2020

# Sample size recalculation in three-arm non-inferiority trials

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

# SAMPLE SIZE RECALCULATION IN THREE-ARM NON-INFERIORITY TRIALS

by

**LANYU LEI**

B.S., University of Science and Technology of China, 2006
M.S., University of Illinois at Urbana-Champaign, 2010

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

Approved by

First Reader _____

Joseph M. Massaro, Ph.D.
Professor of Biostatistics

Second Reader _____

Ralph B. D'Agostino, Ph.D.
Professor of Biostatistics

Third Reader _____

Michael P. LaValley, Ph.D.
Professor of Biostatistics

## DEDICATION

I would like to dedicate this work to the ladies of my life. First, I dedicate this to my mother. I could not have completed this without her love and support. Second, I dedicate this to my wonderful daughters, Irene and Elena. I hope they understand one day why Mommy spend so much time working and studying. Whenever I felt frustrated, their smiles are the best medicine to cure my problems and brighten my day.

## ACKNOWLEDGEMENTS

I cannot adequately express my appreciation to my academic and thesis advisor, Prof. Joseph Massaro. He has been a teacher and mentor to me for many years, and more than that, he is a role model to me. I admire his ability to communicate the big picture of any complicated designs to the audience with various background with ease. I can't say enough thanks to him for his flexibility, guidance and great patience with me throughout this long PhD journey.

I also wish to thank the other members of my committee, Dr. Ralph D'Agostino, Dr. Michael Lavalley, Dr. Zhigen Zhao and Dr. Howard Cabral, for their willingness to serve in those roles. Their helpful guidance and probing questions have improved the quality of the dissertation.

**SAMPLE SIZE RECALCULATION IN**

**THREE-ARM NON-INFERIORITY TRIALS**

**LANYU LEI**

Boston University Graduate School of Arts and Sciences, 2020

Major Professor: Joseph M. Massaro, Professor of Biostatistics

ABSTRACT

The three-arm non-inferiority trials include an experimental treatment arm, an active comparator arm and a placebo arm. Such a design allows evaluating the assay sensitivity by testing the superiority of active comparator over placebo, and is a preferred choice when the constancy of the treatment effect is questionable under the current medical setting, or when there is no consensus on the magnitude of a clinically relevant treatment effect. This dissertation, investigating sample size recalculation in active- and placebo-controlled trials at the interim, is composed of three chapters. The first chapter summarizes for the three-arm non-inferiority design, including hypotheses formulations, testing procedures and a few important features. We also compare statistical power between the two forms of the non-inferiority test: fixed margin test and effect preservation test. In the second chapter, the commonly used group sequential designs and sample size recalculation methods are reviewed, and two methods (Method I and Method II) are applied to the three-arm non-inferiority trials with normally distributed endpoint. Method I which was proposed in previous publications, does not require unblinding the experimental treatment effect. However, it has a limitation that the sample sizes of the three arms

must meet specific ratios in order to control Type I error. This thesis extends this method to a broader range of sample size allocations. Method II is based on the concept of promising zone of the conditional power. It is a modified group sequential design and recalculate the sample size only when the interim result is promising. The overall type I error control, power and efficiency of the two methods (I vs. II) are compared under various scenarios through simulations. It is found that the method I provide a substantial power gain when the initial active comparator effect is over-estimated at the design stage. However, it cannot handle the uncertainties in the experimental treatment effect. In contrast, the Method II controls type I error at all investigated sample size allocations. It provides moderate power gain if the preserved effect is over-estimated. Once the interim result is promising the recalculated sample size can increase the final rejection probability considerably. When the experimental treatment effect is under-estimated at the design stage, the average sample size can reduce dramatically compared to fixed sample design due to the high probability of rejection at the interim stage. The third chapter investigates how the two methods perform for the binary endpoint. The formula of conditional power for binary outcome is derived. The statistical properties including the variance estimation, the type I error control and actual power are investigated through simulations.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER I: THREE-ARM NON-INFERIORITY TRIALS

## 1.1 Introduction and Background

In randomized clinical trials (RCT), there are two major approaches to demonstrate the efficacy of a novel medical therapy: to prove the superiority of the experimental treatment over placebo, or to prove the superiority or non-inferiority of the experimental treatment to an active comparator. Placebo-controlled RCTs are often used for indications when no known effective and safe therapy is available. When there is proven efficacious therapies on the market, placebo-controlled RCTs are often not ethically justifiable. In some cases, if patients will not be harmed by deferral of therapy, placebo-controlled therapy may also be permitted. There are a few limitations on the RCT with placebo as the single control. Although one can demonstrate that the treatment difference between the experimental treatment and placebo is significantly different from zero, it is sometimes not clear whether the observed magnitude of the difference is clinically significant. For example, in many preventive vaccine studies, the primary endpoint is the immunogenicity (the vaccine's ability to stimulate an immune response) which is often characterized by the concentration of a certain antibody in the blood sample. Biologically the antigen in the vaccine product would induce antibody production whereas the placebo rarely has immune response. More often than not the concentration needed to provide effective protection is not clear. In such case showing the superiority to placebo is not sufficient evidence to prove the efficacy. It is often required by the regulatory agency that the primary evidence being the non-inferiority to an active comparator

or a pre-specified positive threshold. In contrast if the superiority is not shown, it is important to find out whether the non-significance is truly due to no efficacy or possibly be due to low sensitivity of the assay used to measure the efficacy. Such limitations can be circumvented by having an active comparator in the same trial. To support the marketing approval of a new therapy for an indication with existing therapy available, the strongest evidence is showing superiority of experimental treatment in efficacy compared to the active comparator. However, in many therapeutic areas the proven therapies are highly effective leading to great challenge for the new therapies to show superiority. In such case, if there is sufficient evidence showing that the new therapy is not materially worse than a standard therapy (i.e., the active comparator) and benefits patients in other aspects, such as better safety profile or lower cost, it is still worth considering having such novel therapies on market. Trials designed to show that a new treatment is 'not unacceptably worse' than the current standard therapy is called non-inferiority trials. Since the introduction of non-inferiority trials in the mid-1990s there has been controversy about the validity and interpretation on such trials, as its design is complicated and is founded on assumptions that are difficult to verify[1,2,3]. A major concern on the conventional non-inferiority trials is how the assay sensitivity and the constancy assumption[2] of the active comparator can be evaluated. As a result, the three-arm non-inferiority design, including an experimental treatment, an active comparator (reference treatment) and a placebo, was brought up. The design has been referred as "gold standard design"[4] and is recommended by regulatory agency

in the United States[5] as well as in Europe [6]. The three-arm design is preferred for the medical indications where the constancy assumption is questionable, or the assay sensitivity is critical, or when there is no consensus on the magnitude of a clinically relevant treatment effect[7]. Such situations include but not limited to:

(1) There is doubts in the efficacy of active comparator due to relatively large fluctuation in the efficacy estimates observed in different trials, or because the active comparator is established long ago so that the historical finding may not be valid or accurate in the present medical setting.

(2) The active comparator effect is "weak" and thus it might be difficult to justify a negligible loss of efficacy in the present medical setting.

(3) Constancy assumption is not valid due to varying response to both active comparator and to placebo. For instance, depression studies usually have such issues.

(4) In a placebo-controlled clinical trial, incorporating an active comparator group may be needed when there is difficulty defining what constitutes a clinically relevant difference in efficacy by looking at the difference between experimental treatment and placebo alone.

(5) Superiority to placebo might be less meaningful if the standard treatment significantly outperforms the experimental treatment.

### 1.2 Hypothesis Formulation and Testing Procedure

Let $\mu_T$, $\mu_A$ and $\mu_P$ represent the population means of the treatment effect under the experimental treatment, active comparator and placebo arms, respectively. Without loss of generosity, it is assumed that larger value means better treatment result. The following four hypothesis tests could form a closed testing procedure:

(1) $H_{0,TP}^{(s)}$: $\mu_T \leq \mu_P$ versus $H_{1,TP}^{(s)}$: $\mu_T > \mu_P$

(2.1) $H_{0,TA}^{(n)}$: $\mu_T \leq \mu_A - \Delta$ versus $H_{1,TA}^{(n)}$: $\mu_T > \mu_A - \Delta$

where $\Delta$ is a pre-specified fixed value called margin. When $\Delta$ is replaced with a proportion of mean difference between active comparator and placebo, $\lambda(\mu_A-\mu_P)$, the form of hypothesis (2.1) became

(2.2) $H_{0,TA}^{(n)}: \frac{\mu_T-\mu_P}{\mu_A-\mu_P} \leq 1-\lambda$ versus $H_{1,TA}^{(n)}: \frac{\mu_T-\mu_P}{\mu_A-\mu_P} > 1-\lambda$

This form is often referred to as effect preservation test or fractional test and will be discussed in chapter 1.3.

(3) $H_{0,AP}^{(s)}$: $\mu_A \leq \mu_P$ versus $H_{1,AP}^{(s)}$: $\mu_A > \mu_P$

If (2.1) or (2.2) is rejected, i.e., the non-inferiority is demonstrated, one may be interested in testing if the treatment test is superior to the active comparator.

(4) $H_{0,TA}^{(s)}$: $\mu_T \leq \mu_A$ versus $H_{1,TA}^{(s)}$: $\mu_T > \mu_A$

There has been controversy amongst authors regarding which hypotheses ought to be tested and in what hierarchy[4,7–10]. In general, $H_{0,TA}^{(n)}$ is the primary test to establish non-inferiority. When the non-inferiority test $H_{0,TA}^{(n)}$ is formulated as a fixed margin

test, there is consensus that the rejection of the $H_{0,TP}^{(s)}$: $\mu_T \leq \mu_P$ is considered the utmost important step since superiority of test of treatment over placebo is a pre-requisite for all the rest hypotheses tests. Whether $H_{0,AP}^{(s)}$ needs be rejected is the most controversial. Koch and Rohmel[4,8] suggested that the mandatory requirement of assay sensitivity is ill founded, because if active comparator fails to show superiority to placebo and in the meanwhile experimental treatment is superior to placebo, it should be considered an additional strength rather than a reason to doubt the experimental treatment. Therefore they proposed that in most cases the non-inferiority trial can be termed successful as soon as it can be demonstrated that the experimental treatment is superior to placebo and the experimental treatment is non-inferior to the reference. Hauschke and Pigeot[7] argued that under a few medical settings when the active comparator is weak or represents a traditional standard with doubts in efficacy, the efficacy of the experimental treatment over placebo can be claimed if $H_{0,TP}^{(s)}$ and $H_{0,TA}^{(n)}$ are rejected. But in other scenarios when the active comparator is a well-established treatment, the assay sensitivity should be a mandatory condition to demonstrate the validity of the study.

The controversy on the testing procedures had been focused on the hypotheses with fixed margin non-inferiority test though. When the non-inferiority test is formulated as effect preservation test, the rejection of $H_{0,AP}^{(s)}$ is a prerequisite to ensure that the denominator of $\frac{\mu_T-\mu_P}{\mu_A-\mu_P}$ is not zero. In such case, if both $H_{0,AP}^{(s)}$ and $H_{0,TA}^{(n)}$ are rejected, we could then derive that $\mu_T - \mu_P > (1-\lambda)(\mu_A - \mu_P) > 0$, i.e., it is not necessary to

test the $H_{0,TP}^{(s)}$ in this sense. It is also shown in the appendix I that when $\lambda$ is no greater than 0.5, the power is almost always dominated by the non-inferiority test, meaning there is little chance that when the non-inferiority is demonstrated the hypothesis test of superiority of experimental treatment over placebo fails.

In this thesis, when the null hypothesis for the non-inferiority test is $\frac{\mu_T-\mu_P}{\mu_A-\mu_P} \leq 1-\lambda$, the testing procedure adopted will be a two-step sequential test: $H_{0,AP}^{(s)}$ followed by $H_{0,TA}^{(n)}$.

### 1.3 Type I Error Control

With more than one tests, we are confronted with the multiplicity issues. To control the familywise error rate (FWER) the testing procedure must be carefully planned, and adjustment should be made if necessary. As discussed in 1.2, due to a natural hierarchy between the null hypotheses to be tested, control of type I error in three-arm non-inferiority studies is often achieved by following a fixed sequence procedure[11] and by considering the logical interrelations between the null hypotheses under consideration.

The fixed sequence procedure provides a straightforward but powerful approach for controlling overall type I error when there are K (K≥2) hypotheses. It assumes that the null hypotheses $H_{01}$, …, $H_{0K}$ have a priori order so that lower index corresponds to higher importance. The procedure tests hypotheses in the order of low index to high index, each at level α, and stops at the first non-rejected

hypothesis. Let I be the set of indices of all true null hypotheses { $H_{0,i}$ }, and k* be the lowest index of true null hypotheses. Given that the testing sequence is fixed, the $H_{0,k^*}$ has to be rejected before rejecting any true null hypothesis, therefore Prob(At least one $H_{0,i}$ is rejected) ≤ Prob($H_{0,k^*}$ is rejected)=α.

### 1.4 Power Function and Sample Size Determination

The power function and sample size determination depend on the distribution of outcome, hypothesis formulation and testing procedure. For continuous outcomes, we assume the observations under experimental treatment, active comparator and placebo are mutually independent and normally distributed with mean ($\mu_T$ , $\mu_A$, $\mu_P$) respectively and with common variance $\sigma^2$. The fixed sequence test procedure is used to control the family-wise type I error. If the non-inferiority test takes the form of fixed margin test, the first step tests the superiority of experimental treatment over placebo; If the non-inferiority test takes the form of effect preservation test, the first step tests the assay sensitivity, i.e., the superiority of active comparator over placebo.

#### *1.4.1. Fixed Margin Test*

The fixed margin approach is largely used in the conventional two-arm non-inferiority trials. The formulation of key hypotheses in three-arm trial as discussed above include the superiority test of experimental treatment over placebo and the non-inferiority test:

(1) $H_{0,TP}^{(s)}$: $\mu_T \leq \mu_P$ versus $H_{1,TP}^{(s)}$: $\mu_T > \mu_P$

(2) $H_{0,TA}^{(n)}$: $\mu_T \leq \mu_A - \Delta$ versus $H_{1,TA}^{(n)}$: $\mu_T > \mu_A - \Delta$

Assume $n_T$ , $n_A$ and $n_P$ subjects are the sample size for the experimental treatment, active comparator and placebo arms respectively, i.e., the total sample size N= $n_T + n_A + n_P$; The allocation of the three treatments follow the relationship of $\frac{n_A}{n_T} = c_A, \frac{n_P}{n_T} = c_P$; Let $\bar{X}_T$ , $\bar{X}_A$ and $\bar{X}_P$ be the sample means. The test statistics can be expressed as follows:

$$T_{TP}^{(s)} = \frac{\bar{X}_T - \bar{X}_P}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_P}}} \qquad (1.4.1)$$

$$T_{TA}^{(n)} = \frac{\bar{X}_T - \bar{X}_A + \Delta}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_A}}} \qquad (1.4.2)$$

When $\sigma$ is known, or $\sigma$ is unknown but sample size is large enough, the overall power for the testing procedure is calculated as

$$1 - \beta = P(\left\{T_{TP}^{(s)} > Z_{1-\alpha}\right\} \cap \left\{T_{TA}^{(n)} > Z_{1-\alpha}\right\} | \delta_{TP}, \delta_{TC}) \qquad (1.4.3)$$

Per Bonferroni inequality, the lower bound of (1.4.3) can be found as

$$\begin{aligned} 1 - \beta &= P\left(\left\{T_{TP}^{(s)} > Z_{1-\alpha}\right\} \cap \left\{T_{TA}^{(n)} > Z_{1-\alpha}\right\} | \delta_{TP}, \delta_{TA}\right) \\ &= 1 - P\left(\left\{T_{TP}^{(s)} \leq Z_{1-\alpha}\right\} \cup \left\{T_{TA}^{(n)} \leq Z_{1-\alpha}\right\} | \delta_{TP}, \delta_{TA}\right) \\ &\geq 1 - P\left(T_{TP}^{(s)} \leq Z_{1-\alpha} | \delta_{TP}\right) - P\left(\left\{T_{TA}^{(n)} \leq Z_{1-\alpha}\right\} | \delta_{TA}\right) \\ &= 1 - \beta_{TP} - \beta_{TA} \end{aligned} \qquad (1.4.4)$$

$\beta_{TP}\ and\ \beta_{TA}$ are respectively the type II error for the two hypotheses, $Z_{1-\alpha}$ is the $1-\alpha$ percentile of normal distribution and is the critical value for each test. For example, if the power of first and second step are both powered at 80%, the lower bound for overall power is 60%. It is nice to have a boundary, however, as shown in the example, it is too conservative to be used in real clinical trials.

By examining the vector **T**= $(T_{TP}^{(s)}, T_{TA}^{(n)})'$, we find that both statistics are linear combination of the means of three groups. The group means all follow normal distributions, therefore any linear combination of $T_{TP}^{(s)}$ and $T_{TA}^{(n)}$ are normally distributed. That is to say, the vector **T**= $(T_{TP}^{(s)}, T_{TA}^{(n)})'$ follows bivariate normal distribution. Hence, it is possible to calculate the overall power directly which will lead to a more efficient sample size design. Specifically, the expectation of the vector **T** is

$$\mathrm{E}\left(\left(T_{TP}^{(s)}, T_{TA}^{(n)}\right)'\right) = \left(\frac{\mu_T-\mu_P}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{C_P}}}, \frac{\mu_T-\mu_A+\Delta}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{C_A}}}\right)', \qquad (1.4.5)$$

The covariance between the two statistics is

$$\mathrm{Cov}\left(\left(T_{TP}^{(s)}, T_{TA}^{(n)}\right)'\right)$$

$$= \sigma^{-2}\sqrt{\left(\frac{1}{n_T}+\frac{1}{n_P}\right)^{-1}\left(\frac{1}{n_T}+\frac{1}{n_A}\right)^{-1}})\ Cov(\bar{X}_T-\bar{X}_A, \bar{X}_T-\bar{X}_P)$$

$$= \sigma^{-2}\sqrt{\frac{n_T n_P}{n_T+n_P}\frac{n_T n_A}{n_T+n_A}}\frac{\sigma^2}{n_T}$$

$$= \sqrt{\frac{c_A c_P}{(1+c_P)(1+c_A)}} \qquad (1.4.6)$$

The three arms are independent and thus the covariance between the means of any two arms is zero. Therefore we have $\mathbf{T} \sim$ Bivariate Normal($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ ), where

$$\boldsymbol{\mu} = \left( \frac{\mu_T - \mu_P}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_P}}}, \frac{\mu_T - \mu_A + \Delta}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_A}}} \right)', \ \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \sqrt{\frac{c_A c_P}{(1+c_P)(1+c_A)}} \\ \sqrt{\frac{c_A c_P}{(1+c_P)(1+c_A)}} & 1 \end{pmatrix}$$

The power function can then be derived as

$$1 - \beta = \Phi_{\boldsymbol{\Sigma}}\left( \frac{\mu_T - \mu_P}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_P}}} - z_{1-\alpha}, \frac{\mu_T - \mu_A + \Delta}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_A}}} - z_{1-\alpha} \right) \qquad (1.4.7)$$

where $\Phi_{\boldsymbol{\Sigma}}$ is the CDF of bivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$**.** The sample size $n_T$ can then be resolved by grid searching the bivariate normal distribution probability function. The total sample size needed to achieve a power of $1 - \beta$ is $N = n_T(1 + c_A + c_P)$.

### *1.4.2 Effect Preservation Test*

The effect preservation approach is also referred to as effect retention test, or fractional test. Rather than assuming a fixed margin, the purpose of this approach is to show that the experimental treatment preserves a fraction $(1 - \lambda)$ of the active comparator effect relative to placebo. The attractive properties of this test was discussed in several works[12,13]. First of all, such design circumvents the issue of

problematic specification of an absolute non-inferiority margin. Such fixed margin was usually estimated based on historical studies. However, there is often difficulty to determine the pre-specified margin due to (1) few pioneer studies on the experimental treatment; (2) the variability of the response from trial to trial; or (3) differences between the analysis populations and medical environment from historical samples and the current study. In addition, after non-inferiority is proved at a pre-specified criterion (e.g., λ =0.2), this design allows a decrease of λ in a continuous manner until statistical significance no longer applies. With $\lambda \leq 0$, one shows that the experimental treatment is "nominally superior" to the reference treatment. The minimum value of λ for the achieved statistical significance at level α is the lower 1- α confidence bound for $\frac{\mu_T-\mu_P}{\mu_A-\mu_P}$.

The formulation of primary hypotheses can be written as follows:

(1) $H_{0,AP}^{(s)}$: $\mu_A \leq \mu_P$ versus $H_{1,AP}^{(s)}$: $\mu_A > \mu_P$

(2) $H_{0,TA}^{(n)}$: $\frac{\mu_T-\mu_P}{\mu_A-\mu_P} \leq 1-\lambda$ versus $H_{1,TA}^{(n)}$: $\frac{\mu_T-\mu_P}{\mu_A-\mu_P} > 1-\lambda$

By re-arranging null hypothesis in step 2 to $\mu_T - \mu_P - (1-\lambda)(\mu_A - \mu_P) \leq 0$, the test statistics can be defined as $T_{TA}^{(n)} = \overline{X}_T - \overline{X}_P - (1-\lambda)(\overline{X}_A - \overline{X}_P) = \overline{X}_T - (1-\lambda)\overline{X}_A - \lambda X_P$. $T_{TA}^{(n)}$ is a linear combination of normal distributed variables, therefore it follows the normal distribution with mean $\mu_T - (1-\lambda)\mu_A - \lambda\mu_P$ and variance $\frac{\sigma^2}{n_T}(1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P})$. $T_{AP}^{(s)}, T_{TA}^{(n)}$ follows bivariate normal distribution based on the

same logic used in 1.4.1. Under the commonly used alternative hypothesis of $\mu_T = \mu_A$, the overall power function can be derived as

$$1-\beta = \Phi_{\boldsymbol{\Sigma}}\left(\frac{\Delta_{\mathrm{AP}}}{\frac{\sigma}{\sqrt{n_T}}\sqrt{\frac{1}{c_A}+\frac{1}{c_P}}} - z_{1-\alpha},\ \frac{\lambda\Delta_{\mathrm{AP}}}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P}}} - z_{1-\alpha}\right) \qquad (1.4.8)$$

Where $\Delta_{\mathrm{AP}} = \mu_A - \mu_P$, $\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{1} & \frac{\frac{\lambda}{c_P}-\frac{1-\lambda}{c_A}}{\sqrt{(\frac{1}{c_A}+\frac{1}{c_P})(1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P})}} \\ \frac{\frac{\lambda}{c_P}-\frac{1-\lambda}{c_A}}{\sqrt{(\frac{1}{c_A}+\frac{1}{c_P})(1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P})}} & \mathbf{1} \end{bmatrix}$, $\Phi_{\boldsymbol{\Sigma}}$ is the CDF of bivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

Mathematically $\lambda$ can take any positive values. Whereas in real world it is almost impossible for a new treatment to get approval if it preserves less than half of the control effect. Therefore all the discussions throughout this thesis will be limited on $\lambda \leq 0.5$. Schwartz and Denne[14] did simulations to show that to achieve the same nominal power only a fraction (<50%) of sample size for non-inferiority test is required for the test of assay sensitivity when $\lambda \leq 0.5$ when the treatment allocation is at $c_A = 1-\lambda$ and $c_P = \lambda$. Therefore in their design the sample size calculation is dominated by the non-inferiority test. Here we will prove this in theory and also extend to see if this conclusion also applies to other treatment allocations.

The sample size needed for the non-inferiority test and for the assay sensitivity test in the experimental treatment arm are respectively

$$n_T^{NIF} = \frac{\left\{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}\right\}\sigma^2(z_{1-\alpha} + z_{1-\beta})^2}{\lambda^2 {\Delta_{AP}}^2}$$

$$n_T^{AP} = \frac{\left\{\frac{1}{c_A} + \frac{1}{c_P}\right\}\sigma^2(z_{1-\alpha} + z_{1-\beta})^2}{{\Delta_{AP}}^2}$$

Therefore, by taking the ratio, we have

$$\frac{n_T^{NIF}}{n_T^{AP}} = \frac{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}}{\left\{\frac{1}{c_A} + \frac{1}{c_P}\right\}\lambda^2}$$

$$= 1 + \frac{1 + \frac{1}{c_A} - \frac{2\lambda}{c_A}}{\left\{\frac{1}{c_A} + \frac{1}{c_P}\right\}\lambda^2} \qquad (1.4.9)$$

When $\lambda < 0.5$, the ratio is always greater than 1.

Specifically at the allocation of $c_A = 1 - \lambda$ and $c_P = \lambda$,

$$\frac{n_T^{NIF}}{n_T^{AP}} = \frac{1 + 1 - \lambda + \lambda}{\left\{\frac{1}{1-\lambda} + \frac{1}{\lambda}\right\}\lambda^2} = \frac{2(1-\lambda)}{\lambda} = \frac{2}{\lambda} - 2$$

As $0 < \lambda \leq 0.5$, we have $2 \leq \frac{n_T^{NIF}}{n_T^{AP}} < \infty$, i.e., the sample size needed for the assay sensitivity is $\leq 50\%$ of the sample size for non-inferiority test.

**Table 1** Ratio of sample size required to test non-inferiority over sample size for the assay sensitivity test at the same nominal power

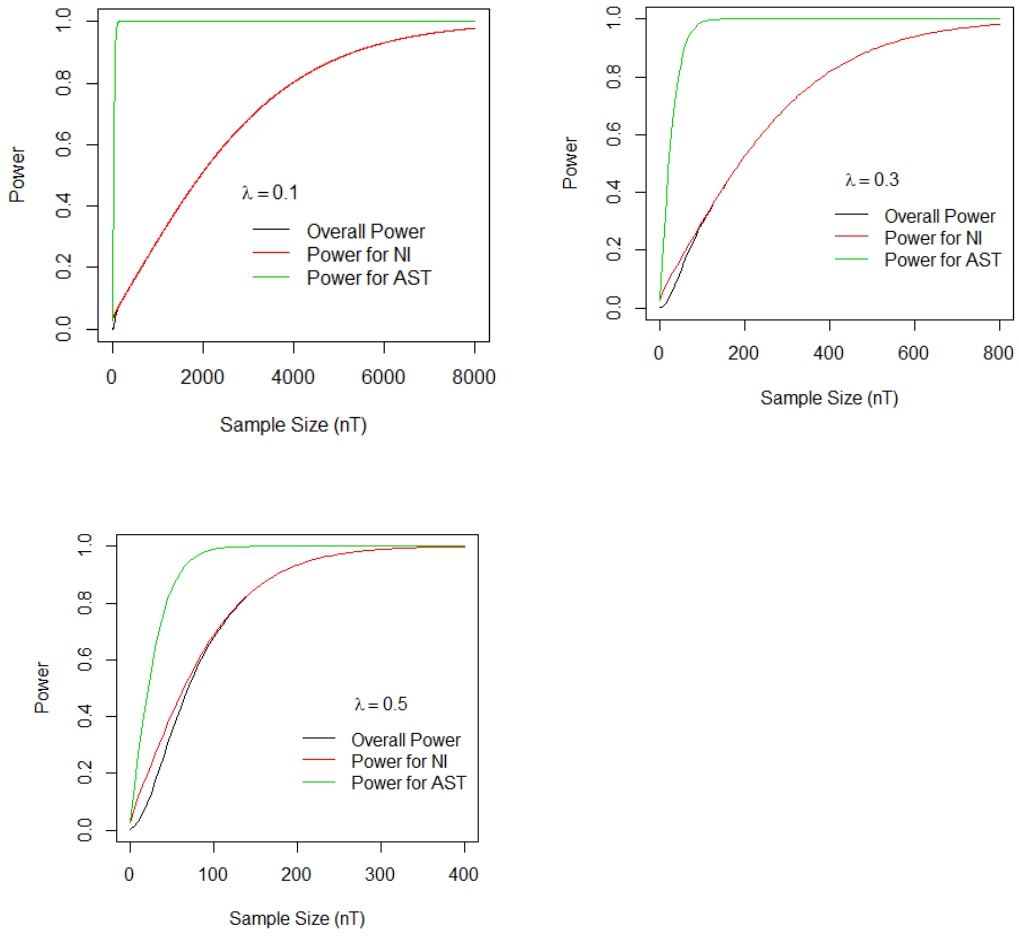
| **Allocation** | **λ** | | | | |
|---|---|---|---|---|---|
| $n_T : n_A : n_P$ | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** |
| **1:1:1** | 91 | 21 | 8.8 | 4.75 | 3 |

| | | | | | |
|---|---|---|---|---|---|
| **2:2:1** | 61 | 14 | 6.2 | 3.5 | 2.3 |
| **3:3:1** | 46 | 11 | 4.9 | 2.9 | 2 |
| **4:4:1** | 37 | 9 | 4.1 | 2.5 | 1.8 |
| **5:5:1** | 31 | 7.7 | 3.6 | 2.25 | 1.7 |
| **2:1:1** | 66 | 14.8 | 6 | 3.2 | 2 |
| **4:2:1** | 44 | 10 | 4.3 | 2.4 | 1.7 |

We cannot enumerate all possible treatment allocations, but Table 1 listed out the ratios for the most commonly seen allocations, including the balanced design ($n_{\mathrm{T}}: n_A: n_P = 1: 1: 1$) and partially balanced designs ($n_{\mathrm{T}}: n_A = 1: 1$). It is observed that the ratio decreases as λ increases. When λ ≥ 0.4, the sample size for the test of assay sensitivity is less than half of that for the test of non-inferiority test. Only at λ = 0.5 and at allocations where the treatment is at least four times placebo the ratios are lower than 2, but still close to 2.

In addition, the power functions were plotted and compared. It is shown that the power for the assay sensitivity test is always dramatically higher than the power for the non-inferiority test. As the power goes higher the overall power function curve overlap with the power function curve for the non-inferiority test.

**Figure 1 Plots of power function curves for the non-inferiority test (NI), the assay sensitivity test (AST) and overall power for both tests**




Due to this observation, the sample size for such study is dominated by the three-arm non-inferiority test. And this conclusion will be the basis for the initial sample size calculation for adaptive design discussed later.

The discussion so far has all assumed that the test statistics are normally distributed. However, the sample size is not always big enough to apply the Central

Limit Theory. Pigeot and et al.[10] published their work that is applicable to more generalized scenarios. In their study, when $\sigma^2$ is unknown and replaced with sample variance, non-central T-test with degree freedom of $n_T + n_C + n_P - 3$ was used to calculate the exact power and sample size for the non-inferiority test. Under various parameter combinations explored, sample size per group compared to normal approximation always differed by one only.

### 1.*4.3 Comparison of Power for Fixed Margin Test and Effect Preservation Test*

It is noteworthy that fixed margin test and effect preservation test can be transformed to one another. By replacing Δ with $\lambda(\mu_A - \mu_P)$, the fixed margin test then takes the form of effect preservation test. Therefore there has been interest in comparing the power of the two tests at equivalent margin. Let's use $1 - \beta^F$ to denote power for the fixed margin test, and $1 - \beta^E$ for effect preservation test. Assume $\mu_T - \mu_A$, we have

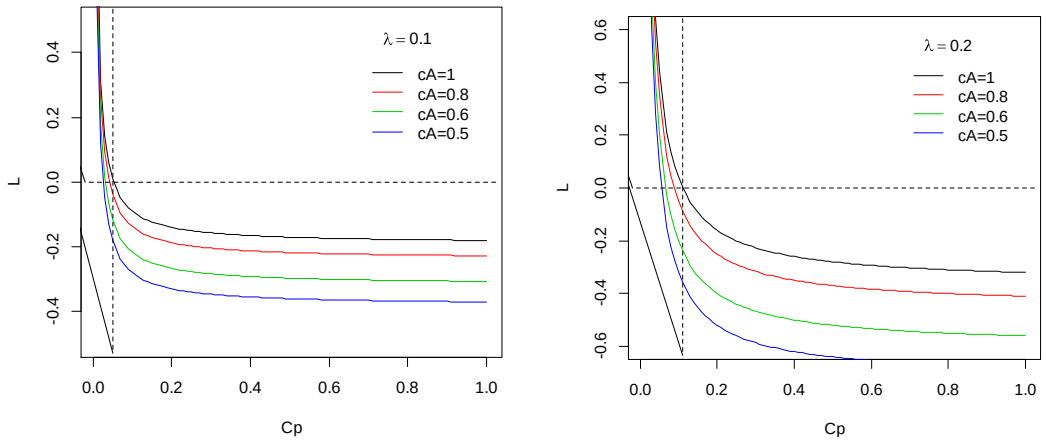
$$1 - \beta^F = \Phi\left(\frac{\Delta}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_A}}} - z_{1-\alpha}\right) = \Phi\left(\frac{\lambda(\mu_A-\mu_P)}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_A}}} - z_{1-\alpha}\right) \qquad (1.4.3.1)$$

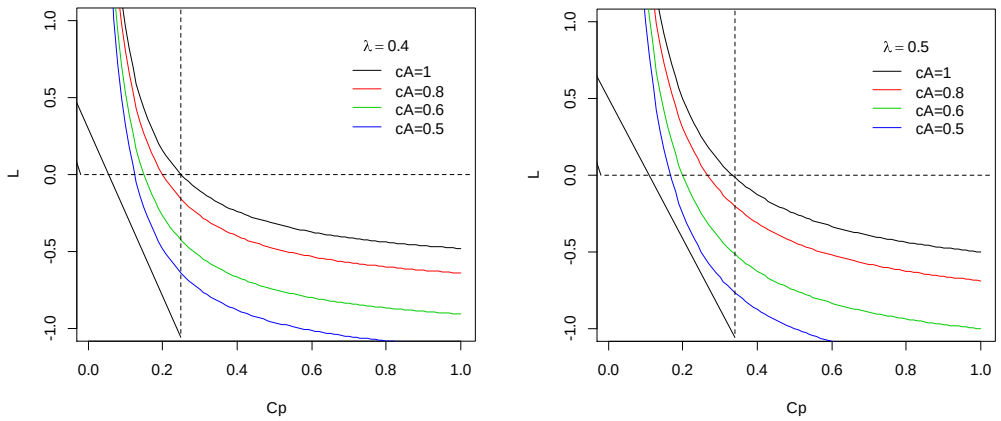
$$1 - \beta^E = \Phi\left(\frac{\lambda(\mu_A-\mu_P)}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P}}} - z_{1-\alpha}\right) = \Phi\left(\frac{\lambda(\mu_A-\mu_P)}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_A}+\lambda(\frac{\lambda}{c_A}-\frac{2}{c_A}+\frac{\lambda}{c_P})}} - z_{1-\alpha}\right) \qquad (1.4.3.2)$$

By comparing (1.4.3.1) and (1.4.3.2) we can find that whether $1-\beta^E$ is larger or smaller than $1-\beta^F$ depends on the sign of $\frac{\lambda}{c_A}-\frac{2}{c_A}+\frac{\lambda}{c_P}$. Let L=$\frac{\lambda}{c_A}-\frac{2}{c_A}+\frac{\lambda}{c_P}$. It can be easily derived that when $\frac{c_A}{c_P}=\frac{2-\lambda}{\lambda}$ the power for the two forms with numerically equal margin are the same. When $\frac{c_A}{c_P}<\frac{2-\lambda}{\lambda}$, L is negative, i.e., the power for the effect preservation test is higher; When $\frac{c_A}{c_P}>\frac{2-\lambda}{\lambda}$, L is positive, i.e., the power for the fixed margin test is higher. For $\lambda$=0.1, 0.2, 0.3, 0.4 and 0.5, the effect preservation test wins this power contest when the ratio of active comparator to placebo is no greater than 19, 9, 5.6, 4, and 3.

The plots below showed when the sign of L switches at different combination of $c_A, c_P$ and $\lambda$.

**Figure 2** L=$\frac{\lambda}{c_A}-\frac{2}{c_A}+\frac{\lambda}{c_P}$ over $c_P$ at different combinations of $\lambda$ , $c_A$






As indicated in the equation, at fixed value of $\lambda$ and $c_A$, L is a monotone decreasing function of $c_P$. In real world it is hard to justify any $\lambda$ values greater 0.5, and $c_A$ is often between 0.5 and 1. In such case the biggest $c_P$ that keeps L positive occurs at $c_A = 1$ for all values of $\lambda$.

### 1.5 Optimal Sample Size Allocation

It is well known that in two-arm studies equal allocation is used most frequently because it provides higher efficiency than unequal allocation. Whereas in the three-arm trials, in certain circumstances unequal allocation may be preferred choice.

(1) Given limited budget, one treatment may be much more expensive than the alternative, the trial can enroll more subjects if more people are allocated to the cheaper treatment.

(2) There are more unknown things about the new treatment. Assigning more subjects to the new treatment allows for higher power to detect any unknown side-effects.

(3) When placebo group is introduced to evaluate assay sensitivity, it is often desirable to keep the placebo group minimum for ethical reasons.

(4) In three-arm designs, with pre-specified power, the most efficient allocation that results in the smallest total sample size is not equal allocation. Instead, it can be derived under different constraints and is referred to as optimal allocation. Pigeot et al.[10] looked into optimal allocation for the effect preservation non-inferiority test. As the total sample size can be calculated using the formula (based on the $n_T^{NIF}$ derived in 1.4.2)

$$N = \frac{\left\{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}\right\}\sigma^2\left(z_{1-\alpha} + z_{1-\beta}\right)^2(1 + c_A + c_P)}{\lambda^2{\Delta_{AP}}^2}$$

Under the restriction of $c_A = 1$(i.e., equal sample size of treatment and active comparator), by taking the first and second derivative the equation above, the minimum of N is achieved at $c_P = \frac{\sqrt{2}\lambda}{\sqrt{(1-\lambda)^2+1}}$. For example, $\lambda = 0.3$ will lead to an restricted optimal allocation of 2.9:2.9:1, which means only 15% of the analysis population will receive placebo. Without restriction, the minimum N is achieved at $c_A = 1 - \lambda$ and $c_P = \lambda$. In such case, $N = 2n_T$.

## CHAPTER II: GROUP SEQUENTIAL DESIGN AND SAMPLE SIZE RECALCULATION

### 2.1 Introduction and Background

In traditional clinical trials, the sample size has to be determined in advance with pre-specified significance level, power and treatment effect to be detected, and data has to be collected before analysis can proceed. However, there may be large uncertainty in key factors that drive the sample size estimation. In addition, the conventional fixed sample design may not be efficient as the recruitment of patients usually takes a period of time and the data become available steadily and sequentially. For some rare diseases, the recruitment could take years. Sequential trial designs were introduced which allowed the accumulated data to be analyzed at a series of planned interim analyses during the course. Sequential trials can be adaptive or non-adaptive. Non-adaptive sequential designs are often called group sequential designs. The trial can be stopped at any planned interim analysis for excessive efficacy or futility. Adaptive designs allow more flexibility, permitting changes to some design features at some points during the trial. Allowing sample size adaptation helps mitigate the risk of under-powered or over-powered study.

### 2.2 Group Sequential Design

In the group sequential designs repeated significance testing will be performed. Repeated testing would inflate the overall type I error if the same critical values as in fixed sample design are used in the hypothesis testing at all stages. Armitage, McPherson and Rowe are the first authors who investigated this issue

systematically[15]. They considered sequential observations of binomial, normal and exponential distribution forms, and concluded that when null hypothesis is true the probability of obtaining a significant result goes above the nominal significance level at repeated tests on the accumulated data. Table 2 below showed the overall type I error for repeated tests in sampling from a normal distribution with known variance.

**Table 2 Overall type I error for repeated tests**

| Number of repeated tests at 2-sided 0.05 level | Overall Significance Level if Null Hypothesis of No Treatment Difference is True |
|---|---|
| 1 | 0.05 |
| 2 | 0.08 |
| 3 | 0.11 |
| 4 | 0.13 |
| 5 | 0.14 |
| 10 | 0.19 |
| 20 | 0.25 |
| 50 | 0.33 |

In the group sequential setting with K groups of observations in two-arm studies, without loss of generosity, the responses of treatment and control groups are assumed to be normally distributed with means $\mu_T$ and $\mu_C$ respectively and known common variance; let the mean difference $\theta = \mu_T - \mu_C$ . Under null hypothesis $= 0$ , the test statistic at each stage is $T_k$ which follows standard normal distribution, the

critical value is $c_k$, the overall type I error is the probability of rejecting null hypothesis at least once. It follows straightforwardly that

$$P(\cup_{k=1}^{K}\{T_k > c_k\}) = 1 - P(\cap_{k=1}^{K}\{T_k < c_k\})$$

To avoid the inflation of type I error, the critical value $c_k$ must be adjusted to a value bigger than $\Phi^{-1}(1-\alpha)$. Pocock (1977)[16] followed the same method of numerical quadrature in the work of Armitage et al. to solve the critical value on equally spaced information levels. In this situation the critical value is a constant for all stages. However this may not be the most practical approach considering that the early stages had small sample size and thus is more difficult to declare statistical significance. Therefore, O'Brien and Fleming (1979) proposed a procedure with rejection criteria that goes more stringent over the stages. This procedure has conservative stopping boundary values at very early stages, and boundary values close to the fixed design at the final stage. In this situation, the boundary at stage k is proportional to $\sqrt{K/k}$. This approach is recommended by FDA in their 2010 Guidance on Adaptive Designs. Wang and Tsiatis[17], Emerson and Fleming[18] and Pampallona and Tsiatis[19] generalized the Pocock and O'Brien-Fleming methods to the power family, which use boundary values of $(\frac{k}{K})^{-\rho}C$, where $C$ is a derived constant. $\rho = 0$ is the case of Pocock method and $\rho = 0.5$ is the case of O'Brien-Fleming method.

All these above methods require pre-specifying the total number of decision times K. The timing of interim analyses is information time rather than calendar time which creates challenges in scheduling in practice. Also it is possible that a trial may change the frequency of data review due to slower recruitment or other reasons. To add more flexibility, Lan and DeMets[20] proposed a procedure based on an error spending function $\alpha^*(t)$, where t is the information fraction at each stage, which can be approximated by the proportion of sample size at the interim analysis out of overall sample size. Under O'Brien & Fleming philosophy, $\alpha^*(t) = d(1 - \Phi\left(\frac{Z_\alpha}{\frac{d}{\sqrt{t}}}\right))$, where d=2 for one-sided test with significance level $\alpha$ and d=4 for two-sided test with significance level $\alpha$. Under Pocock philosophy $\alpha^*(t) = \alpha\ln[1 + (e - 1)t]$.

At each interim analysis, if the significance is claimed with the family-wise type I error controlled, the trial can be stopped for excessive efficacy. That will not only cut the cost but also benefit the patients so that they could have access to the novel effective treatment earlier. In contrary, if the interim result turns out to be not promising at all, the trial may also be terminated for futility. To quantify this stopping criterion, the concept of conditional power was brought about. Conditional power is defined as the statistical power to reject the null hypothesis conditioned on the data observed at the interim stage. With pre-determined critical value or the significance level calculated from alpha spending function, the conditional power

can be calculated at each interim analysis and can be used as an important parameter to guide the next step of a trial.

### 2.3 Sample Size Recalculation

Sample size recalculation is a type of adaptive design that adds additional flexibility to group sequential design. In the initial design stage there is often considerate uncertainty associated with the initially assumed treatment effect due to lack of knowledge or data from past studies, medical practice improvement, patient population difference and so on. In such cases the initial sample size calculated at the design stage may not provide adequate power or lead to inefficient overpowered studies. To mitigate such risk, a number of methods have been proposed to re-calculate the sample size at interim stages. In this approach, one starts out with a relatively small initial sample size based on optimistic estimation of treatment effect and adapt the sample size if the interim analysis indicate that the initial assumption overestimate or underestimate the treatment effect.

The most commonly used sample size re-calculation(SSR) methods include Cui-Hung-Wang (CHW) method[21] and conditional power method[22,23,24].

Cui-Hung-Wang (CHW) method[21]

Cui et al developed an adaptive design that allow re-calculating the sample size at the interim stage by substituting the assumed treatment effect with the estimate from the interim data. With simulations, they showed that the family-wise type I

error can substantially increase if the final test is performed as in fixed design. Therefore they devised a modified final test statistics, which is a weighted average of the interim test statistic $Z_1$ and initial final test statistic $Z_2$. It was demonstrated that using the modified final test statistic the type I error probability is well preserved at the target level. Their approach can be generalized to any group sequential test based on the repeated significance test that can be asymptotically expressed as a Brownian motion process. The downside is that as patients evaluated after the interim stage received less weight at the final analysis, it violates the sufficiency principle. Therefore, Chen et-al.[25] proposed in their work a modified weighted Z-statistic method. The null hypothesis can be rejected if both the weighted and un-weighted Z-statistic at the final stage are greater than the critical value. The type I error rate for this modified approach is less than the nominal $\alpha$. In their investigation, the power loss due to the additional requirement is ignorable.

Denne Method[22]

In this method, the sample size recalculation is based on the idea of conditional power. The conditional power is defined as the probability of rejecting the hypothesis in the final stage conditioned on the data observed in the interim stage. The conditional power for two-arm two stage group sequential design can be calculated from the equation

$$CP_{\theta}(n_2,\ c_2|z_1) = 1 - \Phi\left[\frac{c_2\sqrt{n_2} - z_1\sqrt{n_1} - \frac{n_2 - n_1}{\sqrt{2\sigma^2}}\theta}{\sqrt{n_2 - n_1}}\right]$$

where $\theta$ is the treatment difference, $n_1$ and $n_2$ are respectively sample size at the first and second stage, $z_1$ and $c_2$ are respectively the critical value for the first and second stages.

Initially at the design stage, the final critical value is set at $\widetilde{c_2}$ for a pre-specified significance level. The sample size at first and second stage are calculated based on the assumed treatment effect and variance. At the interim stage, a so called "targeted" sample size ($n_t$) is calculated by substituting in the sample variance estimated from the interim data. Then the final sample size $n_2$ and critical value $c_2$ to preserve overall type I error is found to meet the equation

$CP_{\theta=0}(n_t, \widetilde{c_2}|z_1) = CP_{\theta=0}(n_2, c_2|z_1)$ and

$CP_{\theta=\Delta}(n_2,\ c_2|z_1)$ = desired conditional power

This method preserves the type I error and improves unconditional power compared to CHW method. However, it is not widely used in practice because its complex statistical adjustment leads to difficulties in interpreting and presenting trial findings in the regulatory submissions.

Promising Zone Method[23,26,24]

Chen *et al*. [23] further developed the conditional power method and showed that if the interim results are promising, the sample size can be increased without adjusting the final hypothesis test critical value. In their work it was found that the conventional hypothesis test can be performed without inflating overall type I error if the conditional power is greater than 50%. This finding makes the application of the sample size adaptation design more promising because the final stage hypothesis test is non-controversial, intuitive, and more interpretable to clinicians and other non-statistician background reviewers. Gao *et al.*[26] extended this finding to a broader range of conditional power and called that range promising zone. They proved in theory that when the conditional power falls into a certain range, the final critical value can be adjusted to a value lower than the original planned value to obtain a size α test. In such case, keep the original planned critical value will preserve the overall type I error. They also showed how to find out the thresholds of this range. Mehta *et al.*[24] made that work more accessible to practitioners by presenting it in the context of two-stage designs, tabulating explicit cut-off values for the promising zone under different adaptive rules based on conditional power, and using two actual case studies to demonstrate how to apply and interpret this method.

**CHAPTER III SAMPLE SIZE RECALCULATION IN THREE-ARM NON-INFERIORITY TRIALS**

## 3.1 Effect Preservation Test, Normal Endpoint

As discussed in chapter I, when the non-inferiority takes the form of effect preservation test, rejection of the assay sensitivity test is a mathematical pre-requisite for rejection of the non-inferiority test. Therefore, in this discussion, the hypotheses testing procedure will include

$H_{0,AP}^{(s)}$: $\mu_A \leq \mu_P$ versus $H_{1,AP}^{(s)}$: $\mu_A > \mu_P$

$H_{0,TA}^{(n)}$: $\frac{\mu_T - \mu_P}{\mu_A - \mu_P} \leq 1 - \lambda$ versus $H_{1,TA}^{(n)}$: $\frac{\mu_T - \mu_P}{\mu_A - \mu_P} > 1 - \lambda$

and fixed sequence testing will be used at each stage to preserve the stage-wise type I error.

The formula of

$$N = n_T(1 + c_A + c_P) = \frac{\left\{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}\right\}\sigma^2\left(z_{1-\alpha} + z_{1-\beta}\right)^2(1 + c_A + c_P)}{\lambda^2 {\Delta_{AP}}^2}$$

discussed in chapter 1.4 will be used for the initial sample size calculation. In practice there is little interest in a new treatment that preserves less than 50% effect of the active comparator, therefore all the discussions below will be restricted at $\lambda < 0.5$.

### 3.1.1 Method 1: Two-Stage SSR Based on the Observed Treatment Effect of the Active Comparator Effect ($\widehat{\Delta}_{AP}^{(1)}$)

Schwartz and Denne[14] proposed a two-stage SSR method for three-arm Non-Inferiority Trial in such setting. In their method, at the interim stage, the sample size will be re-calculated through replacing the initial $\Delta_{AP}$ with $\widehat{\Delta}_{AP}^{(1)}$ computed from the interim data. At the final stage, the assay sensitivity and non-inferiority tests will be performed sequentially at a significance level of α.

*3.1.1.1 Overall Type I Error Preservation*

**<u>At Optimal allocation</u>**

It is proved by Schwartz and Denne[14] that this procedure preserves overall type I error at α when the sample size allocation is at $n_A$=k(l-λ)$n_T$, and $n_P$ =kλ$n_T$ (k=1 is optimal allocation). The proof is based on the notion that recalculating the sample size using an estimate of a parameter that is independent of the first-stage test statistic will not substantially inflate the type I error at the final stage.

Proof:

Let $Z^{(1)}$ and $Z^{(2)}$ be the test statistics at the interim and final stage respectively, $Z^{(2-1)}$ be the test statistic constructed with only the data at the second stage.

$$Z^{(1)} = \frac{\bar{X}_T^{(1)} - (1-\lambda)\bar{X}_A^{(1)} - \lambda\bar{X}_P^{(1)}}{\frac{\sigma}{\sqrt{n_T^{(1)}}}\sqrt{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}}}$$

$$Z^{(2)} = \frac{\bar{X}_T^{(2)} - (1-\lambda)\bar{X}_A^{(2)} - \lambda\bar{X}_P^{(2)}}{\frac{\sigma}{\sqrt{n_T^{(2)}}}\sqrt{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}}}$$

$$Z^{(2-1)} = \frac{\bar{X}_T^{(2-1)} - (1-\lambda)\bar{X}_A^{(2)} - \lambda\bar{X}_P^{(2)}}{\frac{\sigma}{\sqrt{n_T^{(2)}}}\sqrt{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}}}$$

$Z^{(2)}$ can be written as a weighted average of $Z^{(1)}$ and $Z^{(2-1)}$

$$Z^{(2)} = \sqrt{\frac{n_T^{(1)}}{n_T^{(2)}}}Z^{(1)} + \sqrt{\frac{n_T^{(2-1)}}{n_T^{(2)}}}Z^{(2-1)} \qquad (3.1.1)$$

under $H_0$, $Z^{(1)}$, $Z^{(2)}$ and $Z^{(2-1)}$ all follows standard normal distribution. $Z^{(2-1)}$ is independent of $\hat{\Delta}_{AP}^{(1)}$ as it is solely formed from the data at the second stage,

$$Overall\ Type\ I\ error = P\left(Z^{(2)} > z_{1-\alpha} | \hat{\Delta}_{AP}^{(1)}, H_0\right)$$

, and

$$Z^{(1)} = \frac{\bar{X}_T^{(1)} - (1-\lambda)\bar{X}_A^{(1)} - \lambda\bar{X}_P^{(1)}}{\frac{\sigma}{\sqrt{n_T^{(1)}}}\sqrt{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}}} \sim N(0,1) \text{ under } H_0$$

$$\hat{\Delta}_{AP}^{(1)} = \bar{X}_A^{(1)} - \bar{X}_P^{(1)} \sim N\left(\Delta_{AP}, \frac{\sigma^2}{n_T^{(1)}}\left(\frac{1}{c_A} + \frac{1}{c_P}\right)\right)$$

Any linear combination of $Z^{(1)}$ and $\hat{\Delta}_{AP}^{(1)}$ are linear combination of $\bar{X}_A^{(1)}$, $\bar{X}_T^{(1)}$ and $\bar{X}_P^{(1)}$, which follows normal distribution. Therefore $Z^{(1)}$ and $\hat{\Delta}_{AP}^{(1)}$ are jointly bivariate

normal.

Let C= > 0

$$Cov\left(Z^{(1)}, \hat{\Delta}_{AP}^{(1)}\right) = \left(\frac{\sigma}{\sqrt{n_T^{(1)}}}\sqrt{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}}\right)^{-1} Cov\left(\bar{X}_T^{(1)} - (1-\lambda)\bar{X}_A^{(1)} - \lambda\bar{X}_P^{(1)}, \bar{X}_A^{(1)} - \bar{X}_P^{(1)}\right)$$

$$= \frac{\left[-\frac{(1-\lambda)}{c_A} + \frac{\lambda}{c_P}\right]}{\sqrt{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}}} \qquad (3.1.2)$$

At optimal allocation, $c_A = 1 - \lambda$, $c_P = \lambda$, $Cov\left(Z^{(1)}, \hat{\Delta}_{AP}^{(1)}\right) = 0$ according to (3.1.2)

therefore $Z^{(2)}$ is independent with $\hat{\Delta}_{AP}^{(1)}$ as well. $P\left(Z^{(2)} > z_{1-\alpha} | \hat{\Delta}_{AP}^{(1)}, H_0\right) =$

$P\left(Z^{(2)} > z_{1-\alpha} | H_0\right) = \alpha$

**<u>At other allocations</u>**

In this thesis, it is further demonstrated that such two-stage SSR will not have type I error inflation when the allocation met the condition: $\frac{c_P}{c_A} \geq \frac{\lambda}{(1-\lambda)}$.

Proof:

Under this condition,

$$Cov\left(Z^{(2)}, \hat{\Delta}_{AP}^{(1)}\right) = Cov\left(\sqrt{\frac{n_T^{(1)}}{n_T^{(2)}}} Z^{(1)}, \hat{\Delta}_{AP}^{(1)}\right) = \frac{\sqrt{\frac{n_T^{(1)}}{n_T^{(2)}}}\left[-\frac{(1-\lambda)}{c_A} + \frac{\lambda}{c_P}\right]}{\sqrt{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}}} \qquad (3.1.3)$$

At the condition of $\frac{c_P}{c_A} \geq \frac{\lambda}{(1-\lambda)}$, $Cov\left(Z^{(2)}, \hat{\Delta}_{AP}^{(1)}\right)$ is negative. Let $\rho =$

$\frac{Cov\left(Z^{(2)}, \hat{\Delta}_{AP}^{(1)}\right)}{\sqrt{Var(Z^{(2)}) * Var(\hat{\Delta}_{AP}^{(1)})}} \leq 0$. Given that the hypothesis for assay sensitivity must be

rejected before non-inferiority test can be performed, we know $\hat{\Delta}_{AP}^{(1)} > z_{1-\alpha}\sqrt{\frac{2\sigma^2}{n_T^{(1)}}} > 0$.

As the $Z^{(2)}$ and $\hat{\Delta}_{AP}^{(1)}$ jointly follow bivariate normal distribution, we can derive the conditional distribution

$$Z^{(2)} | \hat{\Delta}_{AP}^{(1)}, H_0 \sim N\left(\rho \frac{\sigma_{Z^{(2)}}}{\sigma_{\hat{\Delta}_{AP}^{(1)}}} \hat{\Delta}_{AP}^{(1)},\ 1-\rho^2\right) \qquad (3.1.4)$$

We then can derive the upper limit of the rejection probability at the final stage conditioned on $\hat{\Delta}_{AP}^{(1)}$ and $H_0$

$$P\left(Z^{(2)} > z_{1-\alpha} | \hat{\Delta}_{AP}^{(1)}, H_0\right)$$

$$= P\left(\frac{Z^{(2)} - \rho \frac{\sigma_{Z^{(2)}}}{\sigma_{\hat{\Delta}_{AP}^{(1)}}} \hat{\Delta}_{AP}^{(1)}}{\sqrt{1-\rho^2}} > \frac{z_{1-\alpha} - \rho \frac{\sigma_{Z^{(2)}}}{\sigma_{\hat{\Delta}_{AP}^{(1)}}} \hat{\Delta}_{AP}^{(1)}}{\sqrt{1-\rho^2}} \middle| \hat{\Delta}_{AP}^{(1)}, H_0\right)$$

$$= 1 - \Phi\left(\frac{z_{1-\alpha} - \rho \frac{\sigma_{Z^{(2)}}}{\sigma_{\hat{\Delta}_{AP}^{(1)}}} \hat{\Delta}_{AP}^{(1)}}{\sqrt{1-\rho^2}}\right)$$

Under the condition of $\rho \leq 0$, $\hat{\Delta}_{AP}^{(1)} > 0$, we have $z_{1-\alpha} - \rho \frac{\sigma_{Z^{(2)}}}{\sigma_{\hat{\Delta}_{AP}^{(1)}}} \hat{\Delta}_{AP}^{(1)} \geq z_{1-\alpha}$

Also $\sqrt{1-\rho^2} \leq 1$, therefore $\frac{z_{1-\alpha} - \rho \frac{\sigma_{Z^{(2)}}}{\sigma_{\hat{\Delta}_{AP}^{(1)}}} \hat{\Delta}_{AP}^{(1)}}{\sqrt{1-\rho^2}} > z_{1-\alpha}$,

$$1-\Phi\left(\frac{z_{1-\alpha}-\rho\frac{\sigma_{Z^{(2)}}}{\sigma_{\hat{\Delta}_{AP}^{(1)}}}\hat{\Delta}_{AP}^{(1)}}{\sqrt{1-\rho^2}}\right)\leq 1-\Phi(z_{1-\alpha})=\alpha$$

## Simulation Results

Simulations also demonstrated that when the allocations yield non-positive correlation between $Z^{(1)}$ *and* $\hat{\Delta}_{AP}^{(1)}$, increasing sample size would not inflate the overall type I error. In contrary, the inflation can be observed when $\frac{c_P}{c_A}<\frac{\lambda}{(1-\lambda)}$. The false rejection rate in simulations with 100,000 trials under different $\lambda$s and allocations are shown in table 3 below.

**Table 3 False rejection rate in simulations under different allocations.**

| | | $n_T:n_A:n_P$ | | | |
|---|---|---|---|---|---|
| $\lambda$ | Optimal 1:(1-$\lambda$): $\lambda$ | 1:1:1 | 2:1:1 | 2:2:1 | 4:4:1 |
| 0.1 | 2.47% | 2.39% | 2.35% | 2.40% | 2.46% |
| 0.2 | 2.49% | 2.29% | 2.30% | 2.39% | 2.51% |
| 0.3 | 2.52% | 2.34% | 2.36% | 2.43% | 2.67% |
| 0.4 | 2.46% | 2.43% | 2.32% | 2.59% | 2.87% |
| 0.5 | 2.54% | 2.50% | 2.54% | 2.64% | 3.04% |

Note: Simulations performed at $\lambda=0.3$; $\Delta_{\text{AP,Assumed}}=0.6$, nominal $\alpha$=2.5%, number of simulations=100,000

*3.1.1.2 Power and Efficiency*

Schwartz and Denne[14] showed in simulation that at the optimal allocation, when the $\Delta_{AP}$ is over-estimated at the initial stage, leading to inadequate power, this SSR method can increase the actual power close to the desired value. For instance, when the true $\Delta_{AP}$ was only $1/\sqrt{2}$ times the assumed value at the design stage, the sample size given by fixed design provides only 57% power. In contrast, by recalculating the sample size at the interim stage while half of the initially planned subjects are enrolled, the power goes up to 75-79%, depending on the value of λ.

The efficiency of a study is often characterized by the average sample size. It was shown by simulations that, compared the average sample size for fixed study with the same power, this two-stage SSR method needs 2- 40% more subjects. The smaller value of λ, the more efficient this method is.

### 3.1.2 Method 2: SSR When the Conditional Power Falls Into Promising Zone

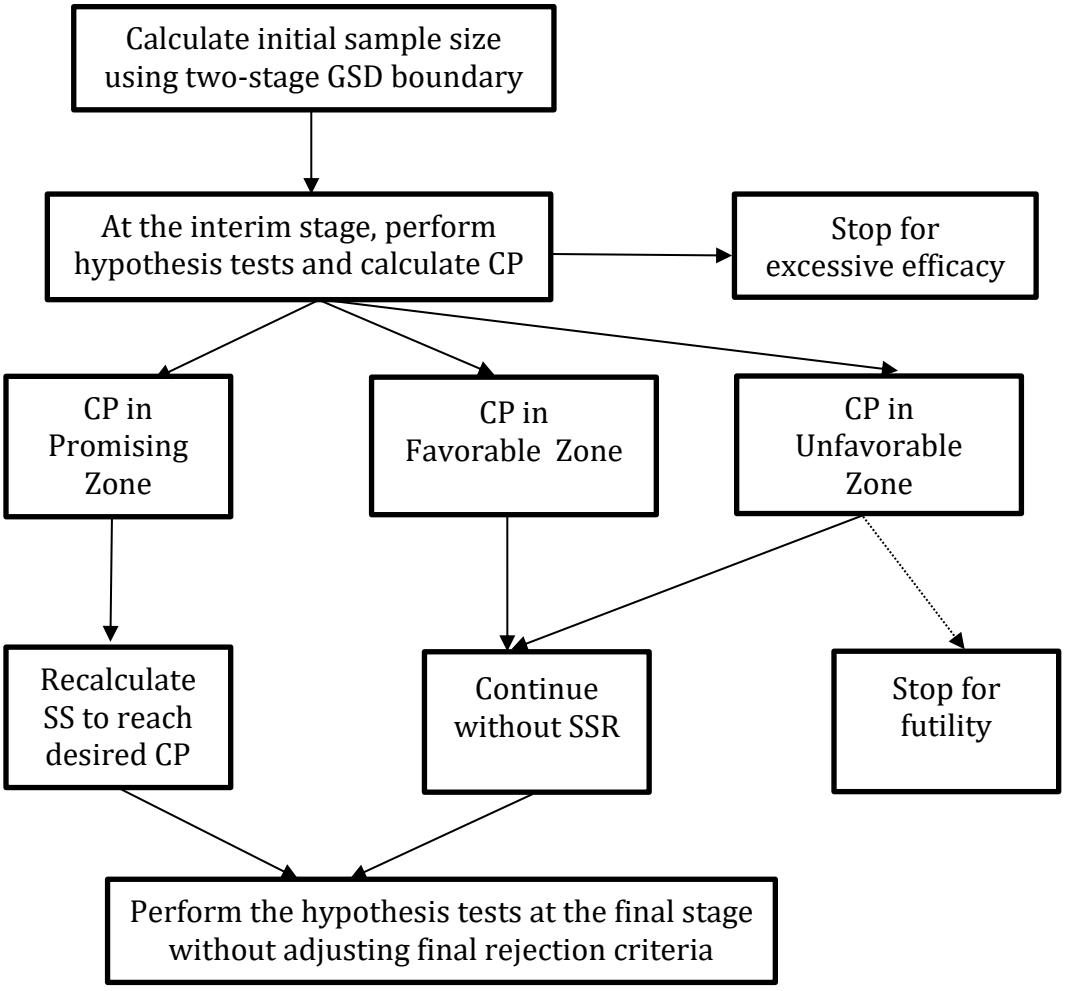
The two-stage SSR method based on $\hat{\Delta}_{AP}^{(1)}$ works well where there is large uncertainty of the assay sensitivity. In Schwartz and Denne's work, the performance of the methods were evaluated under the assumption that experimental treatment effect is the same as the active comparator. That is not always true in non-inferiority study. When there is also great uncertainty in estimating the experimental treatment effect, which is common in clinical studies, there is a need for an SSR

method that allows for adjusting the sample size based on interim results of the experimental treatment arm.

A conditional power based SSR method was then developed for three-arm non-inferiority trials in this dissertation. It borrows the idea of promising zone to simplify the testing procedure and for easier interpretation.

This SSR design requires testing the hypothesis at the interim stage, therefore the group sequential design boundaries, such as Pocock boundary, O'Brien-Fleming boundary and etc. will be used to calculate the initial sample size. At the interim stage, the hypothesis will be tested against pre-specified critical value at the first stage. If both rejected, the study success may be claimed for excessive efficacy. Otherwise, the conditional power for the non-inferiority test is calculated. If it is within the promising zone, the sample size will be recalculated to achieve a desired conditional power, often set as the same as the initial nominal power. Otherwise the study will go on without sample size recalculation. One could also pre-specify a lower bound for the conditional power to allow the study stop for futility. At the final stage, the hypotheses can be performed without critical value adjustment. The flow chart below showed the whole procedure.

**Figure 2 Flow chart of method 2: two-stage SSR design based on conditional power**



*3.1.2.1 Conditional Power for Effect Preservation Non-Inferiority Test*

A key elements for this design is the conditional power calculation, which is used to guide the directions after interim analysis, and is the basis for sample size re-calculation. The analytic formula of conditional power for the non-inferiority test can be obtained as

$$CP^{NIF} = 1 - \Phi\{\frac{c_2\sqrt{n_T^{(2)}} - z_1\sqrt{n_T^{(1)}}}{\sqrt{n_T^{(2)} - n_T^{(1)}}} - \frac{\theta}{\sigma\sqrt{\frac{1}{n_T^{(2)} - n_T^{(1)}}\left\{1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}\right\}}}\} \quad (3.1.5)$$

Where $\theta = \mu_T - (1-\lambda)\mu_A - \lambda\mu_P$, which equals $\lambda\Delta_{AP}$ under alternative hypothesis; $n_T^{(1)}$ is the number of subjects in the experimental treatment group in stage 1 (interim stage), and , $n_T^{(2)}$ is the total number subjects from stage 1 and 2 combined; $c_2$ is the critical value at the final stage; $z_1$is the calculated value of non-inferiority test statistic at the interim stage. The derivation of the analytic form of the conditional power is shown as follows:

For simplicity, let M=$1 + \frac{(1-\lambda)^2}{c_A} + \frac{\lambda^2}{c_P}$

The test statistic at stage 1 is $z_1 = \frac{\bar{X}_T^{(1)} - (1-\lambda)\bar{X}_A^{(1)} - \lambda\bar{X}_P^{(1)}}{\sigma\sqrt{\frac{M}{n_T^{(1)}}}}$

By definition conditional power is the probability of rejection at the final stage conditioned on the interim data $z_1$,

$$CP = P\left(\frac{\bar{X}_T - (1-\lambda)\bar{X}_A - \lambda\bar{X}_P}{\sigma\sqrt{\frac{M}{n_T^{(2)}}}} > c_2 | Z_1 = z_1\right)$$

$$= P\left(\frac{n_T^{(1)}}{n_T^{(2)}}\left(\bar{X}_T^{(1)} - (1-\lambda)\bar{X}_A^{(1)} - \lambda\bar{X}_P^{(1)}\right) + \left(1 - \frac{n_T^{(1)}}{n_T^{(2)}}\right)\left(\bar{X}_T^{(2-1)} - (1-\lambda)\bar{X}_A^{(2-1)} - \lambda\bar{X}_P^{(2-1)}\right) > c_2\sigma\sqrt{\frac{M}{n_T^{(2)}}} \mid Z_1 = z_1\right)$$

$$= P\left(\frac{n_T^{(1)}}{n_T^{(2)}} z_1\sigma\sqrt{\frac{M}{n_T^{(1)}}} + \frac{n_T^{(2)} - n_T^{(1)}}{n_T^{(2)}}\left(\bar{X}_T^{(2-1)} - (1-\lambda)\bar{X}_A^{(2-1)} - \lambda\bar{X}_P^{(2-1)}\right) > c_2\sigma\sqrt{\frac{M}{n_T^{(2)}}} \mid Z_1 = z_1\right)$$

$$= P\left(\bar{X}_T^{(2-1)} - (1-\lambda)\bar{X}_A^{(2-1)} - \lambda\bar{X}_P^{(2-1)} > \frac{c_2\sigma\sqrt{n_T^{(2)}M} - z_1\sigma\sqrt{n_T^{(1)}M}}{n_T^{(2)} - n_T^{(1)}}\right) \quad (3.1.6)$$

Let $\theta$=$\mu_T$-$(1-\lambda)\mu_A - \lambda\mu_P$, $E\left(\bar{X}_T^{(2-1)} - (1-\lambda)\bar{X}_A^{(2-1)} - \lambda\bar{X}_P^{(2-1)}\right) = \theta$

$$SD\left(\bar{X}_T^{(2-1)} - (1-\lambda)\bar{X}_A^{(2-1)} - \lambda\bar{X}_P^{(2-1)}\right) = \sigma\sqrt{\frac{M}{n_T^{(2)} - n_T^{(1)}}}$$

Let T=$\frac{\bar{X}_T^{(2-1)} - (1-\lambda)\bar{X}_A^{(2-1)} - \lambda\bar{X}_P^{(2-1)} - \theta}{\sigma\sqrt{\frac{M}{n_T^{(2)} - n_T^{(1)}}}}$ , we know $T \sim N(0,1)$

Therefore (3.1.6) becomes

$$P\left(T > \frac{c_2\sqrt{n_T^{(2)}} - z_1\sqrt{n_T^{(1)}}}{\sqrt{n_T^{(2)} - n_T^{(1)}}} - \frac{\theta}{\sigma\sqrt{\frac{M}{n_T^{(2)} - n_T^{(1)}}}}\right)$$

$$= 1 - \Phi\left(\frac{c_2\sqrt{n_T^{(2)}} - z_1\sqrt{n_T^{(1)}}}{\sqrt{n_T^{(2)} - n_T^{(1)}}} - \frac{\theta}{\sigma\sqrt{\frac{M}{n_T^{(2)} - n_T^{(1)}}}}\right)$$

*3.1.2.2 Promising Zone Determination*

Similar to two-arm two-stage studies discussed in section 2.3, it is found that there exists a promising zone for the conditional power for the three-arm effect preservation test, within which one can increase the sample size without worrying about type I error inflation.

The conditional power formula (3.1.5) is a function of $z_1$, and thus for any observed value of $z_1$, the conditional power for the interim data can be calculated. If it is lower than the nominal power, we can then equate this formula to the nominal power (or any desired value) to solve for a recalculated sample size $\tilde{n}_T^{(2)}$. We could also find an adjusted $\tilde{c}_2$ in pair with $\tilde{n}_T^{(2)}$ so that the conditional error function of $CP_{\theta=0}(\tilde{n}_2, \tilde{c_2}|z_1) = CP_{\theta=0}(n_2, c_2|z_1)$. Next, by plotting the adjusted critical value $\tilde{c}_2$ versus conditional power calculated at $n_2$, we could find that within a certain range of conditional power, the adjusted $\tilde{c}_2$ is lower than the initial $c_2$. It turned out that the promising zone boundaries depend on the nominal power, the timing of the interim analysis ($n_1/n_2$) and the cap for the recalculated sample size ($n_{max}$). The same findings were reported in the work by Mehta and Pocock[24] for two-arm two-stage studies.

In a hypothetical two-stage study design powered at 80%, with the interim stage being performed at 50% of the initial sample size, and the maximum recalculated sample size being four times the initial sample size, the promising zone ranges approximately from 31% to 80% (Figure 3).

**Figure 3 Plot of adjusted final stage rejection criterion versus conditional power for the three-arm non-inferiority test**

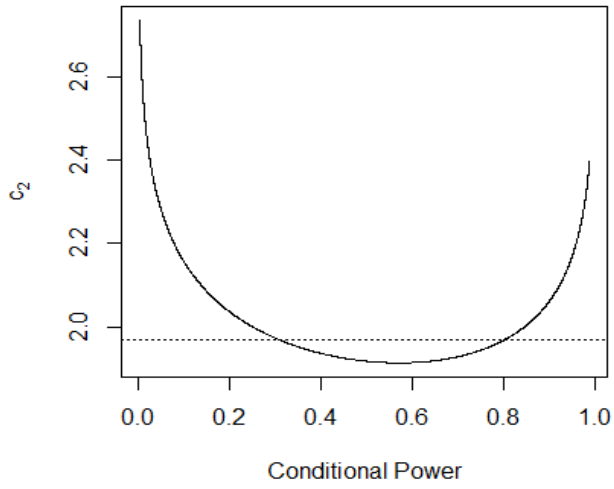



Table 4 summarizes the lower bound of promising zone at different combinations of analysis timing, $n_{max}$ while the nominal power is 80%. The lower bound decreases with increased $n_{max}$ and later interim analysis.

**Table 4. The threshold of promising zones for the three-arm non-inferiority test in the two-stage SSR designs**

| **$n_{max}$ (folds of original n)** | **Fraction of sample size at interim look ($n^{(1)}/n^{(2)}$)** | **Lower bound of promising zone** |
|---|---|---|
| 2 | 0.5 | 36% |
| 2 | 0.75 | 33% |
| 4 | 0.5 | 31% |
| 4 | 0.75 | 30% |
| ∞ | 0.5 | 31% |
| ∞ | 0.75 | 30% |

Note: The lower bounds were calculated for nominal initial power of 80%.

*3.1.2.3 Type I Error Control*

At each stage, we will first perform the assay sensitivity test and will not test the non-inferiority unless we reject the null hypothesis, the type I error is controlled stage-wise. We also showed that for the effect preservation test, the type I error is preserved if the sample size increase happens only when the interim result is promising. This way the entire procedure preserves the family-wise type I error. Simulations were run to look at the probability of null hypotheses rejection in the final stage when the parameters meet null hypotheses condition. Given the nominal type-I error of 2.5% we consider several combinations of $\lambda$ and allocations. For each scenario simulations with 100,000 replicates were performed. The simulated type-I error and power are the proportion of the simulations that cross the boundaries at the final stage. Table 5 below shows that the overall type I error was well preserved at all the investigated scenarios.

**Table 5 Simulation results on the type-I error rate (%) for the proposed SSR design based on the conditional power**

| | $\lambda$ | | | | |
|---|---|---|---|---|---|
| Allocations | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| 1:1:1 | 2.38 | 2.47 | 2.47 | 2.41 | 2.39 |
| 1:(1-λ): λ | 2.53 | 2.39 | 2.47 | 2.40 | 2.39 |
| 2:1:1 | 2.30 | 2.40 | 2.41 | 2.48 | 2.40 |
| 2:2:1 | 2.41 | 2.51 | 2.48 | 2.41 | 2.41 |
| 4:4:1 | 2.39 | 2.42 | 2.39 | 2.25 | 2.22 |

Note: Simulations performed conditioned on $\Delta_{TP} = (1-\lambda)\Delta_{AP}$; nominal $\alpha$=2.5%, 1-β=80%, number of simulations=100,000.

*3.1.2.4 Operational Characteristics by Monte-Carlo Simulations*

In this section, the actual power and average sample size are investigated via Monte-Carlo simulations. First of all, we look into the power property and efficiency of method 2. Then we compare the operational characteristics of this testing methods to method 1 described in chapter 3.1.1.

In many cases the placebo was introduced into such studies because the assay sensitivity is not constant and hard to find a good estimate. Therefore, we first looked at how this SSR design adapt to such uncertainties by looking into the rejection probabilities at each stage at the scenarios when the assay sensitivity is close to true value, overestimated or underestimated. The table 6 used a few examples to show how the proposed SSR design with interim look at 50% information fraction work using simulations. Assuming $\mu_T = \mu_A$, $\lambda = 0.3$, with an overall type I error rate of 0.025 and a nominal power of 80%, when the active comparator assumption at the trial design stage is close to true, the expected probability of claiming study success is 84%. There is 16% chance of rejecting null hypothesis at the interim stage. The probability for the conditional power falling into promising zone is 23%, in which case the sample size will be recalculated. The expected sample size is 6% higher than the sample size needed for a fix sample design. When the active comparator effect was underestimated by 10%, the chance of rejection at the interim stage increase and the chance of conditional power falling into promising and favorable zone is higher. On the contrary, if the assay sensitivity was overestimated by a factor of $\sqrt{2}$, in which case the initial sample size is only half

of the needed sample size, the actual power without sample size recalculation was 51% only. The method 2 SSR increase the actual power by 6% which is moderate compared to a substantial increase of more than 20% in method 1. The reason for the moderate increase is that the method 2 increase the sample size only when the interim result is promising. There is a relatively high chance that the conditional power goes unfavorable for which there would not be interest in investing more. However if the interim result happen to be in the promising zone there is a considerable power gain to 77% after sample size recalculation.

**Table 6 Simulation results on the operational characteristics for the proposed SSR design (method 2) with uncertainties in the assay sensitivity**

| | Actual Power[1] (%) | Probability of null Rejection at the interim (%) | Average sample size of SSR/fixed sample size | Conditional Power Zone | Probability falls into each CP zone (%) | Probability of null Rejection in the zone (%) |
|---|---|---|---|---|---|---|
| $\Delta_{AP,True} = \Delta_{AP,Assumed}$ | 84 (80) | 16 | 1.06 | Unfavorable | 20 | 43 |
| | | | | Promising | 23 | 93 |
| | | | | Favorable | 57 | 95 |
| $\Delta_{AP,True} = 1.1\Delta_{AP,Assumed}$ | 90 (87) | 22 | 1.22 | Unfavorable | 15 | 51 |
| | | | | Promising | 21 | 95 |
| | | | | Favorable | 64 | 97 |
| $\Delta_{AP,True} = \Delta_{AP,Assumed}/\sqrt{2}$ | 57 (51) | 6 | 1.15 | Unfavorable | 40 | 20 |
| | | | | Promising | 26 | 77 |
| | | | | Favorable | 34 | 84 |

1. The numbers are actual power of the SSR design*100 (actual power without SSR).
Note: The numbers in parenthesis are the actual power without sample size recalculation; nominal $\alpha$=2.5%, allocation is 1:1:1, $\lambda = 0.3$, number of simulations=100,000**.**

In therapeutic areas with moderate to big response variations, besides the uncertainties of active comparator effect, it is not surprising that the experimental treatment preserves slightly less or more than 100% of active comparator effect.

However, in the previous work on the three-arm non-inferiority trials, the sample size calculation was always being determined assuming equivalent effect between treatment and active comparator, in such case the actual power will also deviate from the nominal power. Here we are interested in seeing how this method 2 perform when the true experimental treatment effect differs from the active comparator effect.

In table 7, the actual power was computed for each scenario by simulations. The actual power for the fixed sample design was presented in the parenthesis for comparison. In general, when the preserved effect is less than 100% but is over-estimated at the design stage, the initial sample size does not adequately power the study, the method 2 SSR increases actual study rejection probability by 4% to 6%. This magnitude agrees with the observation by Mehta etc[18] when they compared the group sequential design against the design with SSR at promising interim results for two-arm studies. When the experimental treatment is actually superior than the active comparator, the study using the initial sample size will have excessive power. In such case for the method 2 SSR design, in the interim stage there is a good chance of rejecting nulls and claim study success, and also a relatively high probability that the conditional power falls into favorable zone. Table 8 showed average sample size (ASN) and the chance of rejecting the null hypotheses in the interim stage. When the assumptions are close to the true values, there is around 7% increase in the averaged sample size associated with 4% increase of rejection probability compared

to the nominal power. When treatment effect is underestimated, the ASN can be smaller than the sample size for fixed design due to good chance (35% - 49%) to reject the null hypothesis at the interim stage. In all investigated scenarios, the optimal allocation is still more efficient than balanced allocation although with a slightly (~1%) lower chance of interim stage rejection. The sample size allocations investigated are the balanced allocation (1:1:1) and optimal allocation (1:(1- λ): λ). As shown in table 7 their impact on the actual power is negligible. The optimal allocation requires fewer total sample size than balanced allocation in the SSR design as well.

**Table 7 Simulation results on the power (%) for method 2 at different combinations of $\Delta_{TP,True}$, $\Delta_{AP,True}$, and $\Delta_{AP,Assumed}$**

| | | $\Delta_{AP,True} = \Delta_{AP,Assumed}$ | | | $\Delta_{AP,True} = 1.1\Delta_{AP,Assumed}$ | | | $\Delta_{AP,True} = \Delta_{AP,Assumed}/\sqrt{2}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\frac{\Delta_{TP,True}}{\Delta_{AP,True}}$ | Allocations | $\lambda$ | | | $\lambda$ | | | $\lambda$ | | |
| | | 0.2 | 0.4 | 0.5 | 0.2 | 0.4 | 0.5 | 0.2 | 0.4 | 0.5 |
| 1 | 1:1:1 | 84 (80) | 84 (80) | 84 (80) | 90 (87) | 90 (87) | 90 (87) | 57 (51) | 56 (51) | 55 (51) |
| | 1:(1- λ): λ | 84 (80) | 84 (80) | 84 (80) | 90 (87) | 90 (87) | 90 (87) | 57 (51) | 56 (51) | 55 (51) |
| 0.95 | 1:1:1 | 62 (56) | 75 (70) | 78 (73) | 70 (64) | 82 (77) | 84 (80) | 36 (32) | 47 (41) | 48 (44) |
| | 1:(1- λ): λ | 62 (56) | 75 (69) | 77 (73) | 70 (64) | 82 (78) | 84 (80) | 37 (32) | 46 (41) | 48 (44) |
| 1.1 | 1:1:1 | 89 (86) | 96 (94) | 94 (93) | >99 (98) | 98 (97) | 97 (96) | 88 (84) | 80 (75) | 69 (67) |
| | 1:(1- λ): λ | 89 (86) | 96 (94) | 94 (93) | >99 (98) | 98 (97) | 97 (96) | 88 (84) | 79 (75) | 69 (67) |

Note: The numbers in parenthesis are the actual power without sample size recalculation; nominal $\alpha$=2.5%, number of simulations=100,000.

**Table 8 Simulation results on the average sample size and the interim stage rejection probability for method 2**

| $\frac{\Delta_{TP,True}}{\Delta_{AP,True}}$ | Allocations | Sample size for fixed design | $\Delta_{AP,True} = \Delta_{AP,Assumed}$ | $\Delta_{AP,True} = 1.1\Delta_{AP,Assumed}$ | $\Delta_{AP,True} = \Delta_{AP,Assumed}/\sqrt{2}$ |
|---|---|---|---|---|---|
| 1 | 1:1:1 | 1149 | 1229 (16%) | 1187 (22%) | 1354 (5%) |

| | | | | | |
|---|---|---|---|---|---|
| | 1:(1- λ): λ | 970 | 1048 (15%) | 1005 (21%) | 1148 (3%) |
| 0.95 | 1:1:1 | 1149 | 1312 (10%) | 1278 (13%) | 1359 (3%) |
| | 1:(1- λ): λ | 970 | 1112 (9%) | 1082 (12%) | 1148 (2%) |
| 1.1 | 1:1:1 | 1149 | 1032 (38%) | 940 (49%) | 1288 (11%) |
| | 1:(1- λ): λ | 970 | 886 (35%) | 802 (46%) | 1100 (7%) |

Note: λ=0.3, $\Delta_{\text{AP,Assumed}} = 0.6$, nominal $\alpha$=2.5%, number of simulations=100,000.

### 3.1.3 Comparison Between Method 1 and Method 2

In this thesis, first we extended the application scope of the established method 1 by proving that such sample size recalculation will not inflate type I error when the allocation met the condition: $\frac{c_P}{c_A} \geq \frac{\lambda}{(1-\lambda)}$ . When $\lambda$ is no bigger than 0.5, the balanced allocation always meets this condition. This proof adds a great value to this SSR design as balanced allocation is used mostly frequently in real world.

Furthermore, we compare the operational characteristics of the proposed method 2, conditional power based SSR with the method 1. The table 9 listed out which method preserves the type I error within the prespecified 2.5% significance level based on the simulation results at different combinations of λ and sample size allocations. There is no inflation in the type I error for both SSR methods at the optimal allocation (1:(1- λ): λ), the fully balanced allocation (1:1:1) and at 2:1:1 allocation for all λ investigated. Inflations are observed for Method 1 when the placebo to active comparator sample size ratio is less than the cutoff value.

**Table 9 Comparison of method 1 and method 2 at overall type I error control**

| | | $n_T$:$n_A$:$n_P$ | | | |
|---|---|---|---|---|---|
| λ | 1:(1- λ): λ | 1:1:1 | 2:1:1 | 2:2:1 | 4:4:1 |

| | | | | | |
|---|---|---|---|---|---|
| 0.1 | Both | Both | Both | Both | Both |
| 0.2 | Both | Both | Both | Both | Both |
| 0.3 | Both | Both | Both | Both | Method 2 |
| 0.4 | Both | Both | Both | Method 2 | Method 2 |
| 0.5 | Both | Both | Both | Method 2 | Method 2 |

Note: BOTH = The overall type I error was both controlled at 2.5% level; Method 2= The overall type I error was controlled using method 2 and inflation was observed using method 1.
Simulations performed conditioned on $\Delta_{TP}=(1-\lambda)\Delta_{AP}$; nominal $\alpha$=2.5%, number of simulations=100,000.

The power comparisons were then performed at balanced allocations for various combinations of the over-estimated or under-estimated parameters. The simulation results are shown in Table 10. It is found that whenever the assay sensitivity $\Delta_{AP}$ is over-estimated, sample size recalculation based on the observed $\hat{\Delta}_{AP}^{(1)}$ can make up the majority power loss due to the inaccurate estimation of $\Delta_{AP}$. This agrees with the conclusion from the work by Schwartz and Denne[8] where the method 1 is originally proposed. In contrast, the method 2 increase the power moderately, ranges from 4 to 7% for over-estimated parameters. However when the sample size increase happens due to promising interim result for method 2 the power gain is substantial. Besides the sample size allocation restriction, another limitation of method 1 is that it cannot handle the uncertainties from the experimental treatment effect. An underlying assumption for the method 1 to work properly is that the treatment effect equals the active comparator effect. In the pragmatic scenarios where there is deviation from the assumption, one needs re-evaluate its appropriateness. For instance, when $\Delta_{AP,Assumed}$ is accurate at the design stage, $\Delta_{TP}$ preserves 95% of $\Delta_{AP}$, and the non-inferiority threshold is 70% (i.e., $\lambda = 0.3$), with a nominal power of

80%, using method 1 the probability of rejecting the null hypotheses at the final stage is 64%, no change compared to fixed sample design. It is not surprising considering that the sample size recalculation of method 1 is entirely dependent on the estimate of $\Delta_{AP}$ using the interim data. In contrast, method 2 recalculates the sample size to make up the power loss caused by the over-estimation of the preserved fraction of the treatment effect. In a certain scenario, where the experimental treatment effect is better than the active comparator, and the active comparator is under-estimated, method 1 provides an actual power even lower than the fixed sample design. This is because it allows for sample size reduction. Such scenario is not uncommon in the real clinical trials as the clinical conditions improves over time. One way to avoid such issue in method 1 is to not allow reducing sample size and that is probably what regulatory agencies recommend. In terms of efficiency method 2 outperforms method 1 in most cases as it often needs smaller average sample size. This observation is predictable because in method 2 there is chance to stop the study for excessive efficacy at the interim stage.

**Table 10 Comparisons of actual power and average sample size (in the parenthesis) between method 1 and method 2**

| | | Actual Power (Average Sample Size) | | |
|---|---|---|---|---|
| $\Delta_{TP,True}/\Delta_{AP,True}$ | $\Delta_{AP,True}/\Delta_{AP,Assumed}$ | No SSR | Method I | Method II |
| 1 | 1/√2 | 51 (383) | 78 (854) | 57 (445) |
| | 1 | 80 (383) | 77 (421) | 84 (414) |
| | 1.1 | 86 (383) | 78 (344) | 89 (396) |
| 0.95 | 1/√2 | 38 (383) | 62 (854) | 44 (452) |
| | 1 | 65 (383) | 64 (421) | 71 (438) |
| | 1.1 | 73 (383) | 64 (344) | 78 (428) |
| 1.1 | 1/√2 | 75 (383) | 91 (854) | 80 (428) |
| | 1 | 96 (383) | 93 (421) | 97 (346) |

| | | | | |
|---|---|---|---|---|
| | 1.1 | 98 (383) | 93 (344) | 99 (315) |

**Note: $\lambda = 0.3$, $\Delta_{AP,Assumed} = \Delta_{TP,Assumed} = 0.60$, nominal $\alpha$=2.5%, nominal power = 80%, number of simulations=100,000.**

## 3.2 Effect Preservation Test, Binary Endpoint

In many clinical trials the primary endpoint can be the risk of a certain event, such as mortality, stroke and etc. In such studies the primary outcome is collected as a binary variable. The underlying distribution for such endpoint $X_i$ is Binominal $(N_i, \pi_i)$, $i = T, A, P$, where $X_i$ is the number of success in the ith arm, $N_i$ is the number of subjects in the ith arm, and $\pi_i$ is the success rate of each arm. The testing procedures designed for normal endpoints may no longer be appropriate. The most common parametric model for binary outcome is assuming binomial distributions, in which the variance is linked to the mean, as a result, the optimal allocation will no longer take the simple form as in cases with normal endpoint. Another challenge is that there is no analytical solution for the variance estimate of the test statistic. The restricted maximum likelihood estimate has to be calculated using the Newton-Raphson algorithm. These challenges will be discussed in detail in this section, and the performance of the conditional power-based sample size recalculation for such studies will be investigated through Monte Carlo simulations.

### 3.2.1 Hypothesis, Test Statistics and Sample Size

Let $\pi_T$, $\pi_A$, $\pi_P$ denote the success rate of the binary outcome of treatment, active comparator and placebo respectively.

The hypotheses testing procedure will include

$H_{0,AP}^{(s)}$: $\pi_A \leq \pi_P$ versus $H_{1,AP}^{(s)}$: $\pi_A > \pi_P$

$H_{0,TA}^{(n)}$: $\frac{\pi_T - \pi_P}{\pi_A - \pi_P} \leq 1 - \lambda$ versus $H_{1,TA}^{(n)}$: $\frac{\pi_T - \pi_P}{\pi_A - \pi_P} > 1 - \lambda$

Again, due to the natural hierarchy, the fixed sequence testing procedure is the most appropriate approach for the type I error control.

The construction and estimation of test statistics for the binary outcome is similar but more complicated than normally distributed outcome.

Let $\theta = \pi_T - (1 - \lambda)\pi_A - \lambda\pi_p$ , the estimate $\hat{\theta} = P_T - (1 - \lambda)P_A - \lambda P_P$, where $P_i = \frac{X_i}{N_i}$, i=T, A, P

The variance of $\hat{\theta}$

$$Var(\hat{\theta}) = \left\{\pi_T(1 - \pi_T) + \frac{(1-\lambda)^2\pi_A(1-\pi_A)}{c_A} + \frac{\lambda^2\pi_P(1-\pi_P)}{c_P}\right\}\frac{1}{n_T}$$

can have various estimates depending on the estimated ($\pi_T, \pi_A, \pi_P$ ) to be plugged in. Plugging in the maximum likelihood estimates ($P_T, P_A, P_P$ ) yields Wald's statistic. We could also substitute $\pi_i's$ by the restricted maximum likelihood estimate (RMLE), for which the estimates are made under null hypothesis, i.e., $\hat{\pi}_T - (1 - \lambda)\hat{\pi}_A - \lambda\hat{\pi}_p = 0$. The RMLE is identical to the score statistic and is equivalent to the test proposed by Miettinen and Nurminen[27] and Farrington and Manning[28] for two-arm non-inferiority trials. In Kieser and Friede's work (2007)[29] it is found that the RMLE based test statistic is more conservative and is much better at controlling the type I error compared to the Wald-type test statistics.

**Sample size for the non-inferiority test**

Let $\tau_0^2$, $\tau_1^2$ denote $n_T Var(\hat{\theta})$ under null and alternative hypothesis respectively, the power function for the non-inferiority test can then be derived as

$$1-\beta = P\left(\frac{\sqrt{n_T}\hat{\theta}}{\tau_0} > z_{1-\alpha}\middle| H_1\right) = 1-\Phi(\frac{z_{1-\alpha}\tau_0}{\tau_1} - \frac{\sqrt{n_T}\theta_1}{\tau_1})$$

The sample size for the non-inferiority test is calculated as

$$N_{01} = n_T(1+c_A+c_P) = \frac{(z_{1-\alpha}\tau_0 + z_{1-\beta}\tau_1)^2(1+c_A+c_P)}{\theta_1^2}$$

It may be simplified by substituting $\tau_0$ with $\tau_1$ or vise verse to either

$$N_{00} = \frac{(z_{1-\alpha}+z_{1-\beta})^2\tau_0^2(1+c_A+c_P)}{\theta_1^2}$$

or

$$N_{11} = \frac{(z_{1-\alpha}+z_{1-\beta})^2\tau_1^2(1+c_A+c_P)}{\theta_1^2}$$

Simulations were conducted to compare the accuracy and efficiency of the three formula of sample size calculations. We considered a nominal level α=2.5%, desired power=80%, balanced allocation at various combinations of preservation margin ($\lambda$) and success rates ($\pi_T, \pi_A, \pi_p$) . Table 10 shows the calculated sample sizes and the actual power from simulations.

**Table 11 Comparisons of sample sizes and actual power based on different variance values for the test statistics of the preservation non-inferiority test on binary endpoint**

| | | | | $N_{01}$ | | $N_{00}$ | | $N_{11}$ | |
|---|---|---|---|---|---|---|---|---|---|
| λ | $\pi_T$ | $\pi_A$ | $\pi_p$ | Sample size | Actual power | Sample size | Actual power | Sample size | Actual power |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 0.9 | 0.9 | 0.1 | 750 | 0.854 | 816 | 0.883 | 603 | 0.768 |
| 0.1 | 0.9 | 0.9 | 0.3 | 1281 | 0.827 | 1371 | 0.855 | 1080 | 0.747 |
| 0.1 | 0.6 | 0.6 | 0.1 | 4152 | 0.807 | 4173 | 0.811 | 4101 | 0.796 |
| 0.3 | 0.9 | 0.9 | 0.1 | 96 | 0.839 | 114 | 0.903 | 60 | 0.564 |
| 0.3 | 0.9 | 0.9 | 0.3 | 168 | 0.858 | 195 | 0.906 | 114 | 0.661 |
| 0.3 | 0.6 | 0.6 | 0.3 | 1116 | 0.795 | 1125 | 0.806 | 1095 | 0.790 |
| 0.5 | 0.9 | 0.9 | 0.1 | 36 | 0.779 | 45 | 0.878 | 21 | 0.467 |
| 0.5 | 0.8 | 0.9 | 0.1 | 72 | 0.796 | 78 | 0.841 | 54 | 0.667 |
| 0.5 | 0.7 | 0.6 | 0.1 | 60 | 0.782 | 60 | 0.782 | 57 | 0.764 |

Note: Nominal $\alpha$=2.5%, nominal power=80%, number of simulations=10,000.

It can be seen that for all considered parameter combinations the sample size $N_{01}$ gives the power closest to the nominal power. The $N_{00}$ always gives a power bigger than the desired power and is sometimes not as efficient. It can go 20% above the needed sample size. $N_{11}$ showed bad performance when the sample size is small. In certain investigated cases the power for $N_{11}$ went as low as 47%. This finding agrees with the calculations done by Kieser and Friede[29].

**Sample size for the entire test procedure**

In chapter 1.4.2 it is proved that for normally distributed endpoint with common variance, at $\lambda \leq 0.5$, the sample size required for the non-inferiority test is always larger than the sensitivity test. For the binary endpoint due to the link between mean and variance we need re-visit the sample size comparison between the two tests to see if this conclusion is still true. Assuming an equal success rate of experimental treatment and active comparator, the sample size needed for the non-inferiority test and for the assay sensitivity test are respectively

$$n_T^{NIF} = \frac{\left\{(1+\frac{(1-\lambda)^2}{c_A})\pi_A(1-\pi_A)+\frac{\lambda^2\pi_P(1-\pi_P)}{c_P}\right\}\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{\lambda^2{\Delta_{AP}}^2}$$

$$n_T^{AP} = \frac{\left\{\frac{\pi_A(1-\pi_A)}{c_A} + \frac{\pi_P(1-\pi_P)}{c_P}\right\}(z_{1-\alpha} + z_{1-\beta})^2}{\Delta_{AP}{}^2}$$

By taking the ratio between the $n_T^{NIF}$ and $n_T^{AP}$:

$$r = \frac{\left\{\left(1 + \frac{(1-\lambda)^2}{c_A}\right)\pi_A(1-\pi_A) + \frac{\lambda^2\pi_P(1-\pi_P)}{c_P}\right\}}{\lambda^2\left\{\frac{\pi_A(1-\pi_A)}{c_A} + \frac{\pi_P(1-\pi_P)}{c_P}\right\}}$$

$$= \frac{\left(\frac{c_A - 2\lambda + 1}{c_A}\right)\pi_A(1-\pi_A) + \lambda^2\left\{\frac{\pi_A(1-\pi_A)}{c_A} + \frac{\pi_P(1-\pi_P)}{c_P}\right\}}{\lambda^2\left\{\frac{\pi_A(1-\pi_A)}{c_A} + \frac{\pi_P(1-\pi_P)}{c_P}\right\}}$$

$$= \frac{\left(\frac{1 - 2\lambda + c_A}{c_A}\right)\pi_A(1-\pi_A)}{\lambda^2\left\{\frac{\pi_A(1-\pi_A)}{c_A} + \frac{\pi_P(1-\pi_P)}{c_P}\right\}} + 1$$

By observing the above formula, whether r can go below 1 depends on the sign of $1 - 2\lambda + c_A$. Again we are only interested in the scenario with $\lambda \leq 0.5$, in which the first term in the above is positive, therefore the ratio is always greater than 1, suggesting that the sample size required to achieve a certain power is dominated by the non-inferiority test.

### 3.2.2 SSR Methods for Binary outcome

Let $p_T$, $p_A$, $p_P$ be the observed proportion (success rate) of the experimental treatment, active comparator and placebo group $\pi_{T,i}$, $\pi_{A,i}$, $\pi_{P,i}$ be the success rate at the null hypothesis (i=0) or alternative hypothesis (i=1); $\widehat{\theta}^{(k)} = p_T^{(k)} - (1-\lambda)p_A^{(k)} -$

$\lambda \mathrm{p}_P^{(\mathrm{k})}$ (k=1, 2-1, 2) . The superscript (1), (2-1) and (2) respectively indicate the data at the interim stage, post interim and throughout the trial.

The method 1 uses the $N_{01}$ for the initial sample size calculation and replace $\theta_1 = \lambda(\pi_{A,1} - \pi_{P,1})$ with $\lambda(\mathrm{p}_A^{(1)} - \mathrm{p}_P^{(1)})$ for the sample size recalculation. The test procedure stays the same as for the continuous outcome.

*3.2.2.1 Conditional Power Derivation for Binary Outcome*

The method 2 still depends on the conditional power. The conditional power for the non-inferiority test can be derived as follows.

The test statistic for the interim non-inferiority test is denoted as

$\mathrm{z}_1^{NIF} = \frac{\hat{\theta}^{(1)}}{\sqrt{\frac{\hat{\tau}^{(1)^2}}{n_T^{(1)}}}}$, then we have

$$CP(NIF) = P\left( \frac{\hat{\theta}^{(2)}}{\sqrt{\frac{{\tau_0}^2}{n_T^{(2)}}}} > c_2 | Z_1 = \mathrm{z}_1^{NIF} \right)$$

$$= P\left( \frac{n_T^{(1)}}{n_T^{(2)}} \hat{\theta}^{(1)} + \left(1 - \frac{n_T^{(1)}}{n_T^{(2)}}\right) \hat{\theta}^{(2-1)} > c_2 \sqrt{\frac{{\tau_0}^2}{n_T^{(2)}}} \mid Z_1 = \mathrm{z}_1^{NIF} \right)$$

$$= P\left( \frac{n_T^{(1)}}{n_T^{(2)}} \mathrm{z}_1^{NIF} \sqrt{\frac{\hat{\tau}^{(1)^2}}{n_T^{(1)}}} + \frac{n_T^{(2)} - n_T^{(1)}}{n_T^{(2)}} \left(\hat{\theta}^{(2-1)}\right) > c_2 \sqrt{\frac{{\tau_0}^2}{n_T^{(2)}}} \mid Z_1 = \mathrm{z}_1^{NIF} \right)$$

$$= P\left( \hat{\theta}^{(2-1)} > \frac{c_2 \tau_0 \sqrt{n_T^{(2)}} - \mathrm{z}_1^{NIF} \hat{\tau}^{(1)} \sqrt{n_T^{(1)}}}{n_T^{(2)} - n_T^{(1)}} \mid Z_1 = \mathrm{z}_1^{NIF}, H_1 \right)$$

$$= P\left(\frac{\hat{\theta}^{(2-1)} - \theta}{\hat{\tau}^{(1)}/\sqrt{n_T^{(2)} - n_T^{(1)}}} > \frac{c_2\tau_0\sqrt{n_T^{(2)}} - z_1^{NIF}\hat{\tau}^{(1)}\sqrt{n_T^{(1)}}}{\hat{\tau}^{(1)}\sqrt{n_T^{(2)} - n_T^{(1)}}} - \frac{\theta}{\hat{\tau}^{(1)}/\sqrt{n_T^{(2)} - n_T^{(1)}}}\right)$$

$$= 1 - \Phi\left(\frac{c_2\tau_0\sqrt{n_T^{(2)}} - z_1^{NIF}\hat{\tau}^{(1)}\sqrt{n_T^{(1)}}}{\hat{\tau}^{(1)}\sqrt{n_T^{(2)} - n_T^{(1)}}} - \frac{\theta\sqrt{n_T^{(2)} - n_T^{(1)}}}{\hat{\tau}^{(1)}}\right)$$

*3.2.2.2 Testing Procedure and Operational Characteristics by Monte-Carlo Simulations*

The performance of the SSR methods for binary outcome is investigated via Monte-Carlo simulations in this chapter. The familywise type-I error, actual power property and efficiency compared to non-adaptive design are examined at various parameter combinations.

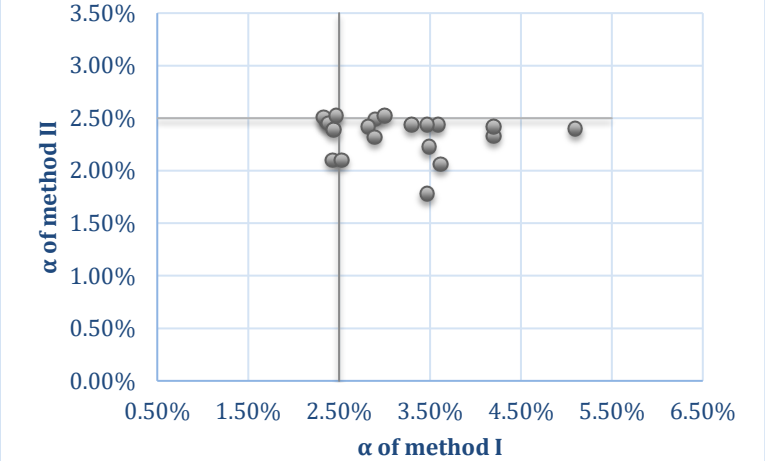
Given the nominal type-I error of 2.5%, study power of 80%, we consider several combinations of $\lambda$ and allocations. For each scenario simulations with 100,000 replicates were performed. The simulated type-I error and power are the proportion of the simulations that cross the boundaries at the final stage.

**Type-I error control**

Due to the nature of the testing procedure, the family-wise type I error of the entire procedure is dominated by the non-inferiority test. The investigated scenarios include the combination of λ=0.05, 0.2, 0.35, 0.5 and commonly used allocations of $n_T$: $n_A$: $n_P$=1:1:1, 2:2:1, 2:1:1, 4:1:1, 1: (1- λ): λ. It is observed from the figure 4 that most of the time the type I error is well preserved for method 2. For method 1, the type I error inflation is observed more frequently and to a larger extent, up to 5%,

even at the sample size allocations with good type I error control for continuous outcomes. It could be due to the non-independence between the mean and variance for the binary outcomes changes the relationship between the final test statistic and the interim $\widehat{\boldsymbol{\Delta}}_{\boldsymbol{AP}}^{(\mathbf{1})}$.

**Figure 4 Simulation results on the type-I error rates (%) for method 1 and method 2 for binary outcomes**




Note: The investigated scenarios include the combination of λ=0.05, 0.2, 0.35, 0.5, and $n_T$: $n_A$: $n_P$=1:1:1, 2:2:1, 2:1:1, 4:1:1, 1: (1- λ): λ

## Power and Efficiency

Simulations were also performed to investigate how the actual power and average sample size for the SSR methods, and how it was compared against fixed sample design while the true success rates deviate from the assumed values. Table 12 listed four representative scenarios. While the true success rates were the same as assumed rate, the method 1 yields almost identical power and average sample size to that of the fixed sample design. Method 2 yields slightly higher probability of final

null hypotheses rejections than the nominal power, with the price of 5% increase of average sample size. The chance of recalculating the sample size is 23%. In the second scenario the actual preserved effect is still 100% but both experimental treatment and active comparator success rates are overestimated by 5% at the design stage. In such case, the actual power for the fixed sample size design is 63%. The method 1 improve the actual power substantially to 77%. Method 2 design increased the actual power to 70%. There is 31% of chance the conditional power would fall into the unfavorable zone. The rejection probability in the promising or favorable zone are respectively 86% and 89%. In the third hypothetical scenario, the success rate in the experimental treatment is overestimated by 1% in the design stage, causing 6% of loss of actual power. The actual power using method 1 design equals the nominal power. It is because the active comparator success rate is accurate and there is very likely the sample size is not adapted. Method 2 increases the average rejection probability to 71%. The last scenario investigates the statistical properties when the assumption of experimental treatment success rate is accurate but the assay sensitivity is overestimated. In such case, the actual preserved fraction is greater than 100%. The method 1 design yields an over-powered study with efficiency even lower than the fixed sample design. In contrast the method 2 design has a 64% chance to stop the study for excessive efficacy, therefore saves the average sample size by 28%.

**Table 12 Simulation results for rejection probability at each conditional power zone for the method 2 applied on binary endpoint**

| True values | | | No SSR | Method 1 | Method 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\pi_T$ | $\pi_A$ | $\varepsilon$ | Actual Power x 100% (N) | Actual Power x 100% (N) | Actual Power x 100% (N) | Interim rej. prob. | CP Zone | Prob. Falls into each Zone | Rej. Prob. Within each Zone |
| 0.6 | 0.6 | 1 | 80 (2670) | 79 (2765) | 84 (2814) | 16.1% | Unfav. | 20% | 43% |
| | | | | | | | Promising | 23% | 92% |
| | | | | | | | Fav. | 57% | 95% |
| 0.55 | 0.55 | 1 | 64 (2670) | 77 (4044) | 70 (3042) | 9% | Unfav. | 31% | 28% |
| | | | | | | | Promising | 27% | 86% |
| | | | | | | | Fav. | 42% | 89% |
| 0.59 | 0.6 | 0.97 | 65 (2670) | 64 (2765) | 71 (3036) | 10% | Unfav. | 30% | 28% |
| | | | | | | | Promising | 26% | 87% |
| | | | | | | | Fav. | 44% | 89% |
| 0.6 | 0.55 | 1.2 | 96 (2670) | 99 (4047) | 99 (1908) | 64% | Unfav. | 2% | 88% |
| | | | | | | | Promising | 5% | >99% |
| | | | | | | | Fav. | 93% | >99% |

Note: The numbers in parenthesis are the actual power without sample size recalculation; nominal $\alpha$=2.5%, 1-$\lambda$=0.8, Alternative hypothesis: $\pi_{P,a}$=0.3, $\pi_{A,a}$=0.6, $\pi_{T,a} = 0.6$; Nominal $\alpha$ =0.025 (one-sided), nominal power for Non-inferiority test = 80%, $\varepsilon = \frac{\pi_T - \pi_P}{\pi_A - \pi_P}$, balanced allocation, number of simulations=50,000.

## 3.3 Conclusion and Discussion

There are quite a few factors to be considered when designing three-arm non-inferiority trial. According to the findings in this thesis, if there is limited prior information to generate a clinically relevant margin, the effect preservation test is preferred over the fixed margin test because of higher efficiency to claim study success when the margin of $\lambda\Delta_{AP}$ is numerically identical to the fixed margin. The adaptive design with sample size recalculation is advocated for such trial because of the great uncertainties on many aspects, such as the experimental treatment

efficacy, the assay sensitivity and the margin. The sample size recalculation method 1 proposed by Schwartz and Denne does not require unblinding the treatment efficacy result and is likely to be a practical choice whenever applicable because of its straightforward design and interpretation. In this thesis its application was extended to a broader range of sample size allocations than the originally proposed optimal allocation only. Nevertheless, its usage is still limited to two-stage design under certain sample size allocations which does not lead to type I error inflation. In addition, it is not under the group sequential design framework and thus does not allow for the study to stop in the interim stage for over-efficacy or futility. The method 2 introduced to the three-arm non-inferiority setting in this thesis calculates the conditional power for the non-inferiority test at the interim stage. If the conditional power falls into the promising zone the sample size may be re-calculated and the final test can proceed without adjustment. When the assay sensitivity was under-estimated at the design stage or when the assumed experimental treatment effect deviates from the true value, this conditional power approach could either increase the trial efficiency by allowing the study to stop early. The stop criteria would have to be specified at the planning stage and be agreed upon by the study clinician and regulatory reviewer. We also looked into the statistical properties of this method in studies with binary outcomes. The restricted maximum likelihood estimate was used to construct the test statistics. Based on the simulations, it is found that the type I error was preserved in the investigated scenarios. The overall study power increases by 4 to 7% compared to the fixed

sample design. When the efficacy was under-estimated at the design stage, the average sample size can reduce dramatically compared to fixed sample design due to the chances of rejection at the interim stage.

In this work we have assumed the continuous outcomes follow normal distributions with known common variance. In practice the sample variance could be used for the sample size and conditional power calculations. This will introduce deviation from normality and additional variability. Further investigations may be conducted to evaluate the impact of this. In real studies it is often not recommended to reduce the sample size for an over-powered study. Therefore when applying the method 1 in real studies the efficiency may be lower than what we see in this thesis. The limitation of method 2 is the low chance of having the interim results in the promising zone at overly optimistic assumptions, in which case the sample size re-estimation is not triggered at all. One way to get around this may be a combination of method 1 and method 2, i.e., use the sample size allocations that will not inflate type I error for method 1, do the sample size adaptation if the conditional power falls into promising zone as well as below the promising zone but not unacceptably low.

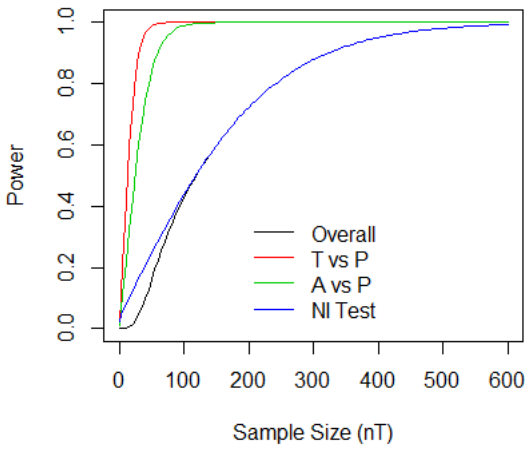
## APPENDIX I Overall Power for Three Tests

$T_{AP}^{(s)}, T_{TP}^{(s)}, T_{TA}^{(n)}$ follows Multivariate Normal distribution. Assume the design is to test $H_{0,AP}^{(s)}$ and $H_{0,TP}^{(s)}$ both at a significance level of $\alpha/2$, and proceed to the test of $H_{0,TA}^{(n)}$ at a significance level of $\alpha$ only when both superiority tests against placebo are successful. This way the overall type I error is well controlled at $\alpha$. Under the alternative hypothesis of $\mu_T = \mu_A$, the overall power function can be derived as

$$1-\beta = \Phi_{\boldsymbol{\Sigma}}\left(\frac{\mu_A-\mu_P}{\frac{\sigma}{\sqrt{n_T}}\sqrt{\frac{1}{c_A}+\frac{1}{c_P}}} - z_{1-\alpha/2},\ \frac{\mu_T-\mu_P}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{1}{c_P}}} - z_{1-\alpha/2},\ \frac{\lambda(\mu_A-\mu_P)}{\frac{\sigma}{\sqrt{n_T}}\sqrt{1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P}}} - z_{1-\alpha}\right)$$

Where $\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{1} & \frac{\frac{1}{c_P}}{\sqrt{(1+\frac{1}{c_P})(\frac{1}{c_A}+\frac{1}{c_P})}} & \frac{\frac{\lambda}{c_P}-\frac{1-\lambda}{c_A}}{\sqrt{(\frac{1}{c_A}+\frac{1}{c_P})(1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P})}} \\ \frac{\frac{1}{c_P}}{\sqrt{(1+\frac{1}{c_P})(\frac{1}{c_A}+\frac{1}{c_P})}} & \mathbf{1} & \frac{1+\frac{\lambda}{c_P}}{\sqrt{(1+\frac{1}{c_P})(1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P})}} \\ \frac{\frac{\lambda}{c_P}-\frac{1-\lambda}{c_A}}{\sqrt{(\frac{1}{c_A}+\frac{1}{c_P})(1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P})}} & \frac{1+\frac{\lambda}{c_P}}{\sqrt{(1+\frac{1}{c_P})(1+\frac{(1-\lambda)^2}{c_A}+\frac{\lambda^2}{c_P})}} & \mathbf{1} \end{bmatrix}$, $\Phi_{\boldsymbol{\Sigma}}$ is the CDF of bivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$.

Assuming the $\mu_T = \mu_A = 2, \mu_P = 0.5, \sigma = 1, \alpha = 0.05, \lambda = 0.3$, the plot of power functions (below) suggest that the overall power is determined by the non-inferiority test. Such trend is representative for all the pragmatic parameter combinations explored. Tuning the parameters with the $\lambda$ capped at 0.5, the point where the overall power curve and noninferiority test power curve converge may vary but always locates before the nominal power.


Power
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
600
Sample Size (nT)
Overall
T vs P
A vs P
NI Test

## APPENDIX II R code for the simulations

### Method 1, continuous outcome

```
##Input Parameters:
#alpha nominal significance level
#beta   nominal type II error (i.e., 1 – power)
# F  information fraction, i.e., the fraction of sample size at the interim stage
#muT, muA and muP true efficacy mean of the experimental treatment arm, active comparator arm and placebo arm respectively
#deltaAP.d assumed efficacy difference between active comparator and placebo
#sigma  common standard deviation of the three arms
#lambda fraction margin of the effect preservation test
#cA     the ratio of sample size in the active comparator arm over the experimental treatment arm
#cP     the ratio of sample size in the placebo arm over the experimental treatment arm
#Nsimu number of simulations

##Output
#pctRej percentage calculated as the number of rejections at the final stage among all the total number of simulated trials, i.e., the actual power by simulation
#avg_nT.r average number of recalculated sample size of the experimental treatment arm

Method1 <- function(alpha=0.025, beta=0.2, F, detAP.d, muT, muA, muP, sigma, lambda, cA, cP, Nsimu) {
 Za <- qnorm(1-alpha)
 Zb <- qnorm(1-beta)
# Initial sample size based on the deltaAP at the design stage #
 nT <- ceiling((Za+Zb)^2*sigma^2*(1+(1-lambda)^2/cA+lambda^2/cP)/
(lambda^2*detAP.d^2))
 nA <- ceiling(cA*nT)
 nP <- ceiling(cP*nT)

 nT.1 <- ceiling(nT*F)
 nA.1 <- ceiling(nA*F)
 nP.1 <- ceiling(nP*F)

 set.seed(123456)
 AllRej=0
 AllnT.r=NULL
 for (N in 1:Nsimu)
 {
  XT1 <- rnorm(n=nT.1, mean=muT, sd=sigma)
  XA1 <- rnorm(n=nA.1, mean=muA, sd=sigma)
  XP1 <- rnorm(n=nP.1, mean=muP, sd=sigma)

  detAP.1=mean(XA1)-mean(XP1);
```

```
   nT.r <- ceiling((Za+Zb)^2*sigma^2*(1+(1-lambda)^2/cA+lambda^2/cP)/
(lambda^2*detAP.1^2))
   nT.r=min(4*nT, nT.r)
   if (nT.r<=nT.1) (nT.r=nT.1+1)
   nA.r <- ceiling(cA*nT.r)
   nP.r <- ceiling(cP*nT.r)


   XT2 <- rnorm(n=nT.r-nT.1, mean=muT, sd=sigma)
   XA2 <- rnorm(n=nA.r-nA.1, mean=muA, sd=sigma)
   XP2 <- rnorm(n=nP.r-nP.1, mean=muP, sd=sigma)

   XTbar=mean(c(XT1, XT2))
   XAbar=mean(c(XA1, XA2))
   XPbar=mean(c(XP1, XP2))

   z.AP.f=(XAbar-XPbar)/sigma/sqrt(1/nA.r+1/nP.r)
   Rej.AP=z.AP.f>Za

   Z.TA.f=(XTbar-(1-lambda)*XAbar-lambda*XPbar)/sigma/sqrt(1/nT.r+(1-
lambda)^2/nA.r+lambda^2/nP.r)
   Rej.TA=(Z.AF.f>Za)
   Rej=min(Rej.AP, Rej.TA)

   AllRej=AllRej+min(Rej.AP, Rej.TA)
   AllnT.r=cbind(AllnT.r, nT.r)
 }
 pctRej=AllRej/Nsimu
 avg_nT.r=mean(AllnT.r)
 return(list(pctRej, avg_nT.r))
}
```

#Example 1: Simulate overall Type I error for $\lambda$=0.5 and $\mu_T = \mu_P + \lambda(\mu_A - \mu_P)$
Method1(detAP.d=1.5 F=0.5,, muT=1.25, muA=2, muP=0.5, sigma=1, lambda=0.5, cA=1, cP=1, , Nsimu=100000)

#Example 2: Simulate actual power for $\lambda$=0.1 and $\mu_T = \mu_A$
Method1(detAP.d=1.5, F=0.5, muT=2, muA=2, muP=0.5, sigma=1, lambda=0.1, cA=1, cP=1, , Nsimu=100000)

## Method 2, continuous outcome

##Input Parameters:
#alpha nominal significance level
#beta  nominal type II error (i.e., 1 – power)
# F  information fraction, i.e., the fraction of sample size at the interim stage
# CP.u upper bound of the promising zone of the conditional power

```
# CP.l lower bound of the promising zone of the conditional power
#muT, muA and muP true efficacy mean of the experimental treatment arm, active comparator arm and placebo arm respectively
#deltaAP.d assumed efficacy difference between active comparator and placebo
#sigma  common standard deviation of the three arms
#lambda fraction margin of the effect preservation test
#cA     the ratio of sample size in the active comparator arm over the experimental treatment arm
#cP      the ratio of sample size in the placebo arm over the experimental treatment arm
#Nsimu number of simulations

##Output
#pctRej percentage calculated as the number of rejections at the final stage among all the total number of simulated trials, i.e., the actual power by simulation
#avgss average number of recalculated sample size of the experimental treatment arm
#intrej the probability of rejection at the interim stage for excessive efficacy
#unfzone a list with two elements; the first element is the probability of falling into the unfavorable zone and the second element is the probability of rejection at the final stage in the unfavorable zone
#pprzone a list with two elements; the first element is the probability of falling into the promising zone and the second element is the probability of rejection at the final stage in the promising zone
#pfavzone a list with two elements; the first element is the probability of falling into the favorable zone and the second element is the probability of rejection at the final stage in the favorable zone

Method2 <- function(alpha, beta, F, CP.u, CP.l, detAP.d, muT, muA, muP, sigma, lambda, cA, cP, Nsimu) {
library(ldbounds)
Za <- qnorm(1-alpha)
Zb <- qnorm(1-beta)
obf.bd=bounds(t=c(0.5,1), iuse=1, alpha=alpha)
c1tilda=obf.bd$upper.bounds[1]
c2tilda=obf.bd$upper.bounds[2]

M=1+(1-lambda)^2/cA+lambda^2/cP

nT <- ceiling((c2tilda+Zb)^2*sigma^2*M/(lambda^2*detAP.d^2))
nA <- ceiling(cA*nT)
nP <- ceiling(cP*nT)


nT.1 <- ceiling(nT*F)
nA.1 <- ceiling(nA*F)
nP.1 <- ceiling(nP*F)

set.seed(123456)
```

```
AllRej=0
interim=0
zone=NULL
AllnT.r=NULL

for (N in 1:Nsimu)
{
XT1 <- rnorm(n=nT.1, mean=muT, sd=sigma)
XA1 <- rnorm(n=nA.1, mean=muA, sd=sigma)
XP1 <- rnorm(n=nP.1, mean=muP, sd=sigma)

detAP.1=mean(XA1)-mean(XP1);
z.AP.1=detAP.1/sigma/sqrt(1/nA.1+1/nP.1)
Rej.AP1=z.AP.1>c1tilda

theta1=mean(XT1)-(1-lambda)*mean(XA1)-lambda*mean(XP1)
Z.1=theta1/sigma/sqrt(M/nT.1)
Rej.TA1=Z.1>c1tilda

Rej1=min(Rej.AP1, Rej.TA1)

if (Rej1==1) {
nT.r=nT.1
AllRej=AllRej+Rej1
rejseq=cbind(rejseq, Rej1)
AllnT.r=cbind(AllnT.r, nT.1)
interim=interim+1
zone=cbind(zone, 1)

} else {

#Conditional power
CP=1-pnorm((c2tilda*sqrt(nT)-Z.1*sqrt(nT.1))/sqrt(nT-nT.1)-theta1/sigma/sqrt(M/(nT-
nT.1)))
if (CP<CP.l) {zone=cbind(zone, -1)}
if (CP>CP.u) {zone=cbind(zone, 1)}
if (CP>=CP.l & CP<=CP.u) {
zone=cbind(zone, 0)
CP0=CP
nT.r=nT
while(CP0<1-beta && nT.r<=4*nT){
 nT.r=nT.r+1
 CP0=1-pnorm((c2tilda*sqrt(nT.r)-Z.1*sqrt(nT.1))/sqrt(nT.r-nT.1)-
theta1/sigma/sqrt(M/(nT.r-nT.1)))
 }
} else {
#if not in promising zone, continue w/o changing sample size
```

```
nT.r <- nT
}

nA.r <- ceiling(cA*nT.r)
nP.r <- ceiling(cP*nT.r)

XT2 <- rnorm(n=nT.r-nT.1, mean=muT, sd=sigma)
XA2 <- rnorm(n=nA.r-nA.1, mean=muA, sd=sigma)
XP2 <- rnorm(n=nP.r-nP.1, mean=muP, sd=sigma)

XTbar=mean(c(XT1, XT2))
XAbar=mean(c(XA1, XA2))
XPbar=mean(c(XP1, XP2))

if (Rej.AP1==1) { Rej.AP=1 } else {
z.AP.f=(XAbar-XPbar)/sigma/sqrt(1/nA.r+1/nP.r)
Rej.AP=z.AP.f>c2tilda }

Z.f=(XTbar-(1-lambda)*XAbar-lambda*XPbar)/sigma/sqrt(1/nT.r+(1-
lambda)^2/nA.r+lambda^2/nP.r)
Rej=Z.f>c2tilda

AllRej=AllRej+min(Rej.AP, Rej)
rejseq=cbind(rejseq, min(Rej.AP, Rej))
AllnT.r=cbind(AllnT.r, nT.r)
}
} # end of for loop
pctRej=AllRej/Nsimu
avg_nT.r=mean(AllnT.r)
nT.r_nT=avg_nT.r/nT
unf=sum(zone==-1)
prom=sum(zone==0)
fav=sum(zone==1)
punf=sum(rejseq[which(zone==-1)])/unf
ppr=sum(rejseq[which(zone==0)])/prom
pfav=sum(rejseq[which(zone==1)])/fav

result <- list(power=pctRej, avgss=avg_nT.r, intrej=interim/Nsimu, unfzone=c(unf/Nsimu,
punf), pprzone=c(prom/Nsimu, ppr), pfavzone=c(fav/Nsimu, pfav))
return(result)
}

#Example 1: Simulate overall Type I error for λ=0.5 and μT= μP + λ(μA - μP)
Method1(detAP.d=1.5, F=0.5, CP.l=0.31, CP.u=0.8, muT=1.25, muA=2, muP=0.5, sigma=1,
lambda=0.5, cA=1, cP=1, Nsimu=100000)

#Example 2: Simulate actual power for λ=0.1 and μT= μA
```

Method1(detAP.d=1.5, CP.l=0.31, CP.u=0.8, muT=2, muA=2, muP=0.5, sigma=1, lambda=0.1, cA=1, cP=1, , Nsimu=100000)

## RMLE for the three-arm non-inferiority test with binary outcome

```
##Input Parameters:
#XTe, XAe and XPe: The number of success respectively in the experimental treatment arm,
  active comparator arm and the placebo arm.
#nTe, nAe, nPe: The available sample size respectively in the experimental treatment arm,
  active comparator arm and the placebo arm.
#lambda fraction margin of the effect preservation test

##Output:
# piT.hat, piA.hat and piP.hat The restricted MLE for each arm found by Newton–Raphson
algorithm. If the local restriction is not found after 100,000 iterations the unrestricted MLE
will be given.

RMLE <- function(XTe, XAe, XPe, nTe, nAe, nPe, lambda) {

iA=XAe/nAe
iP=XPe/nPe

if (iA==1 || iA==0) {iA=runif(1, 0.1, 0.9)}
if (iP==1 || iP==0) {iP=runif(1, 0.1, 0.9)}

xi=c(iA, iP)
dd=c(1,1)
iter=0
s1=1
s2=1
s3=1

while ((dd[1]>1e-8 || dd[2]>1e-8)&&(s1>1e-10 || s2>1e-10 || s3>1e-10)) {
iter=iter+1

J=matrix(c(1,1,2,2), nrow=2)
intJ=1
 while (rankMatrix(J)<2) {
  intJ+1
  m11=-XTe*(1-lambda)/((1-lambda)*xi[1]+lambda*xi[2])^2-XAe/xi[1]^2
  m12=-XTe*lambda/((1-lambda)*xi[1]+lambda*xi[2])^2-XPe/xi[2]^2
  m21=(nTe-XTe)*(1-lambda)/(1-(1-lambda)*xi[1]-lambda*xi[2])^2+(nAe-XAe)/(1-
xi[1])^2
  m22=(nTe-XTe)*lambda/(1-(1-lambda)*xi[1]-lambda*xi[2])^2+(nPe-XPe)/(1-xi[2])^2
  J=matrix(data=c(m11, m21, m12, m22), nrow=2, ncol=2)
  if (rankMatrix(J)<2) {xi=runif(2, 0.1, 0.9)}
  }
```

```
f11=XTe/((1-lambda)*xi[1]+lambda*xi[2])+XAe/xi[1]+XPe/xi[2]-(nTe+nAe+nPe)
f21=(nTe-XTe)/(1-(1-lambda)*xi[1]-lambda*xi[2])+(nAe-XAe)/(1-xi[1])+(nPe-XPe)/(1-
xi[2])-(nTe+nAe+nPe)
F=c(f11, f21)



xii=xi-solve(J)%*%F
dd=abs(xii-xi)
xi=xii

if (iter>100000) {break
  piA.hat=XAe/nAe
  piT.hat=XTe/nTe
  piP.hat=XPe/nPe}
if ( xi[1]<1e-4 || xi[2]<1e-4 || xi[1]>1-1e-4 || xi[2]>1-1e-4) {
   iter=iter+1
   xi=runif(2, 0.1, 0.9)
}

piA.hat=xii[1]
piP.hat=xii[2]
piT.hat=(1-lambda)*piA.hat+lambda*piP.hat

totN=nTe+nAe+nPe

s1=abs(XTe/piT.hat+XAe/piA.hat+XPe/piP.hat-totN)
s2=abs((nTe-XTe)/(1-piT.hat)+(nAe-XAe)/(1-piA.hat)+(nPe-XPe)/(1-piP.hat)-totN)
s3=abs(piT.hat-(1-lambda)*piA.hat-lambda*piP.hat)
}

return(list(piT.hat=piT.hat, piA.hat=piA.hat, piP.hat=piP.hat))
```

## **Method 1, binary outcome**

```
##Input Parameters:
#alpha nominal significance level
#beta   nominal type II error (i.e., 1 – power)
# F  information fraction, i.e., the fraction of sample size at the interim stage
#PiT.t, piA.t, piP.t  true success rate of the experimental treatment arm, active comparator
arm and placebo arm respectively
#PiT.a, piA.a, piP.a assumed success rate of the experimental treatment arm, active
comparator arm and placebo arm respectively at the trial design stage
#lambda fraction margin of the effect preservation test
#cA     the ratio of sample size in the active comparator arm over the experimental
treatment arm
```

```
#cP     the ratio of sample size in the placebo arm over the experimental treatment arm
#Nsimu number of simulations
##Output
#pctRej percentage calculated as the number of rejections at the final stage among all the
total number of simulated trials, i.e., the actual power by simulation
#avg_nT.r average number of recalculated sample size of the experimental treatment arm


Method1.bin <- function(piT.a, piA.a, piP.a, piT.t, piA.t, piP.t, F, lambda, cA, cP,
Nsimu=10000) {
  library('gsDesign')
  tao=function (lambda, pT, pA, pP) {
       sqrt(pT*(1-pT)+(1-lambda)^2*pA*(1-pA)/cA+lambda^2*pP*(1-pP)/cP)
       }
   detAP.a=piA.a-piP.a

 tao.0=tao(lambda, pT.0, pA.0, pP.0)
 tao.1=tao(lambda, piT.a, piA.a, piP.a)

 nT <- ceiling((Za*tao.0+Zb*tao.1)^2/(lambda*detAP.a)^2)
 nA <- ceiling(cA*nT)
 nP <- ceiling(cP*nT)

 nT.1 <- ceiling(nT*F)
 nA.1 <- ceiling(nA*F)
 nP.1 <- ceiling(nP*F)

 set.seed(123456)
 AllRej=0
 AllnT.r=NULL
 for (N in 1:Nsimu)
 {
  XT1 <- rbinom(1, nT.1, piT.t)
  XA1 <- rbinom(1, nA.1, piA.t)
  XP1 <- rbinom(1, nP.1, piP.t)

  pThat1=XT1/nT.1
  pAhat1=XA1/nA.1
  pPhat1=XP1/nP.1

  detAP.1=pAhat1-pPhat1;

  nT.r <- ceiling((Za*tao.0+Zb*tao.1)^2/(lambda*detAP.1)^2)

  nT.r=min(4*nT, nT.r)
  if (nT.r<=nT.1) (nT.r=nT.1+1)
```

```
   nA.r <- ceiling(cA*nT.r)
   nP.r <- ceiling(cP*nT.r)


   XT2 <- rbinom(1, nT.r-nT.1, piT.t)
   XA2 <- rbinom(1, nA.r-nA.1, piA.t)
   XP2 <- rbinom(1, nP.r-nP.1, piP.t)

   pT.f=sum(XT1, XT2)/nT.r
   pA.f=sum(XA1, XA2)/nA.r
   pP.f=sum(XP1, XP2)/nP.r

   z.AP.f=testBinomial(sum(XA1, XA2), sum(XP1, XP2), nA.r, nP.r, delta0=0, chisq=0, adj=0,
scale="Difference", tol=.1e-10)
   Rej.AP=z.AP.f>Za

   MLE.f <- RMLE(sum(XT1, XT2), sum(XA1, XA2), sum(XP1, XP2), nT.r, nA.r, nP.r, lambda)
   Z.f=(pT.f-(1-lambda)*pA.f-lambda*pP.f)/sqrt(Sgm(MLE.f$piT.hat, MLE.f$piA.hat,
MLE.f$piP.hat)/nT.r)
   Rej=Z.f>Za

   AllRej=AllRej+min(Rej.AP, Rej)
   AllnT.r=cbind(AllnT.r, nT.r)
  }
  pctRej=AllRej/Nsimu
  avg_nT.r=mean(AllnT.r)
  return(list(pctRej, avg_nT.r))
}
```

#Example 1: Simulate overall Type I error for $\lambda$=0.5 and $\pi_T = \pi_P + \lambda(\pi_A - \pi_P)$
Method1.bin(piT.a=0.8, piA.a=0.8, piP.a=0.2, piT.t=0.5, piA.t=0.8, piP.t=0.2, F=0.5, lambda=0.5, cA=1, cP=1, Nsimu=100000)

#Example 2: Simulate actual power for $\lambda$=0.1 and $\pi_T = \pi_A$
Method1.bin(piT.a=0.8, piA.a=0.8, piP.a=0.2, piT.t=0.8, piA.t=0.8, piP.t=0.2, F=0.5, lambda=0.5, cA=1, cP=1, Nsimu=100000)

## **Method 2, binary outcome**

##Input Parameters:
#alpha nominal significance level
#beta  nominal type II error (i.e., 1 – power)
#F  information fraction, i.e., the fraction of sample size at the interim stage
#CP.u upper bound of the promising zone of the conditional power
#CP.l lower bound of the promising zone of the conditional power
#PiT.t, piA.t, piP.t  true success rate of the experimental treatment arm, active comparator arm and placebo arm respectively

```
#PiT.a, piA.a, piP.a assumed success rate of the experimental treatment arm, active comparator arm and placebo arm respectively at the trial design stage
#lambda fraction margin of the effect preservation test
#cA     the ratio of sample size in the active comparator arm over the experimental treatment arm
#cP     the ratio of sample size in the placebo arm over the experimental treatment arm
#Nsimu number of simulations
##Output
#pctRej percentage calculated as the number of rejections at the final stage among all the total number of simulated trials, i.e., the actual power by simulation
#avgss average number of recalculated sample size of the experimental treatment arm
#intrej the probability of rejection at the interim stage for excessive efficacy
#unfzone a list with two elements; the first element is the probability of falling into the unfavorable zone and the second element is the probability of rejection at the final stage in the unfavorable zone
#pprzone a list with two elements; the first element is the probability of falling into the promising zone and the second element is the probability of rejection at the final stage in the promising zone
#pfavzone a list with two elements; the first element is the probability of falling into the favorable zone and the second element is the probability of rejection at the final stage in the favorable zone

Method2.bin <- function(alpha, beta, F, piT.t, piA.t, piP.t, piT.a, piA.a, piP.a, CP.l, CP.u, lambda, cA, cP, Nsimu) {
library(ldbounds)
Sgm=function(piT, piA, piP) {
sgm = piT*(1-piT)+piA*(1-piA)*(1-lambda)^2/cA+piP*(1-piP)*lambda^2/cP
return(sgm)
}

obf.bd=bounds(t=c(0.5,1), iuse=1, alpha=alpha)
c1tilda=obf.bd$upper.bounds[1]
c2tilda=obf.bd$upper.bounds[2]

tao=function (lambda, pT, pA, pP) {
  sqrt(pT*(1-pT)+(1-lambda)^2*pA*(1-pA)/cA+lambda^2*pP*(1-pP)/cP)
}
tao.0=tao(lambda, pT.0, pA.0, pP.0)
tao.1=tao(lambda, piT.a, piA.a, piP.a)
phi1=piT.a-(1-lambda)*piA.a-lambda*piP.a
nT <- ceiling((c2tilda*tao.0+Zb*tao.1)^2/(phi1)^2)
nA <- ceiling(cA*nT)
nP <- ceiling(cP*nT)

nT.1 <- ceiling(nT*0.5)
nA.1 <- ceiling(nA*0.5)
nP.1 <- ceiling(nP*0.5)
```

```
set.seed(123456)
AllRej=0
interim=0
zone=NULL
AllnT.r=NULL
rejseq=NULL

for (N in 1:Nsimu)
{
XT1 <- rbinom(1, nT.1, piT.t)
XA1 <- rbinom(1, nA.1, piA.t)
XP1 <- rbinom(1, nP.1, piP.t)

pThat1=XT1/nT.1
pAhat1=XA1/nA.1
pPhat1=XP1/nP.1


z.AP.1=testBinomial(XA1, XP1, nA.1, nP.1, delta0=0, chisq=0, adj=0, scale="Difference",
tol=.1e-10)
Rej.AP1= (z.AP.1>c1tilda)

RMLE1 <- RMLE(XT1, XA1, XP1, nT.1, nA.1, nP.1, lambda)
theta1=pThat1-(1-lambda)*pAhat1-lambda*pPhat1
Z.1=theta1/sqrt(Sgm(RMLE1$piT.hat, RMLE1$piA.hat, RMLE1$piP.hat)/nT.1)
Rej.TA1=Z.1>c1tilda

Rej1=min(Rej.AP1, Rej.TA1)

if (Rej1==1) {
nT.r=nT.1
AllRej=AllRej+Rej1
rejseq=cbind(rejseq, Rej1)
AllnT.r=cbind(AllnT.r, nT.1)
interim=interim+1
zone=cbind(zone, 1)
} else {

tao0hat=sqrt(Sgm(RMLE1$piT.hat,RMLE1$piA.hat, RMLE1$piP.hat))
CP=1-pnorm((c2tilda*sqrt(nT)-Z.1*sqrt(nT.1))*tao0hat/tao1/sqrt(nT-nT.1)-
theta1/tao1/(nT-nT.1))
if (CP<CP.l) {zone=cbind(zone, -1)}
if (CP>CP.u) {zone=cbind(zone, 1)}
if (CP>=CP.l & CP<=CP.u) {
zone=cbind(zone, 0)
```

```
#nT.r=nT.1+Sgm(MLE$piT.hat, MLE$piA.hat, MLE$piP.hat)/(theta1)^2*((c2tilda*sqrt(nT)-
Z.1*sqrt(nT.1))/sqrt(nT-nT.1)+Zb)^2
#nT.r=min(nT.r, 4*nT)
nT.r=nT
CP.d=CP
while (CP.d<=1-beta && nT.r < 4*nT ) {
 nT.r=nT.r+1
 CP.d=1-pnorm((c2tilda*sqrt(nT.r)-Z.1*sqrt(nT.1))/sqrt(nT.r-nT.1)-
theta1/sqrt(Sgm(MLE$piT.hat, MLE$piA.hat, MLE$piP.hat)/(nT.r-nT.1)))
}
 #4*nT as upper limit for sample size increase
} else {
#if not in promising zone, continue w/o changing sample size
nT.r <- nT
}

nT.r <- ceiling(nT.r)
nA.r <- ceiling(cA*nT.r)
nP.r <- ceiling(cP*nT.r)


XT2 <- rbinom(1, nT.r-nT.1, piT.t)
XA2 <- rbinom(1, nA.r-nA.1, piA.t)
XP2 <- rbinom(1, nP.r-nP.1, piP.t)

pT.f=sum(XT1, XT2)/nT.r
pA.f=sum(XA1, XA2)/nA.r
pP.f=sum(XP1, XP2)/nP.r

if (Rej.AP1==1) { Rej.AP=1 } else {
z.AP.f=testBinomial(sum(XA1, XA2), sum(XP1, XP2), nA.r, nP.r, delta0=0, chisq=0, adj=0,
scale="Difference", tol=.1e-10)
Rej.AP=z.AP.f>c2tilda }

RMLE.f <- RMLE(sum(XT1, XT2), sum(XA1, XA2), sum(XP1, XP2), nT.r, nA.r, nP.r, lambda)
Z.f=(pT.f-(1-lambda)*pA.f-lambda*pP.f)/sqrt(Sgm(RMLE.f$piT.hat, RMLE.f$piA.hat,
RMLE.f$piP.hat)/nT.r)
Rej=Z.f>c2tilda

AllRej=AllRej+min(Rej.AP, Rej)
rejseq=cbind(rejseq, min(Rej.AP, Rej))
AllnT.r=cbind(AllnT.r, nT.r)
}
} # end of for loop
pctRej=AllRej/Nsimu
avg_nT.r=mean(AllnT.r)
nT.r_nT=avg_nT.r/nT
```

```
unf=sum(zone==-1)
prom=sum(zone==0)
fav=sum(zone==1)
punf=sum(rejseq[which(zone==-1)])/unf
ppr=sum(rejseq[which(zone==0)])/prom
pfav=sum(rejseq[which(zone==1)])/fav

result <- list(power=pctRej, avgss=avg_nT.r, ratio=nT.r_nT, intrej=interim/Nsimu,
          unfzone=c(unf/Nsimu, punf), pprzone=c(prom/Nsimu, ppr),
pfavzone=c(fav/Nsimu, pfav))
return(result)
}
```

## REFERENCES


1. Head SJ, Kaul S, Bogers AJJC, Pieter KA. Non-inferiority study design: lessons to be learned from cardiovascular trials. *European Heart Journal;*2012; 33(11): 1318-1324.
2. D'Agostino R, Massaro J, Sullivan L. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statistics in Medicine;*2003;22:169-186.
3. Garattini S, Bertele V. Non-inferiority trials are unethical because they disregard patients' interests. *The Lancet*; 2007;370:1875-1877.
4. Koch A, Röhmel J. Hypothesis testing in the "gold standard" design for proving the efficacy of an experimental treatment relative to placebo and a reference. *Journal of Biopharmaceutical Statistics;* 2004;14(2):315-325.
5. ICH. Choice of Control Group and Related Issues in Clinical trials. *ICH Expert Work Group;* 2000; E10(July):1-35.
6. CHMP. Guideline on the choice of the non-inferiority margin. 2005
7. Hauschke D, Pigeot I. Establishing efficacy of a new experimental treatment in the "gold standard" design. *Biometrical Journal;* 2005;47(6):782-786.
8. Röhmel J. Discussion on "establishing efficacy of a new experimental treatment in the 'gold standard' design." *Biometrical Journal;* 2005;47(6):790-791.
9. Lewis JA. Discussion on "Establishing Efficacy of a New Experimental Treatment in the 'Gold Standard' Design." *Biometrical Journal;* 2005; 47(6): 787-789.
10. Pigeot I, Schäfer J, Röhmel J, Hauschke D. Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine;* 2003;22(6):883-899.
11. Maurer W, Hothorn L, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. *Biometrie in der Chemisch-Pharmazeutischen Industrie*; 1995;6:3-18.
12. Koch GG, Davis SM, Anderson RL. Methodological advances and plans for improving regulatory success for confirmatory studies. *Statistics in Medicine;* 1998;17(15-16):1675-1690.
13. Koch GG, Tangen CM. Nonparametric analysis of covariance and its role in non-inferiority clinical trials. *Therapeutic Innovation & Reg. Sci;* 1999;33:1145-59.
14. Schwartz TA, Denne JS. A two-stage sample size recalculation procedure for placebo- and active-controlled non-inferiority trials. *Statistics in Medicine;* 2006; 25(19):3396-3406.
15. Armitage P, McPherson CK, Rowe BC. Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society;* 2016;37(1):129-145.
16. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika;* 1977;64(2):191-199.

17. Wang SK, Tsiatis AA, Wang SK. Approximately Optimal One-Parameter Boundaries for Group Sequential Trials, *International Biometric Society Stable*; 2016;43(1):193-199.
18. Emerson SS, Fleming TR. Symmetric Group Sequential Test Designs. *Biometrics*; 1989; 45(3):905-923.
19. Pampallona S, Tsiatis AA. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference;* 1994;42(1):19-35.
20. Lan KKG, DeMets DL. Discrete Sequential Boundaries for Clinical Trials. *Biometrika*; 1983;70(3):659-663.
21. Cui L, Hung HMJ, Wang S-J. Modification of Sample Size in Group Sequential Clinical Trials. *Biometrics;* 1999;55(3):853-857
22. Denne JS. Sample size recalculation using conditional power. *Statistics in Medicin*e; 2001;20(17-18):2645-2660.
23. Chen YHJ, DeMets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicin*e;2004;23(7):1023-1038.
24. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicin*e; 2011;30(28):3267-3284.
25. Li G, Gao S. A group sequential type design for three-arm non-inferiority trials with binary endpoints. *Biometrical Journal;* 2010;52(4):504-518..
26. Gao P, Ware JH, Mehta C. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics;* 2008; 18(6): 1184-1196.
27. Miettinen OS, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine;* 1985;4:213-226.
28. Farrington C. P. and Manning G., Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk. *Statistics in Medicin*e, 1990;9:1447-1454
29. Kieser M. and Friede T., Planning and analysis of three-arm non-inferiority trials with binary endpoints, *Statistics in Medicin*e, 2007; 26:253–273.

**CURRICULUM VITAE**