 0.936 | 0.892 | 0.747 |
| | ½$\theta$ | 0.915 | 0.854 | 0.684 | 0.927 | 0.88 | 0.824 | 0.926 | 0.875 | 0.754 | 0.938 | 0.891 | 0.753 |
| | $\theta$ | 0.9 | 0.834 | 0.63 | 0.928 | 0.885 | 0.824 | 0.927 | 0.876 | 0.761 | 0.935 | 0.894 | 0.761 |

$t_P$ is the time at which design change is made (as a proportion of total sample size enrolled). $H_{AB}: H_A \cap H_B$, $H_A: \theta_A \leq 0$, $H_B: \theta_B \leq 0$ where $\theta_A = \mu_A - \mu_P$ and $\theta_B = \mu_B - \mu_P$. The results are given for different sizes of intra-stage covariance $\sigma_u^2$ $\{0, \frac{1}{4}\theta, \frac{1}{2}\theta, \theta\}$, where $\theta$=0.38 is the true effect size.

**Appendix 4-2. Simulated Power for tests of $H_{AB}$, $H_A$, $H_B$ (given $\mu_B = \mu_A > \mu_P$ is true and n$_A$=n$_B$=120 and n$_P$=t$_P$x 120+120 after 6000 replications)**

| | | | Pooled Placebo | | | Linear Model | | | Fisher's Combination test | | | Weighted Inverse Norm Comb Test | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $t_P$ | Cov $\sigma_u^2$ | $n_P$ | $H_{AB}$ | $H_A$ | $H_B$ | $H_{AB}$ | $H_A$ | $H_B$ | $H_{AB}$ | $H_A$ | $H_B$ | $H_{AB}$ | $H_A$ | $H_B$ |
| 0.1 | 0 | 132 | 0.963 | 0.921 | 0.918 | 0.963 | 0.920 | 0.913 | 0.941 | 0.886 | 0.888 | 0.962 | 0.920 | 0.905 |
| | $\theta$ | 132 | 0.960 | 0.908 | 0.886 | 0.962 | 0.918 | 0.912 | 0.942 | 0.884 | 0.892 | 0.963 | 0.918 | 0.909 |
| 0.3 | 0 | 156 | 0.976 | 0.938 | 0.933 | 0.975 | 0.936 | 0.923 | 0.969 | 0.920 | 0.906 | 0.976 | 0.937 | 0.911 |
| | $\theta$ | 156 | 0.982 | 0.872 | 0.796 | 0.974 | 0.933 | 0.920 | 0.966 | 0.916 | 0.902 | 0.974 | 0.934 | 0.907 |
| 0.5 | 0 | 180 | 0.977 | 0.938 | 0.939 | 0.974 | 0.930 | 0.915 | 0.970 | 0.919 | 0.889 | 0.974 | 0.932 | 0.888 |
| | $\theta$ | 180 | 0.991 | 0.814 | 0.721 | 0.972 | 0.925 | 0.916 | 0.971 | 0.919 | 0.896 | 0.976 | 0.930 | 0.895 |

t$_P$ is the time at which design change is made (as a proportion of total sample size enrolled). $H_{AB} : H_A \cap H_B$, $H_A : \theta_A \leq 0$, $H_B : \theta_B \leq 0$ where $\theta_A = \mu_A - \mu_P$ and $\theta_B = \mu_B - \mu_P$. The results are given for different sizes of intra-stage covariance $\sigma_u^2$ $\{0, \frac{1}{4}\theta, \frac{1}{2}\theta, \theta\}$, where $\theta$ =0.38 is the true effect size. n$_A$, n$_B$, and n$_P$ denote the total per group sample size for treatment groups A, B, and Placebo, respectively.

## CHAPTER 5: Paper 3: Adding a Treatment Arm to an Ongoing Clinical Trial: An Illustrated Example Using NET-PD clinical trial data

Jordan J. Elm, Yuko Y. Palesch, Barbara C. Tilley, Wenle Zhao, Vanessa Hinson, Bernard Ravina

### Abstract

**Background**: In settings where clinical trials require many subjects and long-term follow-up, evaluating more than one drug against a common placebo would be efficient. Yet, in practice, drugs may be at different stages of development (e.g. Phase I, Phase II, or Phase III). We have previously proposed analytical methods for the addition of newly screened drugs into an ongoing clinical trial. **Methods**: The NET-PD FS1 and FS-TOO clinical trials (with identical study designs conducted concurrently) were re-analyzed as a single study where the treatment arms (from FS-TOO) were added to an ongoing study (FS1). For each treatment arm, a futility hypothesis was analyzed using: (1) two-sample *t*-test with the data pooled across stages; (2) linear model adjusting for a fixed cohort effect; (3) inverse Chi-square (Fisher's) combination test[38]; and (4) Weighted-Inverse Normal combination test[40]. The probability of rejecting the null hypothesis was compared for the different testing methods using bootstrapping. **Results**: The linear model and the combination tests performed similarly and were most consistent with the results from the original analysis of these data as separate studies.

**Conclusions:** If a treatment arm is to be added into an ongoing study, then this example illustrates that either a linear model approach or an adaptive combination test is well suited for such a situation.

## 5.1 Introduction

In settings where clinical trials require many subjects and long-term follow-up, it is more efficient to test more than one drug against a common placebo to maximize the chances of finding a successful drug and to minimize the number of subjects randomized to placebo. In practice, several agents of interest may be at different stages of development (e.g. dose finding, screening for futility, or evaluation for efficacy). This is the situation for the NINDS Exploratory Trials in Parkinson's Disease (NET-PD) Program funded by the National Institutes of Neurological Disorders and Stroke (NINDS). During 2003 to 2005, the NET-PD Program sponsored a series of Phase II clinical trials[6,7] of drugs aimed at slowing the clinical progression of PD. In 2007 the first NET-PD large study (LS-1) trial began enrollment of PD patients into a 5-year double-blind Phase III study of creatine versus placebo. The intent of the NET-PD Program is to add to the LS-1 any promising treatments as they become available for Phase III testing, rather than initiate a new large Phase III trial for each treatment. The main savings would be in terms of the number of subjects enrolled into the placebo group that would be used as the referent group for all new treatments. As such, a flexible clinical trial design that is able to accommodate adding treatment arms at different times would be useful.

In order to meet regulatory approval, a clinical trial needs to be conducted following the Good Clinical Practice (GCP) Guidelines and the Food and Drug Administration (FDA) Code of Federal Regulations[76].

Accordingly, in principle, the key features of clinical trial designs (e.g., hypotheses, primary outcome, treatments) should not change once the study has begun. In reality, not all parameters are known when planning a trial, and if the trial is planned with incorrect assumptions about these parameters, the chance of success can be diminished. It may be logistically and/or economically efficient to modify a trial protocol once new information is known. However, these practicalities should not eclipse the potentially adverse effect of design changes on the validity of statistical inference. Some argue that any modification to the protocol needs to be accounted for in the method of analysis since protocol modifications can introduce bias to the conclusions[15,57].

Recent statistical developments of adaptive clinical trial methodology allow investigators some flexibility to modify the study during its course. There now exists a class of adaptive clinical trial designs that encompass various situations (including but not limited to group sequential design, sample size re-estimation design, adaptive dose escalation design, and phase II/III designs). For a simple, single design change such as adding a treatment arm, there are two issues of concern: (1) the inflation of the type I error rate; and (2) the introduction of bias in the scale or shift parameters. Adaptive analytical methods make it possible to change a key feature of the design, and still maintain adequate control of the type I error rate. In the 2006 conference on Adaptive Trial Design, the FDA has indicated its support of such adaptive/flexible designs for some situations and currently is developing guidance documents.[65] In practice, adaptive methods need to be pre-specified rather than applied in an *ad hoc* fashion. In order to gain regulatory approval of the design, the pre-specified adaptive protocol should anticipate and include all possible scenarios.[65] A program such as NET-PD which is conducting a series of clinical trials, at various phases, should design a protocol that allows for adding treatment arms to the ongoing trial, should they become available.

Although many authors have considered multi-armed studies where one or more treatment arms are dropped[1,2,50,77-82], the process of adding a treatment arm into an ongoing clinical trial has received little attention in existing literature. Adaptive combination methodology can be applied for modifying hypotheses mid-study, in particular, for the case of changing the *a priori* ordering of hypothesis or adding a new hypothesis[49]. As such, we can conceptualize adding a treatment arm as a type of adaptive design, in which a new hypothesis is added to the study mid-course.

One type of adaptive methodology applicable to many types of design changes is the p-value combination test. This method controls the type I error rate after a design modification and would be suitable when a treatment arm has been added to a trial. No distributional assumptions for the endpoints are required. The p-values for the data from 2 stages (before and after the design change) or *k* stages are combined across stages[39].

To illustrate the adaptive design methodology for adding a treatment arm to an ongoing trial, we use the data from the NET-PD FS1 and FS-TOO studies. These studies had the same entry criteria, primary outcomes, and were conducted concurrently at more than 40 clinical sites participating in the

NET-PD Program. Enrollment into the first pilot study (FS1) occurred between 5/2003-9/2003 and into the second pilot study (FS-TOO) between 3/2004-7/2004, and subjects were followed-up every 3 months for at least 12 months. The aim of each pilot study was to evaluate the futility of the active treatments. In each study, PD subjects were randomized into one of two active arms or a concurrent placebo. In FS1, 200 subjects were randomized to creatine (n=67), minocycline (n=66), or placebo (n=67). In FS-TOO,[5] 213 subjects were randomized to Coenzyme $Q_{10}$ (n=71), GPI-1485 (n=71), or placebo (n=71). In the original study design, the primary outcome measure was the change from baseline score at 12-months in the Unified Parkinson's Disease Rating Scale (UPDRS). Each active arm was (individually) compared to a pre-specified futility threshold in a one-sample test. Full descriptions of these clinical trials are available.[5-7]

For the primary analysis, each treatment arm was compared to the futility threshold that was obtained from the placebo group data from a large PD study conducted in 1996 -1998, Deprenyl And Tocopherol Antioxidative Therapy Of Parkinsonism (DATATOP). Additionally, both studies included a small placebo group strictly as a calibration control[83]. In other words, the treatment and placebo groups were not compared head to head. One of the major issues in these studies was the shift in the placebo data across studies. In both studies, the mean UPDRS change score in the concurrent calibration placebo group fell outside the 95% confidence interval (CI) for the DATATOP placebo change, suggesting a significant difference. Moreover, a marked difference in the change score was observed between the FS1 and FS-TOO placebo groups.

Because of the similarities in these two studies and since the primary manuscript for FS-TOO included a re-analysis of FS-1, these studies provided an example for the scenario of adding treatment arms to an ongoing study. For the purposes of this paper, we considered the FS-TOO study as if it were added to the FS1 study, rather than a separate study. We examined the impact of adding the three FS-TOO treatment arms into the FS1 study on the probability of rejecting the null hypothesis for each treatment arm comparison versus concurrent placebo under four methods of analysis. Our primary objective was to demonstrate how the testing methods can be applied in a real setting and to examine how these results compared to the original findings for each treatment. This exercise retains the futility hypothesis as the primary research question.

## 5.2 Methods

First, in order to study whether the design modification (of adding FS-TOO treatment arms to FS1) impacted the target population, we estimated the sensitivity index using FS1 and FS-TOO data. The sensitivity index, a measure of change in the signal-to-noise ratio in the actual population studied versus the original target population[84.p.27], can be estimated as $\hat{\Delta} = \frac{1+\hat{\varepsilon}/\hat{\mu}}{\hat{C}}$ where $\hat{\varepsilon} = \hat{\mu}_{Actual} - \hat{\mu}$ and $\hat{C} = \hat{\sigma}_{Actual} / \hat{\sigma}$. If the modification does not affect the target population then $\Delta$ should equal 1.

In the original study design, each individual treatment arm was compared to a fixed futility threshold of 7.5 points in the UPDRS (defined as a ~3 point decrease in the placebo group's annual change of 10.65 observed in DATATOP). For the purposes of this illustration, we performed a *two-sample t-test* where each treatment was compared to a futility threshold defined as a 3 point decrease in the *concurrent* placebo (rather than the DATATOP placebo). The two sample hypothesis was written as $H_0$: $\mu_i \leq \mu_p - \delta'$ versus $H_a$: $\mu_i > \mu_p - \delta'$ where $\delta'$ was defined as 3 UPDRS points as in the original study design (for UPDRS change a lower score is better).

With the given sample size of 67-71 per group, a two sample *t*-test with a 0.10 one-sided significance level has 75% power to detect a difference in means of 3 assuming that the common standard deviation is 9. The type I error rate was set at 0.10 as in the original studies. The stage 1 data were conceptualized as the FS1 study and the stage 2 data were the FS-TOO study.

*Methods of Analysis*

In Chapter 4, we introduced the testing methods one could use when a treatment arm(s) has been added into an ongoing clinical trial. One could (naively) pool the data across stages (before and after an arm has been added) and perform the usual test statistic from a two sample *t*-test. Better yet, one could adjust for the stage effect in a regression (e.g. as a fixed effect), hereafter referred to as the linear model method.

Alternatively, one could take into account this design adaptation in the construction of the test statistic by combining the p-values from the two stages (before and after an arm has been added). There are several ways to define a two-stage test statistic. The most commonly used are: (1) the product of the stagewise p-values (from stages 1 and 2) which is an inverse $\chi^2$ test; and (2) the weighted inverse normal test.

The inverse $\chi^2$ test (ICHI) is formed as follows:

$T_k = \prod_{k=1}^{2} p_k \leq c_\alpha = \exp[-\frac{1}{2}\chi^2_{2k,1-\alpha}]$ where $p_1$ and $p_2$ are the p-values from the first and second stages of data and $\chi^2_{2k}(1-\alpha)$ is the $(1-\alpha)$-quantile of the central $\chi^2$ distribution with $2k$ degrees of freedom[38,39]. The weighted inverse normal test (INORM) is formed as follows: $T_k = \sum_{k=1}^{2} w_k \cdot \Phi^{-1}(1-p_k)$ where $\sum_{k=1}^{2} w_k^2 = 1$ and $w_k = \left(\sum_{k=1}^{2} n_k\right)^{-1/2} \cdot \sqrt{n_k}$ where $n_k$ is the total number of observations at the $k^{th}$ stage. The overall null hypothesis $H$ is rejected at level alpha if $T_k > c$ where $c$ is the classical group sequential stopping boundary[40-42,57, p.101.]

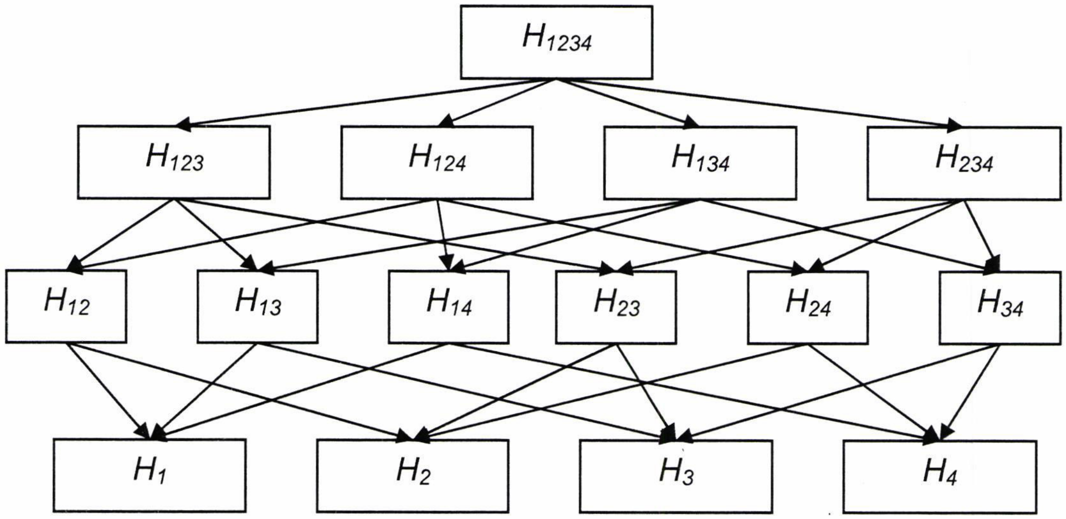
Since the original trial type I error rate was 10%, then for the INORM combination test *c*=0.10, as we will not consider early stopping after stage 1 (at the time of addition of FS-TOO arms). For the ICHI combination test, the 10% type I error rate corresponds to $\alpha_1$= 0.02045 and $\alpha_2$=0.02045 where $\alpha_1$

and $\alpha_2$ are the critical upper limits for stages 1 and 2 (stopping boundaries). (For the ICHI testing procedure, by construction, if the stage 1 p-value is less than $\alpha_2$, then the null hypothesis is rejected at stage 1). For the ICHI combination test, the p-value of the final analysis (for the 2-stage combination test) is given by $p = p_1$ if $p_1 < \alpha_1$ or $p = \alpha_1 + p_1 p_2 \ln\left(\frac{1}{\alpha_1}\right)$ otherwise, where $p_1$ and $p_2$ are the observed p-values for the *t*-tests at stage 1 and stage 2, respectively.

### *Control of Multiplicity*

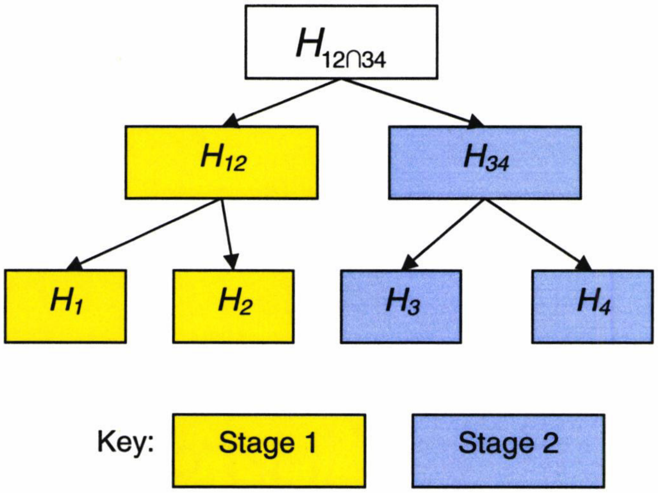
Since there are four treatment-to-placebo comparisons, closed testing methods and Simes' method were used to control the familywise error (FWE) rate strongly at $\alpha$=0.10 (the pre-specified error rate)[19,23]. Closed testing methods involve forming the set of all possible intersection hypotheses among the $H_i$ treatment versus placebo comparisons (the closure of the set). Each intersection hypothesis is tested using an appropriate alpha-level test. Any hypothesis $H_i$ can be rejected with control of the FWE when: (1) the test of $H_i$ is statistically significant; and (2) the test of every intersection hypothesis that includes $H_i$ is statistically significant. The intersection hypotheses were tested using Simes' method. See figure 5-1.

**Figure 5-1. Closed Testing Method for a Trial with Four Treatment versus Placebo Comparisons (Many-To-One)**[2,32]



Closed testing procedures can be combined with adaptive methodology to control multiplicity from *a priori* ordered hypotheses.[48] The closed testing scheme for the adaptive tests is given in figure 5-2. This is simpler than figure 5-1, because at stage 1 only data for treatment 1 and 2 (creatine and minocycline) were available, and at stage 2 only data for treatment 3 and 4 (CoQ-10 and GPI-1485) were available. Thus, at the end of the study, if $H_{12\cap 34}$ is rejected (by an adaptive combination test) then we can step-down to test $H_{12}$ and $H_{34}$ and then the individual hypotheses $H_{1,}$ $H_{2,}$ $H_{3,}$ $H_{4.}$

**Figure 5-2. Closed Testing Scheme for Adaptive Method in this Example**



*Application to the FS1 and FS-TOO Data*

First, the results for each of the possible testing methods were found using the actual data. The following testing methods were performed: (1) two-sample *t*-test with the data pooled across stages; (2) linear model adjusting for a fixed cohort effect; (3) inverse $\chi^2$ (Fisher's) combination test (ICHI)[38]; or (4) Weighted-Inverse Normal combination test (INORM)[40]. Next, we bootstrapped samples (stratified by treatment group and study) to estimate the probability of rejecting the null hypothesis for each treatment comparison under the possible testing methods. Bootstrapping, or repeated sampling with replacement, is a nonparametric method of inference using the empirical distribution. By repeatedly sampling from a given dataset, one can estimate properties of an estimator[85]. Here, we used bootstrapping to estimate the probability of rejecting the null hypothesis defined by the total number of times the bootstrapped samples rejected the null hypothesis (for a given testing method) divided by the total number of samples drawn. Samples of the FS1 and FS-TOO data were drawn with replacement (i.e. bootstrapped) from each treatment group. In this case the number of bootstrapped samples drawn was set at 10,000 since this would provide adequate stability in the estimate properties and could be easily achieved with the given computing resources[85].

### 5.3 Results

The estimated sensitivity index was 1.05 with all six treatment groups combined. This value is close to 1 and is indicative of little change in the target population across studies and treatment groups. However, when examining only the placebo groups, the sensitivity index was 3, indicating a large shift in the target placebo population from FS1 to FS-TOO. Figure 5-3 shows box and whisker plots for the outcome measure (UPDRS change) by treatment group for the bootstrapped samples. The shift in the distribution of the placebo groups between studies remained evident after bootstrapping.

**Figure 5-3. Box and Whisker Plots for Total UPDRS 12 month change by Treatment Group after 10,000 Bootstrapped Samples**

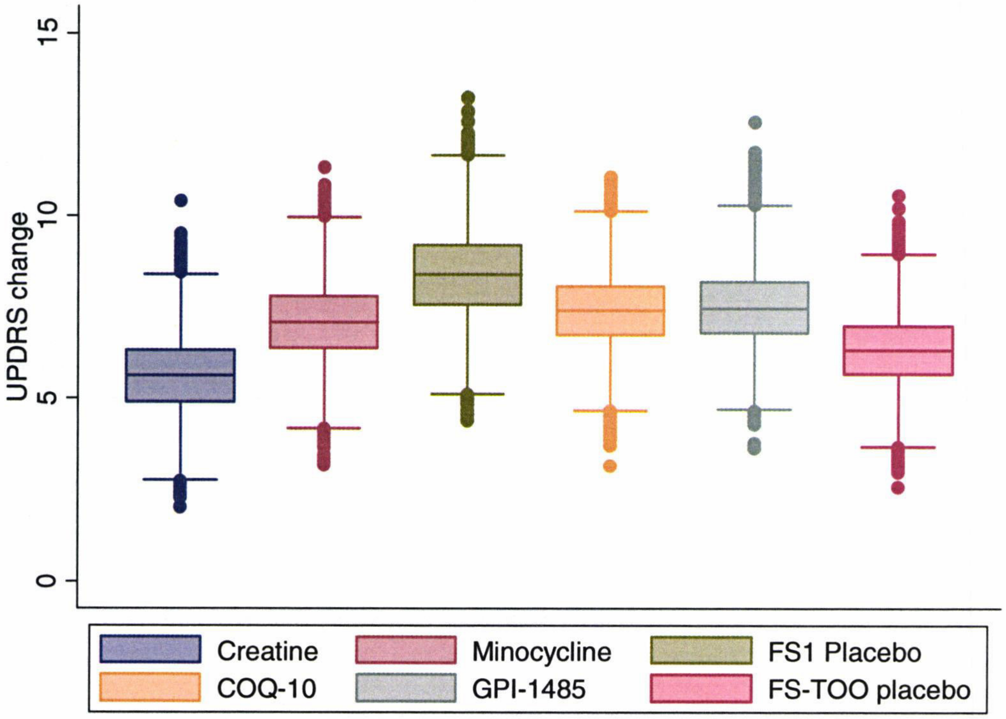


Table 5-1 shows the results for two-sample tests of each treatment versus placebo. The p-value for Hypothesis column gives the unadjusted p-values for $H_1$, $H_2$, $H_3$, and $H_4$ and the p-values (obtained via Simes' method) for the intersection hypotheses. Adjusted p-values for $H_1$, $H_2$, $H_3$, and $H_4$ were found via closed testing methods. Bolded p-values indicate the null hypothesis was rejected at p<0.10, so the treatment was considered futile.

In the first three columns, the results are shown for when the FS-TOO study arms are not added to FS1 ("Separate Studies"). The t-statistic, unadjusted p-values, and multiplicity adjusted p-values are given. When FS-1 and FS-TOO were analyzed as separate studies, we rejected the null hypotheses for GPI-1485 and COQ-10, but not for creatine and minocycline.

Next, in Table 5-1 the results are given for when the data were pooled across stages (studies). The table rows display the results for the intersection hypotheses. In this case, we rejected the null hypothesis for all treatments except for creatine. When a linear model (adjusting for stage) was performed, we rejected the null hypothesis for GPI-1485 and COQ-10, but not for creatine and minocycline.

In Table 5-2 the results are given for the adaptive combination tests. The Stage 1 and Stage 2 p-value columns give the unadjusted p-values for $H_{1,}$ $H_{2,}$ $H_{3,}$ and $H_4$ and the p-values for the intersection hypotheses (obtained via Simes' method). Unadjusted p-values are given for the combination test of $H_{12\cap 34}$ . Then adjusted p-values (via closed testing methods) for $H_{1,}$ $H_{2,}$ $H_{3,}$ and $H_4$ are given. Bolded p-values indicate the null hypothesis was rejected at p<0.10, and the treatment was considered futile. For both adaptive combination tests (ICHI and INORM), we rejected the null hypotheses for GPI-1485 and COQ-10, but not for creatine and minocycline.

**Table 5-1. Non-adaptive Test Statistics, p-value, and conclusion for Treatment 1, Treatment 2, Treatment 3, and Treatment 4 versus Placebo**

| | Separate Studies | | |
|---|---|---|---|
| | T | p-value for Hypothesis | Adjusted p-values |
| $H_1$ | 0.1308 | 0.4480 | 0.4480 |
| $H_2$ | 1.0607 | 0.1454 | 0.2907 |
| $H_{12}$ | - | 0.2907 | |
| $H_3$ | 2.6235 | 0.0048 | **0.0048** |
| $H_4$ | 2.6907 | 0.0040 | **0.0040** |
| $H_{34}$ | - | 0.0048 | |

| | Pool placebo across studies (t-test) | | | Linear Model (adjusting for stage) | | |
|---|---|---|---|---|---|---|
| | T | p-value for Hypothesis | Adjusted p-values | T | p-value for Hypothesis | Adjusted p-values |
| $H_1$ | 0.9536 | 0.1707 | 0.1707 | 0.1360 | 0.4459 | 0.4459 |
| $H_2$ | 2.0560 | 0.0205 | **0.0411** | 1.1040 | 0.1351 | 0.2702 |
| $H_3$ | 2.3399 | 0.0101 | **0.0303** | 2.7458 | 0.0032 | **0.0095** |
| $H_4$ | 2.4220 | 0.0081 | **0.0244** | 2.8213 | 0.0025 | **0.0075** |
| $H_{12}$ | - | 0.0411 | | - | 0.2702 | |
| $H_{13}$ | - | 0.0202 | | - | 0.0063 | |
| $H_{14}$ | - | 0.0163 | | - | 0.0050 | |
| $H_{23}$ | - | 0.0202 | | - | 0.0050 | |
| $H_{24}$ | - | 0.0163 | | - | 0.0063 | |
| $H_{34}$ | - | 0.0101 | | - | 0.0032 | |
| $H_{123}$ | - | 0.0303 | | - | 0.0095 | |
| $H_{124}$ | - | 0.0244 | | - | 0.0075 | |
| $H_{134}$ | - | 0.0152 | | - | 0.0047 | |
| $H_{234}$ | - | 0.0152 | | - | 0.0047 | |
| $H_{1234}$ | - | 0.0202 | | - | 0.0063 | |

Note:
$H_1$=Treatment 1 (Creatine) versus placebo
$H_2$= Treatment 2 (Minocycline) versus placebo
$H_3$= Treatment 3 (COQ-10) versus placebo
$H_4$= Treatment 4 (GPI-1485) versus placebo

- Not Applicable

**Table 5-2. Adaptive Test Statistics, p-value, and conclusion for Treatment 1, Treatment 2, Treatment 3, and Treatment 4 versus Placebo**

|  |  |  | Inverse Chi-square (ICHI) |  |  | Weighted Inverse Normal (INORM) |  |  |
|---|---|---|---|---|---|---|---|---|
|  | Stage 1 p-value | Stage 2 p-value | T | Unadjusted p-value | Adjusted p-values | Z | Unadjusted p-value | Adjusted p-values |
| $H_1$ | 0.4480 | - | - | 0.4480 | 0.4480 | - | 0.4480 | 0.4480 |
| $H_2$ | 0.1454 | - | - | 0.1454 | 0.2907 | - | 0.1454 | 0.2907 |
| $H_3$ | - | 0.0048 | - | 0.0048 | **0.0259** | - | 0.0048 | **0.0132** |
| $H_4$ | - | 0.0040 | - | 0.0040 | **0.0259** | - | 0.0040 | **0.0132** |
| $H_{12}$ | 0.2907 | - | - | 0.2907 |  | - | 0.2907 |  |
| $H_{34}$ | - | 0.0048 | - | 0.0048 |  | - | 0.0048 |  |
| $H_{12 \cap 34}$ | 0.2907 | 0.0048 | 0.0014 | 0.0259 |  | 2.2197 | 0.0132 |  |

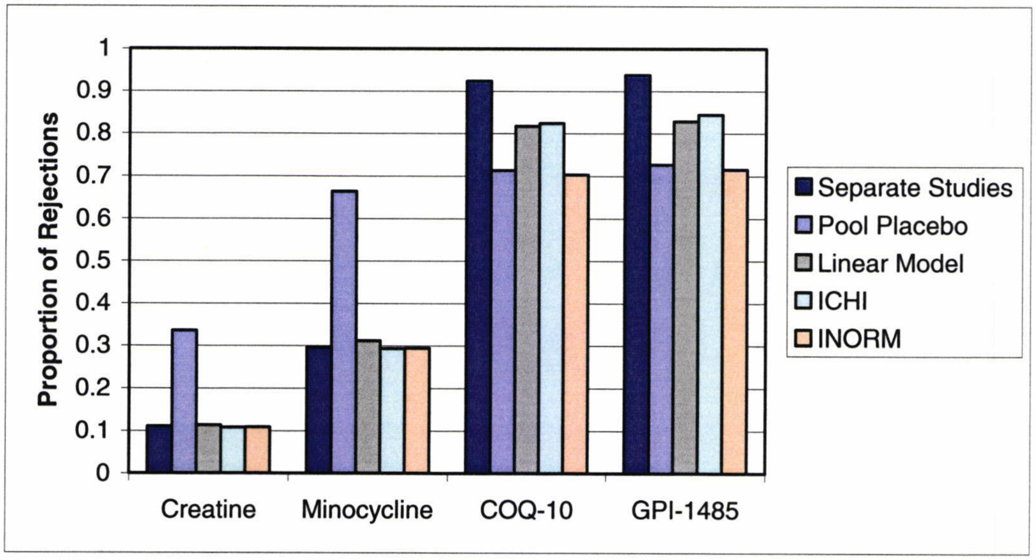
Note:
$H_1$=Treatment 1 (Creatine) versus placebo
$H_2$= Treatment 2 (Minocycline) versus placebo
$H_3$= Treatment 3 (COQ-10) versus placebo
$H_4$= Treatment 4 (GPI-1485) versus placebo

- Not Applicable

Figure 5-4 shows the proportion of rejections of the null hypothesis in 10,000 bootstrapped samples. The bootstrapped results support the findings using the actual data (in Tables 5-1 and 5-2) and are consistent with the original published study findings[7]. The results vary by treatment group. For creatine, all methods, except pooling the placebo, had a proportion of rejections around 0.10, which is what one would expect if the null hypothesis was in fact true. Of note is that for COQ-10 and GPI-1485, all methods had a proportion of rejections greater than 0.70 which would suggest that the true distribution for these treatment groups is from the alternative hypothesis (see **Methods** section). This is consistent with the original FS-TOO paper where these two treatments (COQ-10 and GPI-1485) were found futile[6]. For the minocycline group, the proportion of rejections is fairly low (30%) except when using the pooled method (65%).

Pooling the placebo groups increased the proportion of rejections for both creatine and minocycline, and likewise decreased the proportion of rejections for both COQ-10 and GPI-1485 compared with not combining studies ("Separate Studies"). This was driven by the direction of the mean placebo differences in the two studies (mean of 8.4 versus 6.3 for FS1 and FS-TOO placebos, respectively). The inconsistency of the results with the pooling method indicates (as in Chapter 4) that this method is inappropriate.

**Figure 5-4. Proportion of Rejections of Treatment vs Placebo in a two-sample test after 10,000 bootstrapped samples**



## 5.4 Discussion

For creatine and minocycline, pooling the placebos across studies is the least conservative approach in that the null hypothesis more likely to be rejected (compared with the tests for separate studies or the adaptive tests). Moreover, by pooling the placebo groups, statistical power to reject the null hypothesis is diminished for the two arms that are clearly futile (GPI-1485 and COQ-10). As shown in Chapter 4, the linear model (adjusting for stage) and the adaptive combination tests are better analysis choices when a treatment arm(s) is added to an ongoing trial.

In this example three methods of inference (linear model, ICHI, and INORM) gave similar results. This is partially due to the limitations of the data since there was no stage 2 data for FS1 available or stage 1 data for FS-TOO.

For ICHI and INORM, a 2-stage combination test was used to test the intersection hypothesis $H_{12\cap 34}$ . However, for the individual pairwise hypotheses $H_1$, $H_2$, $H_3$, and $H_4$, data were not available for both stages so the individual hypotheses were tested via a (one stage) *t*-test, rather than a 2-stage combination test. For the linear model, the stage effect was actually a study effect, since all subjects in stage 1 were in FS1 and all subjects in stage 2 were in FS-TOO. Had the studies actually had overlapping enrollment periods, where the FS-TOO arms were introduced after, say, 50% of the FS1 subjects were enrolled, we may have had somewhat different results. Nevertheless, this paper shows how these methods would be applied using real clinical data and the conclusions are consistent with those reported in the primary study papers[6,7].

Any modification to the protocol (even a minor one) has the potential to introduce bias in the study conclusions[84, p.38]. A major modification could have a substantial impact on statistical inferences by shifting the mean response or inflating the variability, even when the type I error rate is controlled at a pre-specified level[84, p.45]. In this example, the sensitivity index indicates a significant shift in the target placebo population from FS1 to FS-TOO. However, all methods used when combining the studies produce conclusions similar to those when the studies are analyzed separately.

The decision to modify a protocol to add a new treatment arm to the study is likely to be based solely on external information. Internal interim effect size is unlikely to play a role in this decision, unless the treatment added was a different dose of the same drug. When only external information is used, then it would be acceptable to apply a linear model approach with an appropriate multiple comparison procedure such as shown here. The adaptive combination tests are also a suitable choice, and are the best choice if internal information contributed at all to the decision to add the arm(s) to the ongoing study[41].

In most Parkinson's disease clinical trials, the primary outcome measure is the Unified Parkinson's Disease Rating Scale (UPDRS). The total score is the sum of 30 ordinal questions on mental, ADL, and motor dysfunction. The questions are completed by a neurologist upon examining the patient and are, by nature, somewhat subjective. The primary outcome for these two studies was defined as the change from baseline to 12 months OR the visit at which the need for symptomatic therapy was determined. Thus, changes in practice of the introduction of symptomatic therapy could also influence the primary outcome measure. A significant shift in the timing of the introduction of symptomatic therapy was observed in FS1 and FS-TOO versus the historical DATATOP dataset[7].

Unpublished data indicate there was increased enthusiasm for the $CoQ_{10}$ drug compared to the GPI-1485 drug among the NET-PD investigators at the outset of FS-TOO, and this is likely to have an impact on subjective outcome measures. Given the subjective nature of the primary outcome and the observed changes in clinical practice, a cohort effect may be a legitimate concern for Parkinson's disease clinical trials.

The NET-PD group is enrolling PD patients into the LS-1 study of creatine versus placebo with 860 subjects/group. The economies of scale by adding another treatment arm and thereby saving on the required placebo subjects are clear.  However, this may not be possible depending on the timing of the availability of future compounds ready to be tested in a Phase III study. If a treatment arm is to be added into an ongoing study, then this example illustrates that either a linear model approach or an adaptive combination test is well suited for such a situation.

# CHAPTER 6. SUMMARY AND CONCLUSIONS

In principle, clinical trials should be conducted according to plan and as specified in the protocol. Any modification made mid-course to the conduct of a trial should be approached with caution, as it may introduce bias into study. That being said, there is often a real need for clinical trial designs that can accommodate mid-course design changes. The motivation for the work here is the need within the NET-PD studies for group sequential multi-armed trials that can handle the introduction of new treatment arms at different times.

As the methodology advances, mid-course design changes (re-designing a trial) are possible. The penalty for performing mid-course design changes (based on internal or external data) are the use of non-standard test statistics which are not sufficient statistics (combination tests of stage-wise p-values)[36]. In general, standard guidance documents and education are needed to fully support the conduct of adaptive clinical trials (particularly phase III confirmatory trials)[36]. With mid-course design changes, there is a need for transparency of the decision process combined with the continued need for blinding of investigators until the study is complete[36]. Documentation of the decision process is essential[11]. Although regulatory agencies have announced their support of such flexible designs, they are in general more conservative in regards to their use[11,14]. As with any new methodological development, it takes time to transfer from development to application without concerns about credibility. There is no doubt that the flexibility of mid-course decision making makes sense ethically and economically, provided the inference will not be affected.

The ethical considerations of conducting group sequential multi-armed clinical trials are significant. By specifying a group sequential design, the investigators have decided *a priori* to stop the trial once efficacy has been determined. Yet as we showed in Chapter 3, when two treatment arms are equally efficacious, it is likely that one, but not the other, will be found efficacious at an interim analysis. There may be ethical concerns about continuing to enroll or follow subjects on placebo to study the second treatment, once efficacy has been established for one treatment. This is the rationale for transitioning into a non-inferiority trial with the first treatment as the standard. Chapter 3 suggests analytical methods for transitioning into a non-inferiority trial including methods to constrain the type I error rate strongly. Approaches to increase the sample size are suggested since sample size necessary for a non-inferiority test is greater than for superiority (given the same power is desired). The results of the Monte Carlo simulation study suggest all methods performed similarly, but empirical power was slightly higher when using an inverse chi-square adaptive testing procedure. When sample size was allowed to be re-estimated based on the observed effect size, then the adaptive procedure required a sample size approximately 50-80% of that necessary for a non-inferiority hypothesis based on the original sample estimates. Thus, the adaptive methods allow for a reduction in total sample size with increased power for testing non-inferiority (compared to a non-adaptive approach). Further research is needed to correct for the biased point estimate and its affect on the inference as noted in Chapter 3. Further research would also be to consider under simulation a linear contrast of the three groups to test for non-inferiority as a percent retention of the (observed) active control effect in the context of sequential testing trial[61,68] and to consider the impact of a cohort effect.

In Chapter 4, we see that both adaptive and non-adaptive analytical methods are possible when a new treatment arm is added mid-study. Methods for controlling the type I error rate strongly in the presence of various sources of multiplicity (multiple treatment arms, interim analyses, and design changes) are reviewed. The possibility of a cohort effect is addressed, and operating characteristics are compared under various levels of a cohort effect. While in most cases, the decision to add a new treatment arm is likely based on external information (such as new results from a pilot study), the possibility of introducing bias still exists (such as due to increased enthusiasm) and the analysis approach adopted should account for that[15]. The results show that simply pooling data, from before and after the design change, would not be prudent. When these methods are applied to real Parkinson's disease trial data in Chapter 5, the conclusions support the findings from Chapter 4 as well as the findings from the primary trial findings[6,6,7,7].

One limitation of this research is the focus on continuous outcomes (the normal case). In Parkinson's disease clinical trials, the standard outcome measure used is the UPDRS, which is continuous. However, the Phase III NET-PD study of creatine is using a global outcome measure. Further research is needed to extend this work for multivariate outcomes. Another limitation is the use of only three-arms and only one interim analysis. The impact of more than three treatment arms and/or more interim analyses was not considered here.

A limitation of the adaptive testing procedure is the restriction to one-sided testing. A publication written in German gives a method to perform doubly one-sided tests, but further research is needed to assess the operating characteristics for this testing approach[74]. Another limitation of adaptive combination tests (either pre-specified or otherwise) is the loss of unbiased point and confidence interval estimation[13]. There have been some developments in this area, but further research is needed[72].

There is limited literature on group sequential multi-armed clinical trials or flexible designs for multi-armed clinical trials. It is widely recommended to perform simulations at the planning stage in order to fully understand the implications of mid-course design changes since inference becomes more complex[12,65]. This work helps to fill this gap and provide some guidance to clinical trialists in the Parkinson's disease area.

## List of References

1. Follmann DA, Proschan MA, Geller NL. Monitoring Pairwise Comparisons in Multi-Armed Clinical Trials. Biometrics 1994; 50(2):325-336.

2. Hellmich M. Monitoring clinical trials with multiple arms. Biometrics 2001; 57(3):892-898.

3. de Rijk MC, Launer LJ, Berger K, Breteler MM, Dartigues JF, Baldereschi M et al. Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. Neurology 2000; 54(11 Suppl 5):S21-3.

4. Fahn S. Description of Parkinson's Disease as a clinical syndrome. Ann N Y Acad Sci 2003; 991:1-14.

5. Ravina BM, Fagan SC, Hart RG, Hovinga CA, Murphy DD, Dawson TM et al. Neuroprotective agents for clinical trials in Parkinson's disease: A systematic assessment. Neurology 2003; 60(8):1234-1240.

6. NINDS NET-PD Investigators. A randomized, double-blind, futility clinical trial of creatine and minocycline in early Parkinson disease. Neurology 2006; 66(5):664-671.

7. NINDS NET-PD Investigators. A Randomized Clinical Trial of Coenzyme Q10 and GPI-1485 in Early Parkinson's Disease. Neurology 2007; 68(1):20-28.

8. Elm JJ, Goetz CG, Ravina B, Shannon K, Wooten GF, Tanner CM et al. A responsive outcome for Parkinson's disease neuroprotection futility studies. Ann Neurol 2005; 57(2):197-203.

9. Gallo P, Maurer W. Challenges in implementing adaptive designs: comments on the viewpoints expressed by regulatory statisticians. Biom J 2006; 48(4):591-597.

10. Wang SJ. Regulatory experience of adaptive designs in well-controlled clinical trials. *Presented at "Adaptive Designs: Opportunities, Challenges and Scope in Drug Development", Washington, DC.* 2006.

11. Hung HM, O'Neill RT, Wang SJ, Lawrence J. A regulatory view on adaptive/flexible clinical trial design. Biom J 2006; 48(4):565-573.

12. Chow SC, Chang M. Adaptive design methods in clinical trials – a review. Orphanet Journal of Rare Diseases 2008; 3(11).

13. Bauer P, Einfalt J. Application of adaptive designs--a review. Biometrical Journal 2006; 48(4):493-506.

14. Koch A. Confirmatory clinical trials with an adaptive design. Biom J 2006; 48(4):574-585.

15. Feng H, Shao J, Chow S. Group sequential test for clinical trials with moving patient population. J Biopharmaceutical Statistics 2007; 17:1227-1238.

16. O'Brien P. The appropriateness of analysis of variance and multiple comparison procedures. Biometrics 1983; 39:787-794.

17. Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 1967; 62:626-633.

18. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 1976; 6:65-70.

19. Simes R. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986; 73(751):754.

20. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988; 75(800):802.

21. Hommel G. A comparison of two modified Bonferroni procedures. Biometrika 1988; 75(383):386.

22. Rom D. A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika 1990; 77(663):665.

23. Marcus R, Peritz E, Gabriel KR. On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. Biometrika 1976; 63(3):655-660.

24. Westfall P, Tobias R, Rom D, Wolfinger R, Hochberg Y. Multiple Comparisons and Multiple Tests Using the SAS system. Cary, NC: SAS Institute Inc, 1999.

25. Westfall PH, Tobias RD, Rom D, Wolfinger RD, Hochberg Y. Multiple Comparisions and Multiple Tests: Using the SAS System. Cary, NC: SAS Institute Inc., 1999.

26. Armitage P, McPherson CK, Rowe BC. Repeated Significance Tests on Accumulating Data. Royal Statistical Society Journal 1969; A(132):235.

27. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35(3):549-556.

28. Pocock S. Interim Analyses for Randomized Clinical Trials: The Group Sequential Approach. Biometrics 1982; 38:153-162.

29. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983; 45:1017-1020.

30. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35(3):549-556.

31. Jennison C, Turnbull B. Group sequential methods with applications to clinical trials. Boca Raton: Chapmann & Hall, CRC, 2000.

32. Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics 1991; 47(2):511-521.

33. Tang D, Geller N. Closed Testing Procedures for Group Sequential Clinical Trials with Multiple Endpoints. Biometrics 1999; 55(4):1188-1192.

34. Gould A, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Commun Stat-Theory Methods (B) 1992; 21:2833-2853.

35. Gould A. Planning and revising the sample size for a trial. Stat Med 1995; 14(9-10):1039-51.

36. Bauer P, Einfalt J. Application of Adaptive Designs – a Review. Biom J 2006; 48(4):493-506.

37. Special Issue: Adaptive Designs in Clinical Trials. Biom.J. 48[4], 487-737. 2006.

38. Fisher R. Statistical Methods for Research Workers. London: Oliver and Boyd, 1932.

39. Bauer P, Kohne K. Evaluation of Experiments with Adaptive Interim Analyses. Biometrics 1994; 50(4):1029-1041.

40. Mosteller F, Bush R. Selected quantitative techniques. In: Lindzey G, editor. *Handbook of Social Psychology*. Cambridge, Massachusetts: Addison-Wesley, 1954: 289-334.

41. Cui L, Hung HMJ, Wang SJ. Modification of Sample Size in Group Sequential Clinical Trials. Biometrics 1999; 55(3):853-857.

42. Lehmacher W, Wassmer G. Adaptive Sample Size Calculations in Group Sequential Trials. Biometrics 1999; 55(4):1286-1290.

43. Proschan MA, Hunsberger SA. Designed Extension of Studies Based on Conditional Power. Biometrics 1995; 51(4):1315-1324.

44. Posch M, Bauer P. Adaptive two stage designs and the conditional error function. Biometrical Journal 1999; 41(6):689-696.

45. Wassmer G, Eisebitt R, Coburger S. Flexible interim analyses in clinical trials using multistage adaptive test designs. Drug Information Journal 2001; 35:1131-1146.

46. Denne JS. Sample size recalculation using conditional power. Stat Med 2001; 20(17-18):2645-2660.

47. Shen Y, Fisher L. Statistical Inference for Self-Designing Clinical Trials with a One-Sided Hypothesis. Biometrics 1999; 55(1):190-197.

48. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. Biom J 1999; 41(3):261-277.

49. Hommel G. Adaptive Modifications of Hypotheses After an Interim Analysis. Biometrical Journal 2001; 43(5):581-589.

50. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. Stat Med 1999; 18(14):1833-1848.

51. Lan KKG, DeMets DL. Changing Frequency of Interim Analysis in Sequential Monitoring. Biometrics 1989; 45:1017-1020.

52. Bauer P, Rohmel J. An adaptive method for establishing a dose-response relationship. Stat Med 1995; 14:1595-1607.

53. Bauer M, Bauer P, Budde M. A simulation program for adaptive two stage designs. Computational Statistics & Data Analysis 1998; 26(3):351-371.

54. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. Stat Med 2003; 22(6):953-969.

55. Hommel G, Lindig V, Faldum A. Two-Stage Adaptive Designs with Correlated Test Statistics. J Biopharm Stat 2005; 15:613-623.

56. Brannath W, Posch M, Bauer P. Recursive combination tests. Journal of the American Statistical Association 2002; 97(457):236-239.

57. Chang M. Adaptive Design Theory and Implementation Using SAS and R. Boca Raton, FL: Chapman & Hall/CRC, 2008.

58. Muller HH, Schafer H. A general statistical principle for changing a design any time during the course of a trial. Stat Med 2004; 23(16):2497-2508.

59. Pullman D, Wang X. Adaptive designs, informed consent, and the ethics of research. Control Clin Trials 2001; 22(3):203-210.

60. Koyama T, Sampson AR, Gleser LJ. A framework for two-stage adaptive procedures to simultaneously test non-inferiority and superiority. Stat Med 2005; 24(16):2439-2456.

61. Kropf S, Hommel G, Schmidt U. Multiple Comparisons of Treatments with Stable Multivariate Tests in a Two-Stage Adaptive Design, Including a Test for Non-Inferiority. Biometrical Journal 2000; 42(8):951-965.

62. Hochberg Y, Tamhane A. Multiple Comparison Procedures. New York: John Wiley & Sons, 1987.

63. Westfall P, Tobias R, Rom D, Wolfinger R, Hochberg Y. Multiple Comparisons and Multiple Tests Using the SAS System. Cary, NC: The SAS Institute, Inc., 1999.

64. Denne JS, Koch GG. Monitoring a clinical trial with multiple hypotheses concerning the treatment effect on a single primary endpoint. Stat Med 2001; 20(19):2801-2812.

65. Hung HMJ, O'Neill RT, Wang SJ, Lawrence J. A Regulatory View on Adaptive/Flexible Clinical Trial Design. Biometrical Journal 2006; 48(4):565-573.

66. Denne JS, Koch GG. Monitoring a clinical trial with multiple hypotheses concerning the treatment effect on a single primary endpoint. Stat Med 2001; 20(19):2801-2812.

67. Wang SJ, Hung HM, Tsong Y, Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. Stat Med 2001; 20(13):1903-1912.

68. Pigeot I, Schafer J, Rohmel J, Hauschke D. Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. Stat Med 2003; 22(6):883-899.

69. Dunnett C, Gent M. An Alternative to the Use of Two-Sided Tests in Clinical Trials. Stat Med 1996; 15:1729-1738.

70. Chang M. Adaptive Design Theory and Implementation Using SAS and R. Boca Raton, FL: Chapman & Hall/CRC, 2008.

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

72. Coburger S, Wassmer G. Conditional Point Estimation in Adaptive Group Sequential Test Designs. Biometrical Journal 2001; 43(7):821-833.

73. Armitage P, McPherson CK, Rowe BC. Repeated Significance Tests on Accumulating Data. Royal Statistical Society Journal 1969; A(132):235.

74. Wassmer G. *Statistische Testverfahren für Gruppensequentielle und Adaptive Pläne in Klinischen Studien.* 2nd ed. ed. Köln: Verlag Alexander Mönch, 2001.

75. Geller N, Pocock SJ. Interim Analyses in Randomized Clinical Trials: Ramifications and Guidelines for Practitioners. Biometrics 1987; 43:213-223.

76. FDA. Code of Federal Regulations Title 21-Food and Drugs, Subchapter D-Drugs for Human Use Part 312 Investigational New Drug Application. 2005.

77. Thall PF, Simon R, Ellenberg S. Two-Stage Selection and Testing Designs for Comparative Clinical Trials. Biometrika 1988; 75(2):303-310.

78. Thall P, Simon R, Ellenberg S. A Two-Stage Design for Choosing among Several Experimental Treatments and a Control in Clinical Trials. Biometrics 1989; 45(2):537-547.

79. Hughes MD. Stopping guidelines for clinical trials with multiple treatments. Stat Med 1993; 12:901-905.

80. Vincent E, Todd S, Whitehead J. A sequential procedure for comparing two experimental treatments with a control. J Biopharm Stat 2002; 12(2):249-265.

81. Stallard N, Todd S. Sequential designs for Phase III clinical trials incorporating treatment selection. Stat Med 2003; 22:689-703.

82. Polansky A. Selecting the best treatment in designed experiments. Stat Med 2003; 22:3461-3471.

83. Herson J, Carter S. Calibrated phase II clinical trials in oncology. 1986;(441):447.

84. Chow SC, Chang M. Adaptive Design Methods in Clinical Trials. Boca Raton, FL: Chapman & Hall/CRC, 2007.

85. Efron B, Tibshirani R. An Introduction to the Bootstrap: Monographs on Statistics and Applied Probability. New York: Chapman & Hall/CRC Press, 1993.