

2013

Sample size re-estimation in active controlled non-inferiority clinical trials using a frequentist approach

<https://hdl.handle.net/2144/49300>

Downloaded from OpenBU. Boston University's institutional repository.

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**SAMPLE SIZE RE-ESTIMATION IN ACTIVE CONTROLLED
NON-INFERIORITY CLINICAL TRIALS
USING A FREQUENTIST APPROACH**

by

WEI GUO

B.S., Wuhan University, 1995
M.S., University of Massachusetts at Amherst, 2003

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2013

© 2013 by
WEI GUO
All rights reserved

Approved by

First Reader

Joseph Massaro, Ph.D.
Professor of Biostatistics

Second Reader

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

Third Reader

Michael Pencina, Ph.D.
Associate Professor of Biostatistics

*This dissertation is dedicated to my husband Qibin,
and sons Ben and Matthew*

ACKNOWLEDGEMENTS

I would be remiss if I did not express my gratitude to many who have been instrumental in guiding me through this endeavor.

I owe a great debt of gratitude to my research advisor Professor Joseph Massaro for all of his support and encouragement during the course of this research. He has demonstrated enormous patience through the past several years, and his skill and insight in approaching our research problems have been illuminating to me. I am very grateful to have had the opportunity to work closely with him and to learn from him.

I would also like to thank my readers: Professor Ralph D'Agostino, and Professor Michael Pencina. In addition to their unwavering support and flexibility, they each provided valuable insight and context for this work that will no doubt help me in my career. I truly hope to continue working with each of them in the years ahead.

I also wish to thank the other members of my committee, Professor Timothy Heeren and Professor Debbie Cheng, for their willingness to serve in those roles. Their helpful guidance and probing questions have improved the quality of this dissertation.

I cannot adequately express my appreciation to my advisor at work, Dr. Ilker Yalcin. He has been a mentor to me, but he has been much more than that. I have observed his wisdom, kindness, and gentleness in action, and I hope to emulate these qualities in my own life.

In addition to the entire BU Biostatistics Department, I would like to thank Professor Howard Cabral for being such a great director of graduate studies, and for the countless times he helped me along the way.

Many others have contributed to my professional growth. Among these are my colleagues at Harvard Clinical Research Institute and Synta Pharmaceutical Inc. I have enjoyed many rewarding opportunities of working with them. Special thanks go to Kim Norris and Anne Harrington for their invaluable assistance in revising and formatting this dissertation.

I also need to thank my husband Qibin for his love and support. He has been so understanding of my lack of free time or energy for household duties, familial time and kids' activities over the last number of years. He has done a tremendous job of keeping our family close and functional. With this time behind us, I look forward to many years of work, growth, and happiness.

Finally, I would like to honor my parents. I have never doubted their unconditional support for me throughout the years. They have clearly demonstrated their love in practical ways and have set a wonderful example for me in many areas of life.

**SAMPLE SIZE RE-ESTIMATION IN ACTIVE CONTROLLED
NON-INFERIORITY CLINICAL TRIALS
USING A FREQUENTIST APPROACH**

(Order No.)

WEI GUO

Boston University Graduate School of Arts and Sciences, 2013

Major Professor: Joseph Massaro, Professor of Biostatistics

ABSTRACT

In active controlled clinical trials, a possible objective is to test a non-inferiority hypothesis that the experimental treatment is therapeutically not inferior to the active control within a pre-defined margin. At the design stage, the misspecification of any design parameters (e.g., the variance or treatment difference for continuous endpoints, control event rate, or non-inferiority margin for binary endpoints) can lead to study power below the desired level. Sample size re-estimation (SSR) procedures protect study power by allowing sample size re-estimation based on an interim analysis using revised estimates of the design parameters.

For continuous endpoints, current approaches to SSR for non-inferiority trials focus on updating the sample size based solely on the estimated variance (blinded or unblinded) at the interim. The SSR using both sample variance and the observed treatment difference at interim in conditional power calculations is used in superiority trials. We have

extended the methodology to non-inferiority trials, quantified the effect on the Type I error rate, and proposed controlling it by modifying the critical value and/or stopping the trial at the interim for futility.

For binary endpoints, current approaches to SSR for non-inferiority trials focus on estimating the event rates (blinded or unblinded) at the interim and update the sample size solely on the estimated event rates at the interim without updating the non-inferiority margin. A procedure that adapts both the absolute non-inferiority margin, and sample size based on the underlying interim observed pooled (blinded) event rate, and updates non-inferiority margin again at the final analysis based on the observed estimate of the event rate in control group at the end of the study is proposed.

Our simulation results show the proposed adaptive procedures for extending a study by adding sample size, if necessary, preserve the overall type I error rate and maintain desired power. Combining sample size re-estimation methods with early stopping rules for continuous endpoints and adapting non-inferiority margins for binary endpoints could increase study flexibility, scope, and efficiency of non-inferiority trials.

The proposed methodologies can be used for designing efficient two-stage non-inferiority trials with sample size re-estimation in active controlled non-inferiority clinical trials.

TABLE OF CONTENTS

Chapter 1: INTRODUCTION.....	1
Chapter 2: LITERATURE REVIEW.....	9
2.1 Introduction and Background.....	9
2.1.1 Fixed sample size design.....	9
2.1.2 Two-stage sample size re-estimation design for a superiority design.....	10
2.2 SSR Methods Used for a Continuous Endpoint.....	12
2.2.1 Two-stage sample size re-estimation using only the interim observed variance and unconditional power calculation for a continuous endpoint	12
2.2.2 Two-stage sample size re-estimation using the interim observed variance and the originally specified treatment difference in a conditional power calculation for a continuous endpoint	14
2.2.3 Two-stage sample size re-estimation using the interim observed variance and treatment difference in a conditional power calculation for a continuous endpoint	16
2.2.4 Two stage sample size re-estimation using the interim observed variance and treatment difference while also permitting early stopping to reject or accept H₀	17
2.3 SSR Methods Used for a Binary Endpoint	18
2.3.1 Sample size estimation in fixed design for a binary endpoint.....	18
2.3.2 Two-stage sample size re-estimation using the blinded overall event rate for a binary endpoint.....	21
2.3.3 Two-stage sample size re-estimation using the unblinded pooled event rate for a binary endpoint	23
2.3.4 Adapting non-inferiority margin with the underlying event rate in the control arm for a binary endpoint	24
2.4 Review of Related Topics	25
2.4.1 Size of first stage samples	25
2.4.2 Choice of non-inferiority margin	26

Chapter 3: NORMALLY DISTRIBUTED ENDPOINT	29
3.1 Introduction and Background.....	29
3.2 Computational Methods	31
3.3 Fixed Sample Size Design.....	33
3.4 Two-stage SSR using only the observed variance and unconditional power calculation	36
3.4.1 Sample size re-estimation based on blinded estimate of the variance	37
3.4.2 Sample size re-estimation based on unblinded estimate of the variance	38
3.4.3 Comparison of the procedures.....	39
3.5 Incorporating conditional power into two-stage SSR in a superiority setting ...	41
3.5.1 Calculating conditional power in a non-inferiority setting	41
3.5.2 Type I error Rate (α) in the non-inferiority setting	43
3.5.3 Determining additional sample size (n2) and critical value adjustment required to maintain nominal Type I error rate	47
3.5.4 General Procedures for SSR using conditional power	49
3.6 Two-stage SSR using the observed variance and the originally specified treatment difference in a conditional power calculation.....	50
3.6.1 No futility stopping at interim and no critical value adjustment at the final analysis	52
3.6.2 With futility stopping at interim and no critical value adjustment at the final analysis	54
3.6.3 No futility stopping at interim and with critical value adjustment at the final analysis	55
3.6.4 With futility stopping at interim and critical value adjustment at the final analysis	56
3.6.5 Comparison of procedures.....	57
3.7 Two-stage SSR using the observed variance and observed treatment difference in a conditional power calculation	60
3.7.1 No futility stopping at interim and no critical value adjustment at the final analysis	62

3.7.2	With futility stopping at interim and no critical value adjustment at the final analysis	63
3.7.3	No futility stopping at interim and with critical value adjustment at the final analysis	64
3.7.4	With futility stopping at interim and critical value adjustment at the final analysis	65
3.7.5	Comparison of procedures.....	66
Chapter 4: BINOMILALLY DISTRIBUTED ENDPOINTS.....		73
4.1	Introduction and Background.....	73
4.2	Computational Methods	76
4.3	Fixed Sample Size Design.....	78
4.3.1	Risk difference approach without adaptive margin.....	78
4.3.2	Relative risk approach	82
4.3.3	Risk difference approach with adaptive non-inferiority margin based on observed control group rate at the end of the study	85
4.3.4	Comparison of fixed designs.....	88
4.4	Design with Blinded Two-Stage SSR	90
4.4.1	Two Stage SSR based on pooled event rate at interim without changing the non-inferiority margin using the risk difference approach.....	93
4.4.2	Two-stage SSR based on pooled event rate at interim using the relative risk approach	94
4.4.3	Two-stage SSR based on pooled event rate at interim using the risk difference approach -- margin was updated under H_a at interim.....	96
4.4.4	Two-stage SSR based on pooled event rate at interim using the risk difference approach -- margin was updated under H₀ at interim	101
4.4.5	Comparison of two-stage SSR designs	101
Chapter 5: SUMMARY AND DISCUSSION.....		107
5.1	Two-stage SSR in normally distributed outcome	107
5.2	Two-stage SSR in binomially distributed outcome.....	110

5.3 Limitation and Future Directions 114

Chapter 6: CONCLUSION 116

Appendix A: TABLES 118

Appendix B: PROGRAMS 176

REFERENCES 190

CURRICULUM VITAE 198

LIST OF TABLES

Table 1:	Parameter values used for the computations of Type I error rate and power for non-inferiority tests with continuous outcome.....	32
Table 2:	Consequences of overestimating treatment difference θ and/or underestimating variance σ^2 with continuous outcomes ($\alpha=0.025$, planned Power=80%)	36
Table 3:	Type I error rates at different true variances for two-stage SSR using only the observed variance and unconditional power calculation in trials with a continuous outcome	39
Table 4:	Power at different true variances for two-stage SSR using only the observed variance and unconditional power calculation in trials with a continuous outcome	40
Table 5:	General Procedures for two-stage SSR using conditional power in trials with a continuous outcome	49
Table 6:	Four different scenarios in two-stage SSR using conditional power in trials with a continuous outcome	51
Table 7:	Final sample size at different true variances in two-stage SSR using the observed variance and the originally specified treatment difference in a conditional power calculation in trials with a continuous outcome under the null hypothesis	58
Table 8:	Type I error rate at different true variances in two-stage SSR using the observed variance and the originally specified treatment difference in a conditional power calculation in trials with a continuous outcome.....	59
Table 9:	Power at different true variances in two-stage SSR using the observed variance and the originally specified treatment difference in a conditional power calculation in trials with a continuous outcome.....	59
Table 10:	Final sample size at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome under the null hypothesis	67
Table 11a:	Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome ($\sigma_{12} = \sigma_{22}$)	67

Table 11b:	Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome ($\sigma_{12} = 0.75 \times \sigma_{22}$) 68
Table 11c:	Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome ($\sigma_{12} = 0.5 \times \sigma_{22}$) 68
Table 11d:	Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome ($\sigma_{12} = 0.25 \times \sigma_{22}$) 69
Table 11e:	Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome (without capping the stage 2 sample size)..... 70
Table 12:	Power at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome 71
Table 13:	Parameter values used for the simulations of Type I error rate and power for non-inferiority tests with binary outcome 78
Table 14:	Power at different true event rates in fixed designs with binary positive outcomes 88
Table 15:	Power at different true event rates in fixed designs with binary negative outcomes 89
Table 16:	General Procedures for blinded two-stage SSR with binary outcomes 90
Table 17:	Type I error rates at different true event rate in relative risk design with binary negative outcomes 96
Table 18:	Type I error rates at different true event rates in risk difference with adaptive margin design with binary positive outcomes 99
Table 19:	Type I error rates at different true event rate in risk difference with adaptive margin design with binary negative outcomes 100
Table 20:	Final test statistics in relative risk approach and procedure with adaptive margin using the risk difference approach..... 102

Table 21:	Type I error rates at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis) – Positive Outcome	102
Table 22:	Type I error rates at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis) – Negative Outcome.....	103
Table 23:	Power at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis) – Positive Outcome.....	104
Table 24:	Type I error rates at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis) – Negative Outcome.....	104
Table 25:	Final sample size at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis).....	105
Appendix A:	TABLES.....	118
Table A3.1:	Type I error rate at different common unknown variances with fixed design	118
Table A3.2:	Power at various common unknown variances and treatment difference with fixed design.....	119
Table A3.3a:	Type I error rate at different common unknown variances when only variance was updated at the interim – Blinded estimate of variance	120
Table A3.3b:	Type I error rate at different common unknown variances when only variance was updated at the interim – Unblinded estimate of variance..	121
Table A3.4a:	Power at different common unknown variances when only variance was updated at the interim – Blinded estimate of variance.....	122
Table A3.4b:	Power at different common unknown variances when only variance was updated at the interim – Unblinded estimate of variance	123
Table A3.5a:	Type I error rate at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$	124

Table A3.5b: Type I error rate at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$	124
Table A3.6: Power at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$	125
Table A3.7: Type I error rate at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment (allowing early stopping for futility) – Assuming $\mu_C - \mu_T = 0$	126
Table A3.8: Power at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment (allowing early stopping for futility) – Assuming $\mu_C - \mu_T = 0$	127
Table A3.9a: Type I error rate at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$	130
Table A3.9b: Type I error rate at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$	130
Table A3.10: Power at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$	131
Table A3.11: Type I error rate at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (allowing early stopping for futility) – Assuming $\mu_C - \mu_T = 0$	132
Table A3.12: Power at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (allowing early stopping for futility) – Assuming $\mu_C - \mu_T = 0$	133

Table A3.13a: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Under $\mu_C - \mu_T$	136
Table A3.13b: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Under $\mu_C - \mu_T$	136
Table A3.14: Power at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Under $\mu_C - \mu_T$	137
Table A3.15: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (allowing early stopping for futility) – Under $\mu_C - \mu_T$	138
Table A3.16: Power at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (allowing early stopping for futility) – Under $\mu_C - \mu_T$	139
Table A3.17a: Type I error rate at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Under $\mu_C - \mu_T$	142
Table A3.17b: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Under $\mu_C - \mu_T$	142
Table A3.18: Power at different common unknown variances when both variance and treatment difference were updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Under $\mu_C - \mu_T$	143
Table A3.19: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP with critical value adjustment (allowing early stopping for futility) – Under $\mu_C - \mu_T$	144

Table A3.20: Power at different common unknown variances when both variance and treatment difference were updated at the interim using CP with critical value adjustment (allowing early stopping for futility) – Under $\mu_C - \mu_T$	145
Table A4.1a: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the fixed sample size design with Risk Difference Approach– Positive Outcome.....	148
Table A4.1b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the fixed sample size design with Risk Difference Approach – Negative Outcome.....	149
Table A4.2a: Power for Blackwelder’s and Farrington-Manning’s tests in the fixed sample size design with different true event rates– Positive Outcome...	150
Table A4.2b: Power for Blackwelder’s and Farrington-Manning’s tests in the fixed sample size design with different true event rates – Negative Outcome	151
Table A4.3a: Type I error rate for Normal Approximation and Farrington-Manning’s tests in the fixed sample size design with relative risk approach– Positive Outcome.....	152
Table A4.3b: Type I error rate for Normal Approximation and Farrington-Manning’s tests in the fixed sample size design with relative risk approach – Negative Outcome.....	153
Table A4.4a: Power for Normal Approximation and Farrington-Manning’s tests in the fixed sample size design with relative risk approach– Positive Outcome	154
Table A4.4b: Power for Normal Approximation and Farrington-Manning’s tests in the fixed sample size design with relative risk approach – Negative Outcome .	155
Table A4.5a: Type I error rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in Risk Difference Approach– Positive Outcome.....	156
Table A4.5b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in Risk Difference Approach – Negative Outcome	157

Table A4.6a: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in Risk Difference Approach– Positive Outcome	158
Table A4.6b: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in Risk Difference Approach – Negative Outcome	159
Table A4.7a: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with fixed margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome	160
Table A4.7b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with fixed margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome	161
Table A4.8a: Power for Blackwelder’s and Farrington-Manning’s tests in the design with fixed margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome	162
Table A4.8b: Power for Blackwelder’s and Farrington-Manning’s tests in the design with fixed margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome	163
Table A4.9a: Type I error Rate for Normal Approximation and Farrington-Manning’s tests in the design with relative risk approach with sample size re-estimation at the interim– Positive Outcome	164
Table A4.9b: Type I error Rate for Normal Approximation and Farrington-Manning’s tests in the design with relative risk approach with sample size re-estimation at the interim– Negative Outcome	165
Table A4.10a: Power for Normal Approximation and Farrington-Manning’s tests in the design with relative risk approach with sample size re-estimation at the interim– Positive Outcome	166
Table A4.10b: Power for Normal Approximation and Farrington-Manning’s tests in the design with relative risk approach with sample size re-estimation at the interim– Negative Outcome	167
Table A4.11a: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome	168

Table A4.11b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome	169
Table A4.12a: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome	170
Table A4.12b: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome	171
Table A4.13a: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome	172
Table A4.13b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome	173
Table A4.14a: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome	174
Table A4.14b: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome	175

LIST OF FIGURES

Figure 1:	The progression of approaches to two-stage SSR in normally distributed outcome.....	31
Figure 2:	The progression of approaches to two-stage SSR in binomially distributed outcome.....	76
Figure 3:	Performance of each procedure discussed in Chapter 3	107
Figure 4:	Performance of each procedure discussed in Chapter 4	111

Appendix B: PROGRAMS

Programs for continuous endpoints	176
Programs for binary endpoints.....	182

Chapter 1: INTRODUCTION

In clinical research, a superiority trial is a common and traditional design for between-treatment comparisons, as it is useful for determining if experimental treatment is clinically and statistically different from the control treatment and, if so, in which direction. This is the standard approach used in most clinical trials, where the hypothesis to be tested is that an experimental treatment is superior to a comparison treatment. However, it has become increasingly important in clinical research to evaluate a different question: determining if the new treatment's efficacy or safety is not clinically worse than that of standard treatment. Traditional superiority significance testing designs and analyses cannot sufficiently address this latter question. Non-inferiority is one of the conceptual research models to address those research questions that involve hypotheses of "no meaningful clinical difference" between examined outcomes.

Non-inferiority designs differ from standard superiority trials in several fundamental ways. The null hypothesis of standard superiority trials is that there is no true difference between the treatments. In other words, unless strong evidence is found indicating the superiority of one treatment over the other, the default conclusion is that there is no evidence a treatment difference exists. A Type I error (probability of this error is denoted as α) erroneously concludes a treatment difference exists, based on the clinical trial sample, when in truth it does not (thus erroneously rejecting the null hypothesis). A Type II error (probability of this error is denoted as β) is a failure to reject the null hypothesis based on the clinical trial sample, when in fact the alternative is true (thus erroneously

missing an actual difference in treatments). In contrast, non-inferiority trials have a null hypothesis that the new treatment is inferior to the active standard therapy by at least a certain amount, a pre-specified non-inferiority margin (Δ). The alternative hypothesis to be proven is that the experimental treatment is inferior to the active control by less than that same amount (Δ). The Type I error for a non-inferiority trial is the probability of erroneously concluding non-inferiority when the novel treatment truly is inferior to the control and a Type II error is the probability of erroneously failing to reject the null hypothesis when the novel treatment is truly non-inferior.

Non-inferiority clinical trial design is becoming increasingly common and relevant in randomized controlled trials. Many randomized clinical trials for a new experimental product are conducted versus an active standard therapy as the control agent for testing an experimental treatment since effective treatments are available and thus placebos are often considered unethical – at least for longer treatment periods. The active control treatment, being a standard therapy for the disease under study, should have previously been shown to be efficacious as compared to the placebo¹. In this situation it is often sufficient for regulatory and marketing purposes to demonstrate that the new treatment is not clinically inferior to an active standard therapy by more than a pre-specified amount, the so-called non-inferiority margin (Δ)². As D'Agostino et al. pointed out, active control trials become more popular as more effective treatments become available for many diseases³.

In active controlled clinical trials without a placebo arm, a possible objective is to test a non-inferiority hypothesis that the experimental treatment is therapeutically not inferior to the active control within a pre-defined non-inferiority margin. At the initial study design stage, crucial required information used to determine sample size, such as expected event rate (dichotomous endpoint) or expected mean (continuous endpoint) in the active treatment group, may not be available, or may be estimated but with a high degree of uncertainty. When this occurs, it may be prudent to check the validity of those assumptions using interim data from the study and make a midcourse adjustment to the study design (“adapting the study”) if necessary in order to obtain the desired power. Much work has been done on the statistical methodology of adapting a superiority clinical trial in this manner^{4,5,6}.

Blackwelder summarized the statistical algorithms for assessing non-inferiority on an outcome variable as follows⁷. Without loss of generality, assuming the outcome is “positive” (larger values of the outcome indicate a more beneficial response to treatment), the null and alternative hypotheses for proving non-inferiority from a “difference” perspective (as opposed to a ratio perspective) are:

$$H_0: C - T \geq \Delta \text{ (T is inferior to C)} \quad (1)$$

$$H_1: C - T < \Delta \text{ (T is not inferior to C)}$$

Here, T and C represent a parameter (e.g., mean for continuous endpoint, proportion of an event for a binary endpoint) of the efficacy endpoint variable for the experimental treatment and the active control, respectively. $\Delta > 0$ is the pre-specified fixed non-

inferiority margin; that is, how much C can exceed T with T still being considered clinically non-inferior to C. The null hypothesis states that the active control C exceeds the experimental treatment T by at least Δ ; if this cannot be rejected, then the experimental treatment is considered inferior to the active control with respect to efficacy. The alternative hypothesis states that the active control may indeed have better efficacy than the experimental treatment, but by no more than a clinically relevant margin, Δ . In such a case, we reject the null hypothesis and conclude the experimental treatment is not clinically or therapeutically inferior to the active control.

Once the non-inferiority trial design is deemed appropriate and favorable for the targeted clinical trial objective, it is crucial to plan the sample size to be sufficient to ensure a desirable level of power. When planning a clinical trial, the investigators are often quite uncertain about the sizes of parameters (C and T) needed for sample size calculations. In a non-inferiority design, the sample size required to reject the null hypothesis is sensitive to the following parameters: the Type I error probability (α), Type II error probability (β), (and hence is sensitive to power for rejecting the null hypothesis, where power is defined as $1-\beta$), the non-inferiority margin (Δ), and the assumed value of T and C (or, alternatively, the assumed difference of T-C in trials with a continuous endpoint). Additionally, for continuous distributions, the power depends on the assumed within group variance of the endpoint variable.

Due to various reasons, such as incomplete knowledge regarding the disease and treatment under study, the assumed values of the treatment group binary endpoint rates or

continuous endpoint means and variances are often difficult to specify with reasonable accuracy before the trial begins. Therefore we often encounter uncertainty in the design of a clinical trial at its initial stage. If these parameters are mis-specified, the sample size calculated may yield to a study without adequate power, though that fact may never be known to us until the end of the trial at the earliest in a fixed design trial without any interim analysis.

One of solutions to this problem is (A) mid-course inspection of effect sizes or assumed treatment group rates/means, variance (either pooled blinded or unblinded) during the ongoing trial, (B) modify the design assumptions accordingly, and (C) re-calculate the sample size required to obtain adequate power according to these modified assumptions. For example for a continuous outcome, the treatment effect size and its associated variation can be estimated at the interim, then the original sample size could be revised based on these interim estimates. At the final stage, the analysis will include the data from both stages (pre-interim and post-interim stages). The primary concern with such an approach is the potential increase of the probability of a Type I error.

In this dissertation research, we focus on one of the common design adjustments: sample size re-estimation based on interim data. This approach has been applied substantially to superiority trials^{8,9}. We will discuss these adjustments for non-inferiority clinical trials. The discussion is applied to both continuous and binary endpoints.

For the continuous endpoint, the following questions will be investigated and answered:

1. At an interim point in a trial, if the observed pooled blinded estimate of within-group variance reflects a larger-than-anticipated value at the initial design stage, how can we use SSR to ensure adequate power to claim non-inferiority if non-inferiority truly exists, while maintaining the nominal Type I error?
2. At an interim point in a trial, if it appears that the treatment difference assumed at the initial design stage was too optimistic and that a smaller difference is more realistic but still clinically relevant, then sample size increase is desired to achieve the target power. What, if any, are the necessary statistical modifications for the final test statistic in order to maintain power and preserve the Type I error in non-inferiority clinical trials?
3. Combining early stopping for futility with SSR to increase study flexibility and efficiency. How will the early stopping of the clinical trial for futility in above scenarios affect the probability of a Type I error and power in a non-inferiority setting?

In non-inferiority trials with binary endpoints, in contrast to conventional superiority studies, the proper determination of the clinically acceptable absolute margin (Δ) in the set of hypothesis (1) is a vital problem. It is commonly accepted that Δ has to be stated in advance of the study and must not be a post hoc decision that occurs after the data have been analyzed. In ideal situations, where event rates for control group are well-established, a fixed Δ may be appropriate. However, fixed Δ s may be difficult to justify in situations in which the expected event rates are difficult to estimate. For example, if at the clinical trial design stage, the expected (positive) event rate for the control arm was

assumed to be 90%, it may be reasonable to set the non-inferiority margin to be 10% as is often done in anti-infective trials where the outcome is infection cure. Specifically, the null and alternative hypotheses are:

$$H_0: C - T \geq 0.10 \text{ (T is inferior to C)}$$

$$H_1: C - T < 0.10 \text{ (T is not inferior to C)}$$

However, if the control rate is truly only 80%, a 10% margin may be less clinically applicable. Thus, revising the fixed margin to reflect the revised estimate of the control group rate is often desired (e.g., if a 10% margin is appropriate for a 90% control group event rate, then an 8.9% margin, or $80\% \times (10\%/90\%)$, might be appropriate for an 80% control group event rate).

For the binary endpoints, the questions we are trying to answer in a non-inferiority trial based on a risk difference are:

1. At an interim point in a trial, if the observed pooled (blinded) event rate reflects a higher or lower value than anticipated at the initial design stage, how should we adapt the non-inferiority margin and resulting sample size to make sure the study is adequately powered to claim non-inferiority at the end of the trial, while maintaining the nominal Type I error?
2. What are the statistical and clinical implications of changing the absolute margin and sample size based on the interim estimate of pooled event rate, then revising

the margin again, if necessary, based on the estimate of the control group event rate at the end of the study?

Chapter 2: LITERATURE REVIEW

2.1 Introduction and Background

2.1.1 Fixed sample size design

Historically, many studies have taken a “fixed sample size” approach, in which the sample size required to yield adequate power is determined prior to the start of the clinical trial based on assumptions of the true within-treatment variance and the difference between means (in non-inferiority design, we usually assume the means of each treatment are the same) for continuous endpoints, and the true population event rates for binary endpoints (which, again, we usually assume are the same for each treatment group in a non-inferiority design). In a fixed design, there is no interim analysis and the sample size or any assumptions on which it was based are not updated during the course of the study. Advantages of the fixed-sample approach include its straightforward nature and well-known properties, its ease of implementation, clear planning implications with respect to the sample size, maintenance of aggregate blindness of the treatment assignments during the course of the study and hence minimizing bias entering the study, and maintenance of a nominal Type I error rate.

The notable disadvantage here is the vulnerability to the misspecification of these assumptions; depending on the nature of the misspecification, the study could be underpowered. In response to this disadvantage, one might adopt the methods of two-stage sample size re-estimation based on interim assessments of the assumptions and

updating the sample size based on these assessment, which would, on average, yield the “correct” sample size and produce a test with the required power.

2.1.2 Two-stage sample size re-estimation design for a superiority design

The two-stage sample size re-estimation proposed by Wittes and Brittain¹² for a two-group superiority trial can be described as a three-step procedure consisting of initial sample size calculation, sample size review at an interim stage, and final analysis on the entire sample, including patients enrolled prior to the interim analysis. The initial sample size calculation leading to a provisional sample size n_0 per group is carried out on the basis of initial assumptions of the nuisance parameters. For example, studies in earlier phases of the drug development process might inform the choice of initial estimates.

When the data from the patients in first stage $n_1 = \pi n_0$ (e.g., $\pi = 1/2$) are available, then the nuisance parameters are re-estimated in a blinded manner from these observations, which constitute the first-stage samples. These estimates are then used for sample size recalculation. The recalculated final sample size is denoted by \hat{N} . The sample size can then be adjusted following a predefined recalculation rule. For instance, Wittes and Brittain¹² proposed the so called restricted design allowing only upwards adjustments of the initially planned sample size n_0 , i.e., the sample size of the second stage n_2 is given by $n_2 = \max(n_0, \hat{N}) - n_1$. Birkett and Day¹⁰ suggested also allowing downwards adjustments, which is referred to as the unrestricted design, i.e. $n_2 = \max(n_1, \hat{N}) - n_1$. In practice, upper bounds for the adjustments may exist because of budget restrictions or limited recruitment resources. Also, a minimum number of patients is often required to study the treatment’s safety profile. These restrictions can be incorporated in the sample

size recalculation rule. The final analysis includes all $n = n_1 + n_2$ observations. The observations are not analyzed separately for each stage; rather all observations are pooled as in a fixed sample size design.

Despite all the previous theoretical work in SSR and the mathematical theory to be presented in this dissertation, please note that the International Conference on Harmonisation (ICH)¹¹ E9 guideline Section 4.4 states the following about sample size adjustment:

“The steps taken to preserve blindness and consequences, if any, for the Type I error [...] should be explained.”

The guideline raises two concerns for the two-stage sample size re-estimation design, namely blinding and actual significance level. Un-blinding of the allocation to treatment groups for trial participants during the ongoing study is a serious concern because it could cause bias. The actual significance level is also important because control of the Type I error rate is assured for a level- α -test only if the sample size is fixed.

We will now review the sample size recalculation techniques based on data obtained from one interim analysis, blinded or unblinded, to overcome the uncertainty of a priori assumptions of the design parameters by allowing the sample size to be modified through a re-specification of relevant design parameters during the study. These techniques have been widely used in superiority trials, and the dissertation will apply these and other methodologies to non-inferiority trials, as we will discuss later.

2.2 SSR Methods Used for a Continuous Endpoint

2.2.1 Two-stage sample size re-estimation using only the interim observed variance and unconditional power calculation for a continuous endpoint

The methods discussed in this section use the interim estimate of the variance only, but not an unblinded interim estimate of the treatment difference from the first-stage, to recalculate the overall target final sample size for a continuous outcome using unconditional power formulas (i.e., the same power formula that one used to initially design the study). This is a common practice in superiority setting. In the non-inferiority setting, which we will investigate further in this dissertation, we usually assume each treatment group has the same distribution when designing the study, therefore the estimate of the within-group variance is simply based on an interim blinded estimate of both treatments pooled. For example, suppose a non-inferiority trial is designed to allow for sample size re-estimation based on an interim blinded estimate of the within-group variance only. At some point during the study (e.g., after 50% of the patients have been treated and followed), without unmasking the treatment assignments, the within-group variance is estimated at interim, and an updated sample size is calculated using the usual power calculations for non-inferiority based on a risk difference. If the estimated variance reflects a larger than anticipated variance, the study very well may have been underpowered at the original sample size in a fixed design; in a study with SSR based only on the pooled variance, the total sample size would increase to maintain adequate power. At the end of the second stage, the hypothesis test is performed using data from both stages. This approach has close to the required power since sample size

recalculation using $\hat{\sigma}_1^2$ (the interim estimate of the within group variance) ensures that the expected sample size is equal to the sample size required given σ_0^2 (the original assumption of the within-group variance).

For a superiority trial, Wittes and Brittain¹² proposed estimating the within-treatment group variance at an interim time point during the study using a pooled estimate but treating the original samples size $n(\sigma_0^2)$ as a minimum sample size and continuing to take at least $n(\sigma_0^2)$ observations even if the updated sample size $n(\hat{\sigma}_1^2) < n(\sigma_0^2)$, where $n(\sigma_0^2)$ is the sample size based on σ_0^2 the original assumption of within-group variance, $n(\hat{\sigma}_1^2)$ is the proposed new sample size based on $\hat{\sigma}_1^2$, the revised assumption of the within-group variance based on interim data. Wittes et al¹³ have shown that this approach exhibits tighter control of the Type I error rate than the unrestricted one, where the initially calculated target sample size does not act as a lower bound for the recalculated sample size. Gould¹⁴, also focusing on superiority trials, proposed retaining the initial choice of sample size unless $n(\hat{\sigma}_1^2) > \gamma n(\sigma_0^2)$, where $\gamma > 1$, and restricting sample size to a maximum of $\varpi n(\sigma_0^2)$, where $\varpi > 1$, suggesting possible values $\gamma = 1.33$ and $\varpi = 2$ for the two constants. If $n(\hat{\sigma}_1^2)$ is particularly high, investigators may decide to terminate a study on the grounds that it is not feasible to recruit sufficient patients to achieve acceptable power.

In general, the Type I error rate resulting from a sample size increase based on blinded estimation of the variance may differ from the nominal level due to the random nature of

the total sample size (being a function of $\hat{\sigma}_1^2$), despite the variance being estimated based on a blinded interim analysis. While inflation of the Type I error rate is negligible with modestly large sample sizes, some authors propose modifications to the final test to better control the Type I error in the case of small samples; these include adjustments to the variance estimator^{15,16} to the number of degrees of freedom associated with the statistic¹⁷, or to the critical value¹⁸.

Advantages of blinded sample size recalculation include the ability to refine the estimate of power in the absence of an accurate estimate of the within-group variance prior to the study. The procedure may be carried out without a formal interim unblinded analysis since methods for calculating the sample variance without unmasking treatment codes are available^{19,20,21,22}.

2.2.2 Two-stage sample size re-estimation using the interim observed variance and the originally specified treatment difference in a conditional power calculation for a continuous endpoint

In a superiority clinical trial of a continuous endpoint where an interim estimate of the nuisance parameter (such as within-treatment variance for a continuous outcome) and the original protocol-specified difference between means drives the SSR process through conditional power calculations, methodologies are available^{4,8}. Formally, this sample size recalculation uses the “conditional power at θ_0 ,” defined as the probability of rejecting H_0 at the end of the study, conditional on the interim test statistic, under the

originally protocol-specified assumed difference θ_0 between treatment means, with $\sigma^2 = \hat{\sigma}^2$.

In this type of procedure, it is necessary to protect the Type I error rate due to the dependence of the sample size on z_1 , the first-stage test statistic. The method involving a “conditional error function”^{4,8} to adjust the rejection region for the final test statistic is one approach to dealing with the issue.

An early stopping rule for the first stage may further be desirable; for instance, futility might be declared when the sample size required to ensure a reasonable conditional power is deemed unfeasibly large.

An advantage of this procedure is that it uses information on the variance and additionally treats the observed data as fixed rather than random in the sample size recalculation (by conditional on z_1); i.e., instead of averaging across all possible values of z_1 , it fixes power for the observed z_1 .

Furthermore, whereas this approach involves powering the study at the originally specified treatment difference, it may be extended to allow for powering the study based on $\hat{\theta}$, the observed treatment difference at the interim (see Section 2.2.3), which affords $\hat{\theta}$ a more influential role in the sample size recalculation process.

2.2.3 Two-stage sample size re-estimation using the interim observed variance and treatment difference in a conditional power calculation for a continuous endpoint

Several authors^{6,8,9,23,24,25} have proposed for a superiority trial that $\hat{\theta}$ (the interim estimate of the observed difference between treatment means) be used in place of the originally specified difference between treatment means in an interim sample size recalculation procedure. This could be desirable if, at the interim, it appears that the magnitude of θ was too optimistic and that a smaller difference closer to $\hat{\theta}$ is more realistic, but still clinically relevant. Sample size recalculation could then involve recalculating the sample size using $\hat{\theta}$ and the interim estimate of the within-group variance in a conditional power calculation and then determining the sample size required to yield adequate, e.g., at least 80%, conditional power for a significant result at the end of the study. Such a sample size recalculation may result in an increase in sample size depending on the magnitude of treatment difference anticipated in the second stage and the variance estimate. Note that such a sample size re-estimation may or may not require (depending on the magnitude of the sample size increase) an adjustment of the test statistics or final critical value to maintain the Type I error at 0.05 level.

A stopping rule for futility based on conditional power could be added; e.g., increase the sample size to obtain conditional power of 80% if the conditional power is between 50 and 80%. Otherwise if the conditional power is between 10-50%, continue the trial as is, understanding it may be underpowered; if the conditional power is <10%, then stop for futility.

2.2.4 Two stage sample size re-estimation using the interim observed variance and treatment difference while also permitting early stopping to reject or accept H_0

In superiority trials, some recent methods extend the approaches in Section 2.2.3 beyond two stages by combining sample size recalculation using information concerning the observed treatment difference with methods for early stopping^{26,27}. For instance, Cui, Hung, and Wang⁹ incorporated conditional power sample size recalculation using the observed treatment difference at a single interim analysis in the context of a group sequential trial. They demonstrated that so long as the weights for the interim test statistics are fixed (as though there were no sample size recalculation), the remaining sample size can depend on the first stage treatment effect with no Type I error rate inflation.

A disadvantage of this approach is the potential inefficiency caused by fixing the weights when the sample size for the second stage is substantially increased (or decreased); e.g., giving identical weights to two stages when the sample size of one is twice that of the other. As Jennison and Turnbull²⁸ point out, this is the drawback of using a statistic that is not sufficient.

Other approaches include an extension to the p-value combination methodology^{29,30,31,32}, that has the advantage of flexibility, and one that uses a variance spending function^{34,33}, in which no explicit hypothesis testing (except for futility) is performed at interim stages, although if a large treatment difference is observed, the trial terminates for the final

analysis. It should also be noted that all of these procedures can bias the usual final estimator of treatment effect, though adjustments have been proposed^{30,34}.

The approaches have been explored by Jennison and Turnbull²⁸ and have been shown to be equivalent for the case of a single interim analysis at which no early stopping is permitted. They demonstrated that while sample size recalculation using the observed treatment difference substantially increases the power at the true treatment difference, the penalty is a substantial increase in expected sample size if the originally specified treatment difference is greater than the observed treatment difference at interim. Instead, they advocate instead designing a group sequential trial that is larger than minimally required and allowing for early stopping due to efficacy or futility, rather than designing a smaller trial and planning to extend it if necessary. More research is required to determine whether it is possible that careful design could make sample size recalculation procedures comparable to such a group sequential trial in terms of efficiency.

2.3 SSR Methods Used for a Binary Endpoint

2.3.1 Sample size estimation in fixed design for a binary endpoint

Before we review the sample size re-estimation literature for a binary endpoint in the section below, we review the sample size calculations for a fixed design first. In non-inferiority clinical trials, the null hypothesis for binary endpoint is usually specified as a non-zero difference or non-unity relative risk. Standard null hypothesis testing for superiority trials does not apply to non-inferiority trials, where the purpose is to demonstrate that the experimental treatment is not inferior to the control treatment by a

pre-specified clinically relevant margin (the non-inferiority margin). For a positive outcome (the larger values of the outcome indicate a more beneficial response to treatment), the hypotheses to test are:

$$H_0: p_1 - p_2 \geq s_0$$

$$H_a: p_1 - p_2 < s_0$$

Where $s_0 > 0$ is the pre-specified non-inferiority margin.

For a clinical trial with binary endpoints that involves two groups of size, N_1, N_2 , in the predetermined ratio $\theta = N_1/N_2$, and independent response variables $x_1 \sim Bi(N_1, p_1)$ and $x_2 \sim Bi(N_2, p_2)$, suppose that under the null hypothesis the difference of two binomial proportion is s_0 . The sample size calculation is based on the following test statistics:

$z_D = \hat{p}_1 - \hat{p}_2 - s_0$. For large N_1 , z_D is approximately normally distributed. Under the

null hypothesis, the variance of z_D is estimated by $\hat{v}_0 = \left[\tilde{p}_{1D} \tilde{q}_{1D} + \frac{\tilde{p}_{2D} \tilde{q}_{2D}}{\theta} \right] / N_1$. \tilde{p}_{1D} and

\tilde{p}_{2D} are estimates of p_1 and p_2 under the null hypothesis, and $\tilde{q}_{1D} = 1 - \tilde{p}_{1D}$, $\tilde{q}_{2D} = 1 -$

\tilde{p}_{2D} . The null hypothesis then is tested by $(\hat{p}_1 - \hat{p}_2 - s_0)^2 / \hat{v}_0$ with the chi-square

distribution on one degree of freedom. Blackwelder⁷ proposed to replace \tilde{p}_{1D} and \tilde{p}_{2D}

with the observed value \hat{p}_1 and \hat{p}_2 for hypothesis testing and the true values for sample

size calculation. Dunnett and Gent³⁵ suggested estimate \tilde{p}_{1D} and \tilde{p}_{2D} under the null

hypothesis restriction $\tilde{p}_{1D} - \tilde{p}_{2D} = s_0$ subject to the marginal totals remaining equal to

those observed. Farrington and Manning proposed⁶³ that \tilde{p}_{1D} and \tilde{p}_{2D} are taken to be the

restricted maximum likelihood estimators of p_1 and p_2 under the null hypothesis

restriction $\tilde{p}_{1D} - \tilde{p}_{2D} = s_0$. The Farrington and Manning methods were shown to have

stronger control of Type I error rate and reasonable power. The corresponding Farrington

and Manning sample size formulae are valid asymptotically, and were shown to be accurate even when the expected numbers of positive outcomes in each group are small.

The sample size formula was given as following:

$$n_T = \frac{(\phi^{-1}(\alpha)\sqrt{V_0} + \phi^{-1}(\beta)\sqrt{V_1})^2}{(p_1 - p_2 - s_0)^2}$$

$$V_0 = \tilde{p}_1(1 - \tilde{p}_1) + \tilde{p}_2(1 - \tilde{p}_2)$$

$$V_1 = p_1(1 - p_1) + p_2(1 - p_2)$$

where V_0 and V_1 are variances under H_0 and H_a respectively. \tilde{p}_1 and \tilde{p}_2 are Farrington and Manning's large sample approximations of p_1 and p_2 .

Similar considerations apply when the null hypothesis is of a specified relative risk R_0 .

For a positive outcome (the larger values of the outcome indicate a more beneficial response to treatment), the hypotheses to test are:

$$H_0: p_1/p_2 \geq R_0$$

$$H_a: p_1/p_2 < R_0$$

Where $R_0 > 0$ is the pre-specified non-inferiority margin.

In this case, and using the same notation as above, the sample size calculation is based on the following test statistics: $z_R = \hat{p}_1 - R_0\hat{p}_2$. Under the null hypothesis, the variance of z_R is estimated by $\hat{w}_0 = [\tilde{p}_{1R}\tilde{q}_{1R} + (R_0^2/\theta)\tilde{p}_{2R}\tilde{q}_{2R}]/N_1$. The null hypothesis may then be tested by referring the statistic:

$$(\hat{p}_1 - \frac{R_0}{\hat{p}_2})^2 / \hat{w}_0$$

to the chi-square distribution on one degree of freedom.

The asymptotic sample size formula for a test of the null hypothesis is

$$N_1 \geq \left\{ z_\alpha \sqrt{[\bar{p}_{1R}\bar{q}_{1R} + (R_0^2/\theta)\bar{p}_{2R}\bar{q}_{2R}] + z_\beta \sqrt{[p_1q_1 + (R_0^2/\theta)p_2q_2]} \right\}^2 / (p_1 - R_0p_2)^2$$

2.3.2 Two-stage sample size re-estimation using the blinded overall event rate for a binary endpoint

In principle, sample size calculation for binary endpoints faces the same problem as for normally distributed endpoints: The required sample size depends not only on the known values of significance level, power and clinically relevant difference but also on the variance of the endpoint. However, unlike for the normally distributed continuous endpoints, the mean and variance are not distinct parameters. In contrast to the normal case, the variance of the outcome is now directly related to the mean value of the outcome, i.e., overall event rate. As a further difference to normally distributed outcomes, the variance always lies between 0 and 1. Much less literature exists about sample size recalculation procedures for binary response variables as compared to normally distributed data in the superiority setting.

The procedure proposed by Gould¹⁴ makes use of the overall event rate (both treatments combined) observed in the first stage samples, where both the individual treatment allocation and the relative treatment effect remain unknown. Gould investigated the characteristics of this method for superiority trials by simulation studies for the situation

when the chi-square test is used to test the null hypothesis of equal proportions at the final stage.

Friede and Kieser^{36,37} presented exact computations of Type I error and power for the Gould¹⁴ approach, the contrasted Type I error rate to the one of the fixed sample size chi-square test for a wide range of parameter values. They concluded that the actual levels of the chi-square test for the fixed sample size design and the SSR design were very similar. Importantly, the excess in Type I error rate that occurs under the sample size recalculation using Gould approach is not higher than for the fixed sample size design. Shih and Zhao³⁸ described a design where by use of an artificial ‘dummy’ stratification and unbalanced randomization within strata, the event rate in each treatment group can be estimated during the ongoing trial without knowing the treatment codes. This approach is problematic for several reasons. Although the treatment allocation of the patients remains masked, trial personnel might become aware of the treatment difference before the study is finished. This may introduce a bias into selection and/or assessment of future patients unless the same precautions as for interim analyses are taken. On the other side, implementing an IDMC makes sample size adjustment logistically expedient.

Friede, Mitchell and Muller-Velten³⁹ extended the methodology used in superiority trials (blinded sample size recalculation) to non-inferiority trials with binary endpoints. They compared the performance of Type I error rate and power between fixed-size designs and designs with sample size re-estimation. It showed that the sample size re-estimation was

effective in correcting sample size and power of the tests when misspecification of design parameters, i.e., event rate in control arm, occurred with the fixed-size design.

2.3.3 Two-stage sample size re-estimation using the unblinded pooled event rate for a binary endpoint

Herson and Wittes⁴⁰ proposed a procedure for a superiority trial where the event rate of the control group is estimated mid-course based on observed data and where the sample size is adjusted assuming that the true event rate in the control group is identical to the estimated value. There is no inspection of the experimental group event rate.

Friede and Keiser⁴¹ extended the methodology used in superiority trials (t-test for continuous endpoints using both blinded pooled and unblinded variance estimator for sample size recalculation) to non-inferiority trials with a binary endpoint^{42,43}. The exact computations of Type I error and power were presented, and the actual level was contrasted to the one of the fixed sample size chi-square test. They concluded that blinded sample size re-estimation mitigates the effect of initial misspecification of the overall response rate.

Wang, Keller, and Lan⁴⁴ proposed an interim sample size re-estimation procedure via conditional power using unblinded event rate for binary outcome in clinical trials with a non-inferiority objective. The essence of the procedure is that if a trial is extended, the α level will be inflated and that, on the contrary, if a trial is stopped early to claim futility, the α level will be reduced; if inflation does not exceed deflation, then the α level will be maintained.

Gould⁴⁵ compared the blinded and unblinded sample size re-estimation methods for superiority trials and showed that unblinded sample size recalculation methods that only use the event rates to estimate the variance, but do not explicitly use the observed between-group difference can be replaced by blinded methods that protect Type I error rate and provide adequate power, at least when the number of observations on which the interim assessment is made is not too small.

2.3.4 Adapting non-inferiority margin with the underlying event rate in the control arm for a binary endpoint

Phillips⁴⁶ modified the standard test of non-inferiority proposed by Blackwelder⁷ where the non-inferiority margin is fixed, to the test that allows the inferiority margin to vary with the underlying event rates at the end of the study. The derived statistical test was similar to the FDA's Points-to-Consider procedure⁴⁷, which is intended to apply to anti-infective medications. Phillips modified the simple test of non-inferiority to allow non-inferiority margin (Δ) to adapt to the comparator rate by incorporating the comparator rate in the expression for non-inferiority margin Δ , where Δ is a linear function of the true comparator rate, $\Delta = \gamma + b\pi_c$. Then γ and b were chosen to conform to the Point-to-Consider non-inferiority margin. At the final analysis stage, Δ is calculated by replacing π_c with its estimate based on the final observed data. The proposed test is shown to preserve the power over a range of response rates in trials with positive endpoint. However, the author did not discuss the impact on the Type I error rate, nor did he discuss the performance of the procedure applied to a trial with negative endpoints.

2.4 Review of Related Topics

2.4.1 Size of first stage samples

A number of suggestions have been made for the choice of the first stage sample size (n_1) in study with sample size re-estimation, especially for superiority trials. For the two-stage procedure of Stein¹⁵, Seelbinder⁴⁸ proposed a rule for determining n_1 , which is on the basis of the optimality criterion of minimizing the maximum difference between the expectation of the recalculated sample size and the sample size required (if s^2 were known) over a range of s^2 . Moshman⁴⁹ refined this strategy by bounding the probability of a huge sample size by a pre-specified value. For the sample size re-estimation design, Wittes and Brittain¹² choose in their simulations the first-stage sample size as half of the initially calculated size (n_0). The approach of Sandvik et al.⁵⁰ targets selecting n_1 as high as possible while at the same time bounding the probability of recruiting more patients than required if s^2 were known. This method assumes that the data available for the initial sample size calculation are a random sample of the study population to be investigated in the planned study. If this assumption is fulfilled, the sample size calculation for a fixed sample size design without sample size re-estimation can be performed so that the desired power is reached with pre-specified probability^{51,18}. On the other hand, the same result holds true (independent of the validity of the assumption) for designs with a sample size re-estimation⁵². Hence, in the situation considered by Sandvik et al.⁵⁰, the two-stage sample size re-estimation study does not lead to a benefit for the sample size calculation as compared to the fixed sample size design. Singer⁵³ pointed out that the recruitment rate and the follow-up period have to be taken into account when the

sample size is chosen according to the recommendation of Sandvik et al.⁵⁰. This is necessary to avoid recruiting more patients than are actually required, as the enrolment is usually not stopped when n_1 patients have been included. A further proposal was made by Denne and Jennison¹⁷ who considered the ratio of the expectation of the recalculated sample size to the required sample size if s^2 were known. Their strategy results in a value of this ratio that is close to the minimum possible value.

2.4.2 Choice of non-inferiority margin

One of the challenges in non-inferiority trials compared with superiority trials is the choices of the margin. It can be quite difficult to specify an appropriate non-inferiority margin. There are two basic approaches, both of which have serious drawbacks. One approach is to specify the non-inferiority margin on the basis of a clinical notion of a minimally important effect. However, this is clearly subjective, and it is possible with this approach to set the non-inferiority margin to be greater than the effect of the active control compared to a placebo, which could lead to harmful treatments fitting within the definition of non-inferiority.

To avoid this, the non-inferiority margin is often chosen with reference to the effect of the active control in historical placebo-controlled trials. When the non-inferiority margin is chosen in this way, there is some basis on which to claim that a positive non-inferiority trial implies that the new treatment is superior to placebo. However, this claim requires an assumption that the effect of the active control in the current trial is similar to its effect in the historical trials. This assumption can be undermined by differences with respect to

design features (eg., the patient population, dosage regimen of the active control, end-point definition, or concomitant therapies), or by an inconsistency in the effect of the active controls among the historical placebo-controlled trials (beyond that expected by random chance). For this reason, the non-inferiority margin usually includes some type of buffer. Rather than basing it on the full predicted effect of the active control, it is often based on the lower bound of a confidence interval for that effect (accounting for within-trial and trial-to-trial variability), or on preservation of a specific fraction (e.g., 50%) of the effect of the active control.

From a statistical perspective, the margin should be, at the very least, no larger than the worst limit of 95% confidence interval (CI) of standard treatment effect relative to placebo^{54,55}, but it could be smaller to assure that the new treatment has greater than minimal efficacy⁵⁶. One proposal for selecting the margin is to take one-half of the magnitude of the worst limit of this CI—the so-called “50% rule” or “95-95 method” recommended by the FDA^{3,55}. This conservative margin, however, often results in a high “false-negative” rate (Type II error; i.e., low power to demonstrate non-inferiority). In general, the objective should be to limit “false-positive” (Type I) errors by avoiding too liberal a margin and “false-negative” (Type II) errors by avoiding too conservative a margin with respect to a claim of efficacy. The margin is generally set in terms of an absolute or relative difference, the latter being favored over the former given its greater trial-to-trial stability⁵⁶.

Our hope is that this chapter will increase the understanding of all the aspects of sample size re-estimation in non-inferiority clinical trials; it includes the choice of first stage sample size and non-inferiority margin, a variety of available procedures for both continuous and binary endpoints in designs with sample size recalculation in both blinded and unblinded fashion, as well as how those procedures compare to one another and to other designs. Finally, we hope that this chapter will also motivate the development and application of sample size re-estimation methods to be used in non-inferiority clinical trials.

Chapter 3: NORMALLY DISTRIBUTED ENDPOINT

3.1 Introduction and Background

In study design, it is crucial to plan the sample size to be sufficient to ensure a desirable level of power. This requires specification of the alternative hypothesis under which to fix the power, commonly the difference between treatment means that represents the minimum clinically significant difference. Sometimes, however, a difference larger than the minimum clinically significant difference is selected due to optimism surrounding a new therapy, or as a strategy to reduce the sample size and, thereby, costs. This strategy runs the risk of failing to achieve statistical significance due to under-powering the study if this optimism is misplaced and the true treatment difference lies between the optimistic value and the minimum clinically significant difference.

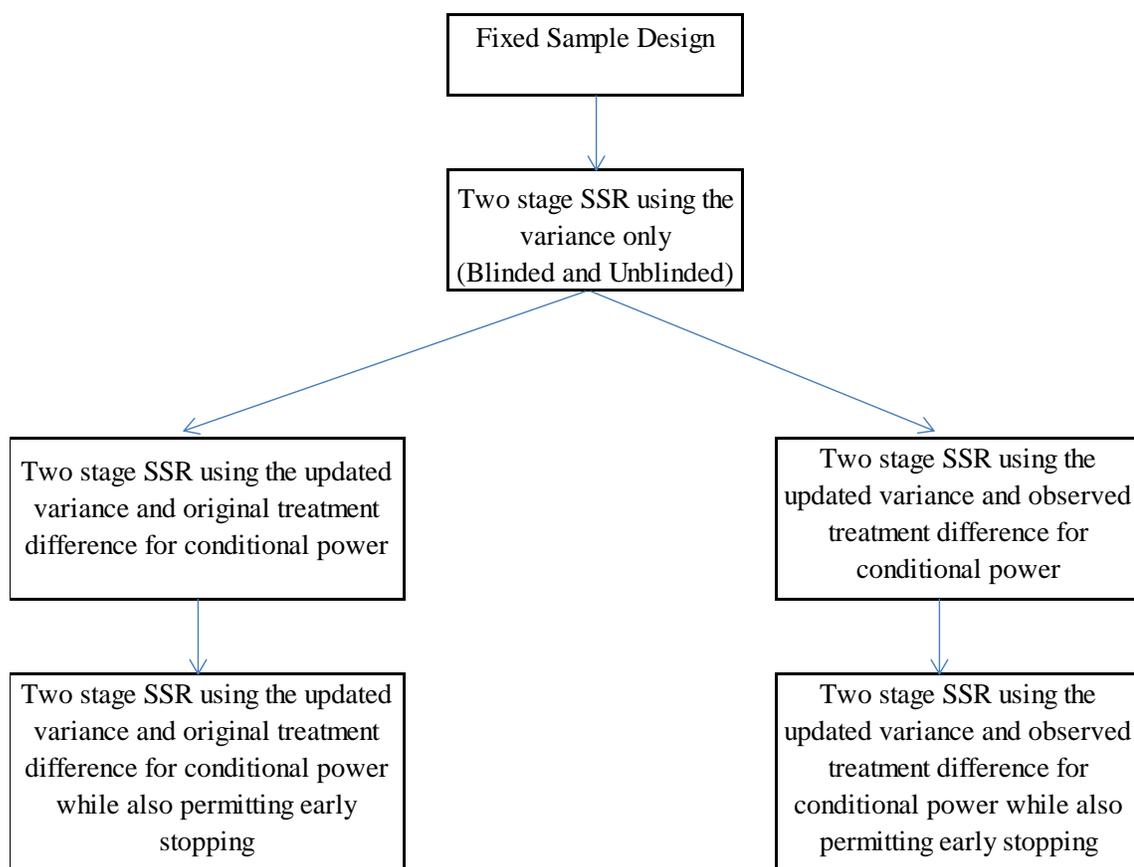
Additionally, for a continuous normally distributed outcome, if an accurate value for the variance is not available, the study may not be precisely powered for a test of the difference in a normally distributed outcome where power depends on the within-group variance (σ^2), in addition to the difference between treatment means. A pilot study may be undertaken to estimate the variance, but this may be problematic due to sampling variability if the pilot is small, or due to bias if the pilot population differs from the study population.

Historically, studies have taken a “fixed-sample” approach, in which the sample size is determined at the outset, based on an assumption of the variance and difference between treatment means obtained from previously collected data. Sample size re-estimation

(SSR) has been proposed to overcome the sensitivity of the power to misspecification of design parameters by allowing the sample size to be modified through a re-specification of these parameters during the study. The rationale is that estimates from within the study are unbiased because they are based on the same study design and population at a concurrent point in time.

Current approaches to SSR for non-inferiority trials focus on estimating the variance (blinded or unblinded) at the interim and updating the sample size based solely on this estimated variance^{41,42}. The SSR using both sample variance and the observed treatment difference at interim in conditional power calculation is used in superiority trials⁴. We extend the methodology to non-inferiority trials, quantify the effect on Type I error rate, and propose controlling it by modifying the critical value and/or stopping the trial at the interim for futility. Figure 1 outlines the progression of the approaches to sample size re-estimation in normally distributed outcome that will be discussed in this chapter.

Figure 1: The progression of approaches to two-stage SSR in normally distributed outcome



3.2 Computational Methods

Simulation studies were carried out to evaluate the performance of the proposed procedures over a range of circumstances likely to be met in most clinical trials, including when the null hypothesis really was true (to assess Type I error rate) and when it really was false (to assess power). All programs for simulations were written in SAS

version 9.3 (SAS Institute, 2004). Under each scenario, 100,000 and 10,000 trials were simulated for Type I error rate and power, respectively.

The parameters required for the specification of model and test ran through all combinations of an extended range of values in simulations are shown in Table 1.

Here the parameter values θ_a and σ_a^2 are the values of the difference between treatment means and the variance of the outcome (assumed to be the same in each group) used for the initial sample size calculation, and the actual values θ and σ^2 are used in the true model, from which the distribution of possible test outcomes is taken.

As a practical consideration, the second stage sample size was restricted so that it could not exceed double the originally designed sample size.

In addition, calculations were made for optimal allocation (i.e., sample sizes are equal in each treatment group), and the interim analysis was conducted at $\frac{1}{2}$ the originally calculated sample size.

Table 1: Parameter values used for the computations of Type I error rate and power for non-inferiority tests with continuous outcome

		Type I error	Power
Nominal significance level	α	0.025	0.025
Target power of the test	$1 - \beta$	0.8	0.8
Non-inferiority margin	Δ	0.2	0.2
Assumed difference between treatment means $\mu_C - \mu_T$ at protocol design stage	θ^a	0	0

Assumed common within-group variance at protocol design stage	σ_a^2	1	1
Actual difference between treatment means $\mu_C - \mu_T$	θ	0.02, 0.05, 0.07, 0.1	0.02, 0.05, 0.07, 0.1
Actual common within-group variance	σ^2	1, 1.2, 1.5, 1.7, 2.0	1, 1.2, 1.5, 1.7, 2.0

3.3 Fixed Sample Size Design

Consider a randomized non-inferiority clinical trial that compares the effect of two independent treatments on a disease based on a normal outcome (x). Let X_{Ti} and X_{Ci} , $i = 1, 2, \dots, n$, (we assume equal sample sizes in each group) denote the responses of patient i in the experimental treatment group and active control group, respectively.

Suppose the responses of patients receiving treatment T are normally distributed with variance σ^2 and mean μ_T , which we write $X_{Ti} \sim N(\mu_T, \sigma^2)$, $i = 1, 2, \dots, n$. Likewise, suppose $X_{Ci} \sim N(\mu_C, \sigma^2)$, $i = 1, 2, \dots, n$, and all observations are independent.

For a positive outcome (without loss of generality, we assume the larger values of the outcome indicate a more beneficial response to treatment), the hypotheses to test are:

$$H_0: \mu_C - \mu_T \geq \Delta$$

$$H_a: \mu_C - \mu_T < \Delta$$

Where $\Delta > 0$ is the pre-specified non-inferiority margin.

With Type I error probability α and power $1 - \beta$ at $\mu_C - \mu_T = \theta$, if n patients are allocated to each treatment group, the standardized statistic

$$Z = \frac{1}{\sqrt{(2n\sigma^2)}} \left(\sum_{i=1}^n X_{Ci} - \sum_{i=1}^n X_{Ti} - \Delta \right)$$

$$\sim N((\theta - \Delta)\sqrt{\{n/(2\sigma^2)\}}, 1)$$

Thus $Z \sim N(0,1)$ under H_0 and the one-sided test with Type I error probability α rejects H_0 if $Z > \Phi^{-1}(1 - \alpha)$, where Φ denotes the standard normal cumulative distribution function (cdf). To satisfy the power requirement, we also need

$$\Pr\{Z > \Phi^{-1}(1 - \alpha)\} = 1 - \beta$$

when $Z \sim N((\theta - \Delta)\sqrt{\{n/(2\sigma^2)\}}, 1)$.

Thus the necessary sample size would be

$$n(\alpha, \beta, \theta, \Delta, \sigma^2) = \{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\}^2 2\sigma^2 / (\theta - \Delta)^2$$

patients per treatment group.

In a fixed-sample design, an estimate of the population variance (σ^2) is formed only once for use in the sample size calculation, and this is prior to any data collection within the planned study. Advantages of the fixed-sample approach include its straightforward nature and well-known properties, its ease of implementation, clear planning implications with respect to the sample size, maintenance of the Type I error, and the maintenance of masking of the treatment assignments during the course of the study.

Example

Assume a clinical trial involves an active control with study objective of non-inferiority.

The endpoint of the trial is a normally distributed response variable. At the planning

stage, we assume that the common within-group variance $\sigma_0^2 = 1$ and expected response for both treatment and control arms are equal, that is $\theta=0$. The non-inferiority margin of $\Delta= 0.2$ was pre-specified and deemed clinically acceptable.

The hypotheses to be tested are

$$H_0: \mu_C - \mu_T \geq 0.2$$

$$H_a: \mu_C - \mu_T < 0.2$$

The trial would be declared as positive at the end of the study at a one-sided 0.025 level of significance if the one-sided upper 95% CI of $\mu_C - \mu_T$ was < 0.2 . In order to achieve a power of 80% with one-sided $\alpha = 0.025$ to demonstrate non-inferiority, a sample size of 393 patients per group would be required.

Since power depends on the true variance and the true difference between treatment means, the misspecification of any of those design parameters will seriously affect the power, and study may not be able to reach a definitive conclusion.

The simulation results in Table A3.1 and Table A3.2 of the Appendix show how the Type I error rate and power is affected for an extended range of values of combination of misspecified variance (σ^2) and treatment difference (θ). Table 2 below highlights some of the findings there and indicates that inappropriate assumptions about any of these factors can lead to an underpowered trial; however, the Type I error rate is not affected by those erroneous assumptions.

Table 2: Consequences of overestimating treatment difference θ and/or underestimating variance σ^2 with continuous outcomes ($\alpha=0.025$, planned Power=80%)

	Type I error Rate	Power
No underestimation of either parameter	0.02508	80%
Underestimate σ^2 alone by 50%	0.02508	46%
Overestimate θ alone by 50%	0.02508	29%
Overestimate θ AND underestimate σ by 50%	0.02508	16%

3.4 Two-stage SSR using only the observed variance and unconditional power calculation

A common difficulty in carrying sample size calculations arises from the lack of an accurate estimate of the true variance of the normally distributed outcome. In order to overcome the vulnerability of misspecification of variance, one might adopt the methods of two-stage sample size re-estimation using the interim sample variance only (blinded or unblinded), which would, on average, yield the “correct” sample size and produce a test with the required power. The methods in this section use the estimate of the variance only (blinded or unblinded), but the observed treatment difference, from the first-stage (interim) to re-estimate the within-group variance and recalculate the overall target final sample size using unconditional power formulas (i.e., the same power formula as used to initially design the study).

Example (continued)

Suppose the non-inferiority study in the above example is instead designed to allow for

two-stage sample size re-estimation based on the variance. After 50% of the patients have been treated and followed, the within group variance is estimated (in a blinded or unblinded manner, as will be discussed below) from observations collected to that point, and an updated sample size is calculated at 80% power. Let s_1^2 indicate the interim estimate of the true within-group variance. Suppose $s_1^2 = 1.5$, then the required sample size is 888 patients per group. Since this estimate reflects a larger than anticipated variance, the study may have been underpowered at the original sample size. At the end of the second stage, the non-inferiority hypothesis test is performed using data from both stages. As we will see, this approach has close to the required power since sample size re-estimation using s_1^2 ensures that the expected sample size is equal to the sample size required given σ^2 (See section 3.4.3).

3.4.1 Sample size re-estimation based on blinded estimate of the variance

According to the proposal of Zucker¹⁶ for superiority trials, the population variance σ^2 is estimated by the total sample variance of the first stage samples ignoring the treatment group indicator. Let us call this s_1^2 . The rationale is that the total variance is similar to the within-group variance as long as the treatment effect is not too large; this is even more justified for the non-inferiority trials, where it is usually assumed the distribution of the normally distributed outcome is the same in each group. Then sample size is recalculated by employing this estimate instead of σ^2 in the usual unconditional power calculations. Specifically, the total variance is estimated by

$$s_{total}^2 = \frac{1}{n_1 - 1} \sum_{j=1}^2 \sum_{k=1}^{n_{ij}} (X_{1jk} - \bar{X}_1)^2$$

Table A3.3a and Table A3.4a in Appendix show the mean and standard deviation of the resulting final sample sizes for various true variance based on blinded variance estimate. As would be expected, the expected sample size is dependent on the true variance and treatment difference. The larger the true variance is than the original assumptions, the more samples are needed. Table A3.3a also shows that the Type I error rates are close to the nominal level in the presence of a blinded estimate of variance.

Table A3.4a gives the expected power for various true values of $\mu_C - \mu_T$ and variance. The blinded sample size re-estimation performs very well if only the variance was mis-specified. However, when the treatment difference was under-estimated as well, the procedure is not able to maintain the desired power.

3.4.2 Sample size re-estimation based on unblinded estimate of the variance

In the original proposal of Wittes and Brittain¹² for a superiority trial, sample size adjustment is on the basis of the pooled variance estimate based on unblinded data.

Specifically, the pooled variance is estimated by

$$s_{pooled}^2 = \frac{\sum_{i=1}^2 (n_{i1} - 1) s_{i1}^2}{\sum_{i=1}^2 (n_{i1} - 1)}$$

where n_{i1} and s_{i1}^2 are the first stage sample size and sample variance, respectively, for treatment group i . Then the pooled variance is assumed to be the within-group variance, and the sample size is re-calculated using un-conditional formulas.

Similar results as shown for the blinded variance estimate approach were found for the unblinded sample size re-estimation procedure. For further details, see Table A3.3b and Table A3.4b in the Appendix A and Section 3.4.3 below.

3.4.3 Comparison of the procedures

Table 3 displays the Type I error rate for both blinded and unblinded procedures at different true variance. Both procedures control the Type I error rate at the nominal level.

There is no reduction in overall sample size when conducting the SSR unblinded as compared to blinded; i.e., The average sample size increases were virtually identical.

Table 3: Type I error rates at different true variances for two-stage SSR using only the observed variance and unconditional power calculation in trials with a continuous outcome

	$\sigma^2 = 0.9$	$\sigma^2 = 0.8$	$\sigma^2 = 0.7$	$\sigma^2 = 0.6$	$\sigma^2 = 0.5$
Blinded	0.02424	0.02402	0.02378	0.02312	0.02240
Unblinded	0.02438	0.02412	0.02356	0.02312	0.02240
	$\sigma^2 = 1.0$	$\sigma^2 = 1.2$	$\sigma^2 = 1.5$	$\sigma^2 = 1.7$	$\sigma^2 = 2.0$
Blinded	0.02441	0.02557	0.02545	0.02504	0.02454
Unblinded	0.02445	0.02557	0.02542	0.02539	0.02438

Table 4 displays the power for both blinded and unblinded procedures at different true variance. Power is greatly stabilized by both procedures. There is no noticeable difference between two procedures.

Table 4: Power at different true variances for two-stage SSR using only the observed variance and unconditional power calculation in trials with a continuous outcome

	$\sigma^2 = 0.9$	$\sigma^2 = 0.8$	$\sigma^2 = 0.7$	$\sigma^2 = 0.6$	$\sigma^2 = 0.5$
Blinded	0.8794	0.9412	0.9798	0.9963	0.9996
Unblinded	0.8784	0.9412	0.9796	0.9962	0.9996
	$\sigma^2 = 1.0$	$\sigma^2 = 1.2$	$\sigma^2 = 1.5$	$\sigma^2 = 1.7$	$\sigma^2 = 2.0$
Blinded	0.8107	0.8033	0.8052	0.8023	0.7991
Unblinded	0.8109	0.8034	0.8050	0.8030	0.7981

In summary, re-estimation of the sample size for normally distributed data on the basis of interim estimate of the within-group variance using either blinded or unblinded estimates preserves the Type I error rate well, and maintains the power to the nominal rate if only the variance was under-estimated. Both procedures are not suitable if the treatment difference was also mis-specified; the power falls substantially below the planned 80% when the treatment difference was under-estimated. If one wants to base SSR only on the variance, we suggest the use of blinded procedure for three reasons: 1) the treatment effect is expected to be quite small in non-inferiority trials and, therefore, under the alternative hypothesis, the within-group variance is quite close to the total variance estimated by one sample variance; 2) Control of the Type I error rate is reasonable; 3) Application of the procedure with the one-sample variance is very simple.

3.5 Incorporating conditional power into two-stage SSR in a superiority setting

Several authors have proposed procedures based on conditional power for superiority trials. Proschan and Hunsberger⁵⁷ proposed a method for designed extension of a study. Their motivation was to reduce the occurrence of studies that come close to, but fail to achieve, statistical significance. Bauer and Kohne²⁴ proposed a general two-stage procedure that combines the separate p-values from the two stages into a global test using Fisher's product criterion. The sample size for the second stage of their procedure can be based upon data from the first stage without inflating the Type I error rate. Lan²³ proposed a two-stage procedure based on the B-value method of Lan and Wittes²³, in which the size of the second stage is fixed so as to ensure sufficient conditional power at $\theta = \hat{\theta}$. Lan showed via simulation that incorporation of a rule for early stopping to accept H_0 ensures that the Type I error rate is controlled over a range of design parameters. Herson⁵⁸ proposed a two-stage procedure also based on the B-value method, which does not strictly condition on the observed data, but uses an estimate of the variance in a conditional power-like formula. Denne⁵⁹ applied conditional power based sample size re-calculation in the context of a two-group error-spending sequential test. We will extend Denne's method to non-inferiority trials.

3.5.1 Calculating conditional power in a non-inferiority setting

Here we will derive the conditional power function for the following non-inferiority null hypothesis for continuous outcome

$$H_0: \mu_C - \mu_T \leq \Delta \quad \text{vs.} \quad H_a: \mu_C - \mu_T > \Delta$$

We desire to test the above null hypothesis at an overall one-sided significance level α .

We shall assume the two groups have the same variance σ^2 .

Suppose that at the interim stage, after n_1 observations per group have been treated and

followed, we compute $z_1 = \sqrt{\frac{n_1}{2\hat{\sigma}_1^2}}(\bar{X}_{C1} - \bar{X}_{T1})$ where $\hat{\sigma}_1^2$ is the estimate of the pooled

variance based on the interim data, \bar{X}_{C1} and \bar{X}_{T1} are the means of the control group and experimental group at the interim, respectively.

We now derive the formula for conditional power. Let $CP_\delta(n_2, z_\alpha | z_1)$ denote the

conditional probability that Z based on $n = n_1 + n_2$ observations per group exceeds z_α ,

given that $Z_1 = z_1$, and given that $(\mu_T - \mu_C)/\sigma = \delta$. At the final stage, the test statistic is

$$Z = \sqrt{\frac{n}{2}} \left(\frac{\bar{X}_C - \bar{X}_T - \Delta}{\sigma} \right) = \sqrt{\frac{n_1 + n_2}{2\sigma^2}} \left(\frac{n_1(\bar{X}_{C1} - \bar{X}_{T1}) + n_2(\bar{X}_{C2} - \bar{X}_{T2}) - (n_1 + n_2)\Delta}{n_1 + n_2} \right)$$

where \bar{X}_{C1} , \bar{X}_{C2} and \bar{X}_C , are the means of the first stage, second stage and final stage

(first and second stages combined) for the control group, respectively. Likewise \bar{X}_{T1} ,

\bar{X}_{T2} and \bar{X}_T , are the means of the first stage, second stage and final stage (first and

second stage combined) for the experimental group, respectively.

Given this formula for Z , we have $CP_\delta(n_2, z_\alpha | z_1) = \Pr(Z > z_\alpha | Z_1 = z_1, \delta)$

$$\begin{aligned}
&= \Pr\left[\frac{n_1(\bar{X}_{C1} - \bar{X}_{T1}) + n_2(\bar{X}_{C2} - \bar{X}_{T2}) - (n_1 + n_2)\Delta}{\sqrt{2\sigma^2(n_1 + n_2)}} > z_\alpha \mid Z_1 = z_1, \delta\right] \\
&= \Pr[n_1(\bar{X}_{C1} - \bar{X}_{T1}) + n_2(\bar{X}_{C2} - \bar{X}_{T2}) - (n_1 + n_2)\Delta > z_\alpha \sqrt{2\sigma^2(n_1 + n_2)} \mid Z_1 = z_1, \delta] \\
&= \Pr\left[\frac{n_2(\bar{X}_{C2} - \bar{X}_{T2}) - n_2\delta\sigma}{\sqrt{2n_2\hat{\sigma}_1^2}} > \frac{z_\alpha \sqrt{2\sigma^2(n_1 + n_2)} - n_1(\bar{X}_{C1} - \bar{X}_{T1}) + (n_1 + n_2)\Delta - n_2\delta\sigma}{\sqrt{2n_2\hat{\sigma}_1^2}} \mid Z_1 = z_1, \delta\right] \\
&= \Pr\left[\frac{n_2(\bar{X}_{C2} - \bar{X}_{T2}) - n_2\delta\sigma}{\sqrt{2n_2\hat{\sigma}_1^2}} > \frac{z_\alpha \sqrt{2\sigma^2(n_1 + n_2)} - z_1 \sqrt{2n_1\hat{\sigma}_1^2} + (n_1 + n_2)\Delta - n_2\delta\sigma}{\sqrt{2n_2\hat{\sigma}_1^2}} \mid \delta\right]
\end{aligned}$$

Treating $\hat{\sigma} = \sigma$ (i.e., calculating conditional power under the assumption that the interim observed standard deviation is the true standard deviation), we get

$$CP_\delta(n_2, z_\alpha \mid z_1) = 1 - \Phi\left[\frac{z_\alpha \sqrt{2(n_1 + n_2)} - z_1 \sqrt{2n_1} + (n_1 + n_2)\delta_\Delta - n_2\delta}{\sqrt{2n_2}}\right]$$

where $\delta_\Delta = \Delta/\sigma$.

3.5.2 Type I error Rate (α) in the non-inferiority setting

The Type 1 error rate is $\int_{-\infty}^{+\infty} CP_0(n_2, z_\alpha \mid z_1)\Phi(z_1)dz_1$, where $\Phi(z_1)$ is the standard normal density function, and $CP_0(n_2, z_\alpha \mid z_1)$ is the conditional Type I error when z_α is used as the final critical value. The following calculation shows that the choice of n_2 can make the overall Type I error rate for the study as high as $\alpha_{\max} = \alpha + \exp(-z_\alpha^2/2)/4$, where α is the original planned nominal significance level.

Under the null hypothesis that $\mu_C - \mu_T \leq \Delta$,

$$CP_0(n_2, z_\alpha | z_1) = \Pr(z > z_\alpha | z_1, \delta = \delta_\Delta) \equiv A(z_1).$$

$$\text{It follows that } A(z_1) = 1 - \Phi\left[\frac{z_\alpha \sqrt{n_1 + n_2} - z_1 \sqrt{n_1} + n_1 \delta_\Delta / \sqrt{2}}{\sqrt{n_2}}\right]$$

Let $R = n_2 / n_1$, then

$$\begin{aligned} A(z_1) &= 1 - \Phi\left[\frac{z_\alpha \sqrt{Rn_1 + n_1} - z_1 \sqrt{n_1} + n_1 \delta_\Delta / \sqrt{2}}{\sqrt{Rn_1}}\right] \\ &= 1 - \Phi\left[(z_\alpha \sqrt{R+1} - z_1 + \sqrt{n_1/2} \delta_\Delta) / \sqrt{R}\right] = 1 - \Phi[f(R); z_1, z_\alpha] \end{aligned}$$

where $f(R) = (z_\alpha \sqrt{R+1} - z_1 + \sqrt{n_1/2} \delta_\Delta) / \sqrt{R}$. We first find the value of R that minimizes $f(R)$ over the range of 0 to z_α , as this will maximize $A(z_1)$ over the range of 0 to z_α , therefore, the maximum value of the Type I error rate will be derived.

$$f'(R) = \frac{\frac{z_\alpha \sqrt{R}}{2\sqrt{1+R}} - \frac{z_\alpha \sqrt{1+R}}{2\sqrt{R}}}{R} + \frac{z_1 - \sqrt{n_1/2} \delta_\Delta}{2R\sqrt{R}} = 0$$

$$\Rightarrow \frac{z_\alpha \sqrt{R}}{\sqrt{1+R}} - \frac{z_\alpha \sqrt{1+R}}{\sqrt{R}} + \frac{z_1 - \sqrt{n_1/2} \delta_\Delta}{\sqrt{R}} = 0$$

$$\Rightarrow \frac{z_\alpha \sqrt{R}}{\sqrt{1+R}} = \frac{z_\alpha \sqrt{1+R} - z_1 + \sqrt{n_1/2} \delta_\Delta}{\sqrt{R}}$$

$$\Rightarrow z_\alpha R = z_\alpha(1+R) - (z_1 - \sqrt{n_1/2\delta_\Lambda})\sqrt{1+R}$$

$$\Rightarrow z_\alpha = (z_1 - \sqrt{n_1/2\delta_\Lambda})\sqrt{1+R}$$

$$\Rightarrow R = \frac{z_\alpha^2}{(z_1 - \sqrt{n_1/2\delta_\Lambda})^2} - 1$$

$$f(R) \text{ is minimized (and } A(z_1) \text{ is maximized) when } R = \frac{z_\alpha^2}{(z_1 - \sqrt{n_1/2\delta_\Lambda})^2} - 1$$

$$\text{For this value of } R, CP_0(n_2, z_\alpha | z_1) = 1 - \Phi\left[\frac{z_\alpha \frac{z_\alpha}{z_1 - \sqrt{n_1/2\delta_\Lambda}} - (z_1 - \sqrt{n_1/2\delta_\Lambda})}{\sqrt{\frac{z_\alpha^2 - (z_1 - \sqrt{n_1/2\delta_\Lambda})^2}{(z_1 - \sqrt{n_1/2\delta_\Lambda})^2}}}\right]$$

$$= 1 - \Phi(\sqrt{z_\alpha^2 - (z_1 - \sqrt{n_1/2\delta_\Lambda})^2})$$

The Type I error rate is $\alpha = \int_{-\infty}^{+\infty} A(z_1)\Phi(z_1)dz_1$, where $n_2 = n_2(z_1)$.

Clearly,

- If $z_1 > z_\alpha$, $n_2 = 0$ is the maximizer, making $A(z_1) = CP_0 = 1$.
- If $z_1 < 0$, $f'(R) < 0$ and $f(R) \rightarrow z_\alpha$ as $f(R) \rightarrow \infty$.
- If $0 \leq z_1 \leq z_\alpha$, $f(R)$ is minimized when $R = \frac{z_\alpha^2}{(z_1 - \sqrt{n_1/2\delta_\Lambda})^2} - 1$

Thus the maximum value of the overall Type I error rate is

$$\int_{-\infty}^0 \alpha \Phi(z_1) dz_1 + \int_0^{z_\alpha} 1 - \Phi(\sqrt{z_\alpha^2 - (z_1 - \sqrt{n_1/2}\delta_\Delta)^2}) \Phi(z_1) dz_1 + \int_{z_\alpha}^{+\infty} \Phi(z_1) dz_1 \quad (\text{a})$$

We can simplify the (a) as follows. Let U and V be independently and identically distributed standard normal variables. Then $U^2 + V^2$ has a chi-squared distribution with 2 degree of freedom, which is exponential with parameter 1/2. It follows that

$$\begin{aligned} \exp(-z_\alpha^2/2) &= \Pr(U^2 + V^2 > z_\alpha^2) \\ &= \Pr(|U| > z_\alpha) + \int_{-z_\alpha}^{+z_\alpha} \Pr(|V| > \sqrt{z_\alpha^2 - u^2}) \Phi(u) du \\ &= 2\alpha + 4 \int_0^{z_\alpha} [1 - \Phi(\sqrt{z_\alpha^2 - u^2})] \Phi(u) du \end{aligned}$$

Using this fact we can simplify expression (a) to

$$\alpha_{\max} = \alpha + \exp(-z_\alpha^2/2)/4$$

The degree of Type I error rate inflation is surprisingly high. For a one-tailed $\alpha = 0.05$ test, suppose we use $z_\alpha = 1.645$ as the final critical value used to declare non-inferiority.

When we evaluate α_{\max} for this value of z_α , we get $\alpha_{\max} = 0.1146$. The actual Type I error rate is more than twice the originally planned 0.05. This more than doubling of the Type I error rate can be avoided by using a re-calculated critical value c value (derived in section 3.5.3), potentially larger than the original z_α , and agreeing not to continue the study unless the conditional power greater than a certain pre-specified cutoff. (e.g., 50%)

3.5.3 Determining additional sample size (n_2) and critical value adjustment required to maintain nominal Type I error rate

Here we describe how to control the Type I error rate at the nominal level in the presence of an SSR to $n=n_1 + n_2$, where n may be larger than the planned sample size at the study design stage in order to maintain desired conditional power.

After n_1 observations per group, we compute the z score and re-estimate the standardized

treatment effect. Usually the empirical estimate, $\delta = \sqrt{2/n_1} z_1 = \left(\frac{(\bar{X}_C - \bar{X}_T - \Delta)}{\hat{\sigma}} \right)$ is

used. Note that the revised standardized treatment difference may differ from that originally hypothesized either because the difference in means is different than expected or because the variance is different than expected, or both. We then determine a critical value $c = c(n_2, z_1)$ and an additional sample size n_2 such that $CP_0(n_2, c | z_1) = A(z_1)$.

We will reject the null hypothesis if Z , computed by using all $n_1 + n_2$ observations, exceeds c .

Setting $CP_0(n_2, c | z_1) = A(z_1)$ and solving for c yields

$$CP_0(n_2, c | z_1) = \Pr(z > c | z_1, \delta = \delta_\Delta) = 1 - \Phi\left[\frac{c\sqrt{n_1 + n_2} - z_1\sqrt{n_1} + n_1\delta_\Delta / \sqrt{2}}{\sqrt{n_2}}\right] = A(z_1)$$

$$\Rightarrow c\sqrt{n_1 + n_2} - z_1\sqrt{n_1} + n_1\delta_\Delta / \sqrt{2} = \sqrt{n_2} z_A$$

$$\Rightarrow c = \frac{\sqrt{n_2} z_A + \sqrt{n_1} z_1 - n_1\delta_\Delta / \sqrt{2}}{\sqrt{n_1 + n_2}}$$

where z_A is a shorthand notation for $z_{A(z_1)}$. For n_2 near 0, c will be close to $z_1 - n_1 \delta_\Delta / \sqrt{2}$, whereas c approaches z_A for very large n_2 .

If we plug c into the expression for $CP_\delta(n_2, c | z_1)$

$$\begin{aligned} CP_\delta(n_2, c | z_1) &= 1 - \Phi\left[\frac{z_\alpha \sqrt{2(n_1 + n_2)} - z_1 \sqrt{2n_1} + (n_1 + n_2)\delta_\Delta - n_2\delta}{\sqrt{2n_2}}\right] \\ &= 1 - \Phi\left(\frac{\sqrt{2n_2}z_A + \sqrt{2n_1}z_1 - n_1\delta_\Delta - \sqrt{2n_1}z_1 + n_1\delta_\Delta + n_2\delta_\Delta - n_2\delta}{\sqrt{2n_2}}\right) \\ &= 1 - \Phi(z_A - \sqrt{n_2/2}(\delta - \delta_\Delta)) \end{aligned}$$

Set $CP_\delta(n_2, c | z_1) = 1 - \beta$, yields

$$n_2 = \frac{2(z_A + z_\beta)^2}{(\delta - \delta_\Delta)^2}$$

$$c = \frac{(z_A + z_1)z_A + \sqrt{n_1/2}(\delta - \delta_\Delta)z_1 - \delta_\Delta(\delta - \delta_\Delta)n_1/2}{\sqrt{n_1(\delta - \delta_\Delta)^2 + (z_A + z_\beta)^2}}$$

If the empirical estimate $\delta = \sqrt{\frac{2z_1^2}{n_1}}$ is used as an estimate of the true effect size,

$$n_2 = \frac{2(z_A + z_\beta)^2}{(\sqrt{2z_1^2/n_1} - \delta_\Delta)^2}$$

$$c = \frac{(z_A + z_1)z_A + \sqrt{n_1/2}(\sqrt{2z_1^2/n_1} - \delta_\Delta)z_1 - \delta_\Delta(\sqrt{2z_1^2/n_1} - \delta_\Delta)n_1/2}{\sqrt{n_1(\sqrt{2z_1^2/n_1} - \delta_\Delta)^2 + (z_A + z_\beta)^2}}$$

3.5.4 General Procedures for SSR using conditional power

Table 5 outlines the general procedures for SSR to obtain a desired conditional power while controlling for Type I error. When early stopping for futility is not permitted, Step 4.1 will be skipped.

Table 5: General Procedures for two-stage SSR using conditional power in trials with a continuous outcome

Step 1: Estimate initial sample size per group $n_0 = n(\alpha, 1 - \beta, \Delta, \sigma_a^2)$ as if it is a fixed sample design under a certain assumption of δ
Step 2: Simulate the pilot $n_1 = \pi n_0$ samples (e.g., $\pi=0.5$) per arm based on original assumptions
Step 3: Estimate test statistic z_1 from first stage samples
Step 4: Decide: <ol style="list-style-type: none"> 1. If $CP_\delta(n_2, z_\alpha z_1)$ is less than a certain cutoff (e.g., 10%), then STOP for futility, ACCEPT H_0 2. If $CP_\delta(n_2, z_\alpha z_1)$ is greater than a certain pre-specified cutoff (e.g., 50%) but less than the desired conditional power (e.g., 80%), then take additional n_2 samples to achieve adequate conditional power (e.g., 80%). <ul style="list-style-type: none"> • If $n_2 \leq 2 \times n_0$ then take additional n_2 samples • If $n_2 > 2 \times n_0$ then take additional $2 \times n_0$ samples
Step 5: Obtain the remaining subjects based on the re-estimated sample size if the decision at step 4 is to take additional samples.
Step 6: Final analysis With all $n_1 + n_2$ samples <ul style="list-style-type: none"> • Hypothesis testing with final critical value adjustment if necessary

In the section below we further inspect the properties (overall Type I error rate and power) of these approaches.

3.6 Two-stage SSR using the observed variance and the originally specified treatment difference in a conditional power calculation

In sample size re-estimation where only the nuisance parameter drives the recalculation process, there is no need to account for the fact that first stage data will contribute to the final test statistics. The results from Section 3.4 shows that the Type I error rate is well controlled because the first stage data was treated as fixed rather than random in the sample size re-estimation process.

In this section we will discuss that the sample size re-estimation using the “conditional power at θ^a ,” defined as the probability of rejecting H_0 at the end of the study, conditioned on the interim test statistic, under the originally protocol-specified assumed treatment difference θ^a , with $\sigma^2 = \hat{\sigma}_1^2$. Properties to be examined include Type I error rate, power and final sample size. Table 6 highlights four different scenarios that will be discussed in this section: 1) design that does not allow futility stopping at interim and no critical value adjustment at the final stage, 2) design that does not allow futility stopping at interim while adjusting the final critical value, 3) design that allows for early stopping for futility at the interim and no critical value adjustment at the final stage, and 4) design that allows for early stopping for futility at the interim while adjusting the final critical value. Finally the four approaches will be compared with respect to Type I error and

power to allow for an illustration of properties and designs that are sensitive to those rules.

Table 6: Four different scenarios in two-stage SSR using conditional power in trials with a continuous outcome

		Early Stopping for futility	
		No	Yes
Critical Value Adjustment	No	w/o futility stopping w/o critical value adj.	w futility stopping w/o critical value adj.
	Yes	w/o futility stopping w critical value adj.	w futility stopping w critical value adj.

Example (continued)

Suppose after approximately 50% of patients are treated and followed; i.e., about 200 patients are enrolled and treated for each treatment group at the interim. After the first stage, $\hat{\theta} = 0.02$, and $\hat{\sigma}_1^2 = 1.2$, so that $z_1 = 0.46$. The conditional power calculation for 393 patients per group at the original specified difference ($\theta^a = 0$) is 53% when assuming $\sigma^2 = 1.2$. An overall sample size of 693 patients per group (an increase of 76% from the original sample size) would be required to obtain 80% conditional power when the observed interim treatment difference is 0.02. This approach assumes the subsequent data come from a treatment difference of $\theta^a = 0$ rather than the observed difference of $\hat{\theta}$.

Simulated Type I error rate and power for each scenario in this examples are:

	Type I Error Rate	Power
w/o futility stopping w/o critical value adj.	0.03129	0.84926
w futility stopping w/o critical value adj.	0.02718	0.79109
w/o futility stopping w critical value adj.	0.02583	0.81703
w futility stopping w critical value adj.	0.02584	0.78291

3.6.1 No futility stopping at interim and no critical value adjustment at the final analysis

Table A3.5a in Appendix A shows, for the various values of true σ^2 the mean and standard deviation of re-estimated final sample size, % of capping ($2 \times n_0$) of second-stage sample size and corresponding Type I error rate across 100,000 simulations when conditional power is calculated under the originally protocol-specified assumed treatment difference θ^a and $\hat{\sigma}_1^2$ is used as the assumption of σ^2 in the conditional power calculation, where there is neither futility stopping at interim nor the critical value adjustment at the final analysis.

As expected, a larger sample size is required and more capping occurs the larger the true variance deviates from the original assumption at the design stage ($\sigma^2 = 1$). The largest mean sample size is approximately 980 due to capping. For cases where true $\sigma^2 \geq 1.5$, almost all the second stage sample sizes are capped at $2 \times n_0$.

Not surprisingly, the Type I error rate here is significantly inflated for those scenarios with a low percentage of capping. For those cases that true $\sigma^2 \geq 1.5$, where almost all the second-stage sample size is capped at $2 \times n_0$, the Type I error rate is moderately inflated. The source of Type I error inflation is from the fact that the second-stage sample size is dependent on the first stage test statistic (z_1). To further illustrate that Type I error rate inflation is less the more the true variance is under-estimated at the design stage is due to capping, Table A3.5b in Appendix A shows the results when the second-stage sample size is capped at $5 \times n_0$. Take the $\sigma^2 = 1.5$ as an example, the percentage of second-stage sample size capping drops from 99% to 61% and Type I error rate increases from 0.02629 to 0.02838 as the second-stage sample size cap increases from $2 \times n_0$ to $5 \times n_0$.

Table A3.6 in Appendix A shows, for various values of true σ^2 , the mean and standard deviation of re-estimated final sample size, % of capping ($2 \times n_0$) of second-stage sample size and corresponding power across 10,000 simulations when conditional power is calculated under the originally protocol-specified assumed treatment difference θ^a , and $\hat{\sigma}_1^2$ is used as the assumption of σ^2 , where there is neither futility stopping at interim nor the critical value adjustment at the final analysis. Power is sensitive to the true variance value, true treatment difference, and the percentage of second-stage sample size capping. When the true variance and true treatment difference are not very different from the original assumptions with a relatively low percentage of capping, the design achieves the target power well (but, again, at the expense of increase Type I error rate).

3.6.2 With futility stopping at interim and no critical value adjustment at the final analysis

Table A3.7 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of early stopping at interim for futility and corresponding Type I error rate when conditional power is calculated under the originally protocol-specified assumed treatment difference θ^a , and $\hat{\sigma}_1^2$ is used as the assumption of σ^2 , where stopping due to futility at interim is permitted if conditional power is less than a pre-specified level (10%, 20% and 30%) but without critical value adjustment at the final analysis.

The final required sample size rises with increasing true σ^2 due to the increased sample size needed to maintain study power. In addition, the final required sample size decreases with increased cutoff for futility stopping due to the increased chance of a correct decision to accept the null and stop the study after the first-stage analysis.

The Type I error rate is moderately inflated for those scenarios with a low percentage of early stopping, say less than 60%. The Type I error rate is at or below the nominal level when a large percentage of early stopping due to futility occurs.

Table A3.8 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of early stopping at interim for futility and corresponding overall power when conditional power is calculated under the original protocol-specified assumed treatment difference θ^a , and $\hat{\sigma}_1^2$ is used as the assumption of σ^2 , where stopping due to futility at interim is permitted but without critical value adjustment at the final analysis.

Power is sensitive to the true treatment difference. When the true treatment difference is

not very different from the original assumption, the design achieves the target power well. The study can be significantly under-powered depending on the true treatment difference. Also, as expected, the scenarios with the lower significance levels also have lower power due to the allowance of early stopping for futility.

3.6.3 No futility stopping at interim and with critical value adjustment at the final analysis

Table A3.9a in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of capping ($2 \times n_0$) of second-stage sample size and corresponding Type I error rate when conditional power is calculated under the originally protocol-specified assumed treatment difference θ^a , and $\hat{\sigma}_1^2$ is used as the assumption of σ^2 , where there is no futility stopping at interim but the critical value is adjusted as discussed above at the final analysis to maintain Type I error rate at the nominal level.

As expected, the larger sample size is required and more capping occurs the more the true variance deviates from the original assumption ($\sigma^2 = 1$). The sample size peaks at approximately 980 due to capping. For those cases where the true $\sigma^2 \geq 1.5$, almost all the second-stage sample size is capped at $2 \times n_0$.

Using the critical value adjustment derived in Section 3.5.3, the Type I error rate is controlled quite well. Similar results are observed when the second-stage sample size is capped at $5 \times n_0$, shown in Table A3.9b in Appendix A.

Table A3.10 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of capping ($2 \times n_0$) of second-stage sample size and corresponding power

when conditional power is calculated under the originally protocol-specified assumed treatment difference θ^a , and $\hat{\sigma}_1^2$ is used as the assumption of σ^2 , where there is no stopping at interim due to futility, the critical value is adjusted at the final analysis. Power is sensitive to the true variance value, true treatment difference and the percentage of second-stage sample size capping. When the true variance and true treatment difference are not very far from the original assumptions with relatively low percentage of capping, the design achieves the target power well.

3.6.4 With futility stopping at interim and critical value adjustment at the final analysis

Table A3.11 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of early stopping at interim if conditional power is less than a pre-specified level (10%, 20%, and 30%) and corresponding Type I error rate when conditional power is calculated under the originally protocol-specified assumed treatment difference θ^a , and $\hat{\sigma}_1^2$ is used as the assumption of σ^2 , where both stopping at interim due to futility is permitted and critical value may be adjusted at the final analysis.

The final required sample size rises with σ^2 due to the increased sample size needed to protect study power. In addition, the final required sample size decreases with increased cutoff for futility stopping due to the increased chance of a correct decision to accept the null and stop at the first-stage analysis.

The Type I error rate inflation observed in its counterpart (section 3.6.2: with futility stopping at interim and no critical value adjustment) is diminished. The Type I error rate is constant over true variance values.

Table A3.12 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of early stopping at interim if conditional power is less than a pre-specified level (10%, 20%, and 30%) and corresponding power when conditional power is calculated under the originally protocol-specified assumed treatment difference θ^a , and $\hat{\sigma}_1^2$ is used as the assumption of σ^2 , where stopping at interim due to futility is permitted and critical value is adjusted at the final analysis. Power is sensitive to the true treatment difference. When the true treatment difference is not very different from the original assumption, the design achieves the target power well. The study can be significantly under-powered depending on the true treatment difference.

3.6.5 Comparison of procedures

In this section, we will compare the four procedures discussed above in terms of final sample size, Type I error rate, and power.

Table 7 displays the values of final sample size for those designs under the null hypothesis. The final sample size rises with true σ^2 for all four scenarios, there is no difference between average sample size with or without critical value adjustment. In addition, designs with possible futility stopping causes a drop in final sample size for all variance values when compared the design without futility stopping due to the chance of a correct decision to accept the null and stop at the first-stage analysis.

Table 7: Final sample size at different true variances in two-stage SSR using the observed variance and the originally specified treatment difference in a conditional power calculation in trials with a continuous outcome under the null hypothesis

	$\sigma^2 = 1.0$	$\sigma^2 = 1.2$	$\sigma^2 = 1.5$	$\sigma^2 = 1.7$	$\sigma^2 = 2.0$
w/o futility stopping w/o critical value adj.	444±93	660±227	929±166	960±118	974±80
w futility stopping w/o critical value adj.	440±89	452±100	472±116	484±127	500±144
w/o futility stopping w critical value adj.	444±93	660±227	929±166	960±118	974±80
w futility stopping w critical value adj.	440±89	452±100	472±116	484±127	500±144

Table 8 displays the simulated values of Type I error rate for the four designs. With no early stopping for futility, the Type I error rate is inflated when there is no critical value adjustment at final analysis. This inflation of Type I error without critical value adjustment diminishes as the percentage of capping of second-stage sample size increases, but in turn, as shown above and as we will discuss, the corresponding power is lower. With no critical value adjustment but with early stopping for futility, Type I error was inflated when the true variance is similar to the variance assumed at the study design stage, but then decreases below the nominal level as the true variance increases due to the capping of the final sample size at $2 \times n_0$. Using the critical value adjustment derived in Section 3.5.3, the Type I error rate is controlled quite well regardless of the true variance of the outcome.

Table 8: Type I error rate at different true variances in two-stage SSR using the observed variance and the originally specified treatment difference in a conditional power calculation in trials with a continuous outcome

	$\sigma^2 = 1.0$	$\sigma^2 = 1.2$	$\sigma^2 = 1.5$	$\sigma^2 = 1.7$	$\sigma^2 = 2.0$
w/o futility stopping w/o critical value adj.	0.03105	0.03129	0.02629	0.02584	0.02556
w futility stopping w/o critical value adj.	0.02992	0.02718	0.01703	0.0122	0.00819
w/o futility stopping w critical value adj.	0.02530	0.02583	0.02515	0.02514	0.02514
w futility stopping w critical value adj.	0.02531	0.02584	0.02536	0.02537	0.02512

Table 9 displays the values of power for those designs. All four scenarios are shown to be vulnerable to misspecification of the variance and treatment difference. When the true variance and/or treatment difference do not deviate much from the original assumptions, the procedures achieve the target power well. The designs with first-stage futility stopping have power at slightly lower levels than their counterparts that do not have futility stopping.

Table 9: Power at different true variances in two-stage SSR using the observed variance and the originally specified treatment difference in a conditional power calculation in trials with a continuous outcome

	$\mu_C - \mu_T$	$\sigma^2 = 1.0$	$\sigma^2 = 1.2$	$\sigma^2 = 1.5$	$\sigma^2 = 1.7$	$\sigma^2 = 2.0$
w/o futility stopping w/o critical value adj.	0	0.88701	0.90478	0.83568	0.74105	0.60209
	0.02	0.82369	0.84926	0.75462	0.65071	0.51484
	0.05	0.68900	0.71773	0.60033	0.49929	0.38496
w futility stopping w/o critical value	0	0.88134	0.85365	0.67537	0.53352	0.37154

adj.	0.02	0.81557	0.78728	0.59477	0.45621	0.31077
	0.05	0.67707	0.64781	0.45648	0.33797	0.22404
w/o futility stopping w critical value adj.	0	0.87327	0.88540	0.79469	0.69430	0.55636
	0.02	0.80062	0.81703	0.70843	0.60356	0.47429
	0.05	0.65116	0.66543	0.55255	0.45805	0.35086
w futility stopping w critical value adj.	0	0.86987	0.84942	0.68429	0.55443	0.40665
	0.02	0.79633	0.77821	0.60744	0.48237	0.34830
	0.05	0.64652	0.63052	0.47363	0.36776	0.26190

In summary, in this type of procedure it is necessary *to* protect the Type I error rate due to the dependence of the sample size on z_1 , the first-stage test statistic. The method that we derived to adjust the rejection region for the final test statistic seems to work well in dealing with this issue.

An early stopping rule due to futility for the first stage may be desirable; for instance, futility might be declared when the sample size required to ensure a reasonable conditional power is deemed unfeasibly large or when conditional power is below a certain level such as 10%.

3.7 Two-stage SSR using the observed variance and observed treatment difference in a conditional power calculation

Several authors^{6,8,9,60,61,62} have proposed for superiority trials that $\hat{\theta}$ (the interim estimate of the observed difference between treatment means) be used in place of the originally specified (θ^a), in an interim sample size recalculation procedure. This could be desirable

if, at the interim, it appears that the magnitude of θ was too optimistic and that a smaller difference closer to $\hat{\theta}$ is more realistic, but still clinically relevant. Sample size recalculation could then involve recalculating the sample size using $\hat{\theta}$ in a conditional power calculation and then determining the sample size required to yield adequate, e.g., at least 80%, conditional power for a significant result at the end of the study. Such a sample size recalculation may result in an increase in sample size depending on the magnitude of the treatment difference anticipated in the second stage.

In this section we will discuss how the sample size re-estimation uses the “conditional power at θ^a ,” defined as the probability of rejecting H_0 at the end of the study, conditional on the interim test statistic, under the observed treatment difference $\hat{\theta}$, with σ^2 set to $\hat{\sigma}_1^2$. Properties to be examined include Type I error rate, power and final sample size in the four different scenarios specified in Section 3.6.

Example (continued)

Suppose after approximately 50% of patients are treated and followed; i.e., about 200 patients are enrolled and treated for each treatment group at the interim. After first stage, $\hat{\theta} = 0.02$, and $\hat{\sigma}_1^2 = 1.2$, so that $z_1 = 0.46$. The conditional power calculation for 393 patients per group at the original specified difference ($\theta^a = 0$) is 53% when assuming $\sigma^2 = 1.2$. An overall sample size of 725 patients per group (an increase of 84% from the original sample size) would be required to obtain 80% conditional power when the observed interim treatment difference is 0.02. This approach assumes the subsequent data

come from the observed treatment difference of $\hat{\theta}$.

Simulated Type I error rate and power for each scenario in this examples are:

	Type I Error Rate	Power
w/o futility stopping w/o critical value adj.	0.03152	0.86005
w futility stopping w/o critical value adj.	0.02077	0.67419
w/o futility stopping w critical value adj.	0.02544	0.82568
w futility stopping w critical value adj.	0.02544	0.68295

3.7.1 No futility stopping at interim and no critical value adjustment at the final analysis

Table A3.13a in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of capping ($2 \times n_0$) of second-stage sample size and corresponding Type I error rate when conditional power is calculated under the observed treatment difference $\hat{\theta}$, with σ^2 set to $\hat{\sigma}_1^2$, where there is neither futility stopping at interim nor the critical value adjustment at the final analysis.

As expected, a larger final sample size is required and the more capping occurs as the σ^2 and/or the assumption of θ move away from the original assumptions ($\sigma^2 = 1$ and $\theta^a = 0$).

Not surprisingly, the Type I error rate here is significantly inflated for all the values of variance.

Table A3.14 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of capping ($2 \times n_0$) of second-stage sample size and corresponding power when conditional power is calculated under the observed treatment difference $\hat{\theta}$, with σ^2 set to $\hat{\sigma}_1^2$, where there is neither futility stopping at interim nor the critical value adjustment at the final analysis. Power is sensitive to the true variance value, true treatment difference and the percentage of second-stage sample size capping. With a target level of 80%, the study can be significantly over- or under powered, depending on the true variance and/or true treatment difference. The under-powering is caused by the capping of second-stage sample size.

3.7.2 With futility stopping at interim and no critical value adjustment at the final analysis

Table A3.15 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of early stopping at interim if conditional power is less than a pre-specified level (10%, 20%, and 30%), and corresponding Type I error rate when conditional power is calculated under the observed treatment difference $\hat{\theta}$, with σ^2 set to $\hat{\sigma}_1^2$, where futility stopping at interim is permitted but without critical value adjustment at the final analysis.

The final required sample size rises with σ^2 due to the increased sample size needed to protect study power. In addition, the final required sample size decreases with increased cutoff for futility stopping due to the increased chance of a correct decision to accept the null and stop at the first-stage analysis.

The Type I error rate is controlled for those scenarios for all the values of variance due to allowing for early stopping for futility at interim. Controlling Type I error by allowing early stopping comes with a price, which is that it leads to the under-powered study. The values for power are shown in Table A3.16 in Appendix A.

3.7.3 No futility stopping at interim and with critical value adjustment at the final analysis

Table A3.17a in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of capping ($2 \times n_0$) of second-stage sample size and corresponding Type I error rate when conditional power is calculated under the observed treatment difference $\hat{\theta}$, with σ^2 set to $\hat{\sigma}_1^2$, where there is no futility stopping at interim but the critical value is adjusted at the final analysis.

As expected, a larger sample size is required and the more capping occurs as the farther away the true variance and/or treatment difference deviates from the original assumptions ($\sigma^2 = 1$ and $\theta^a = 0$)

Using the critical value adjustment derived in Section 3.5.3, the Type I error rate is controlled quite well. Similar results are observed when the second-stage sample size is capped at $5 \times n_0$, shown in Table A3.17b in Appendix A.

Table A3.18 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of capping ($2 \times n_0$) of second-stage sample size and corresponding power when conditional power is calculated under the observed treatment difference $\hat{\theta}$, with σ^2 set to $\hat{\sigma}_1^2$, where there is no futility stopping at interim but adjusting critical value at

the final analysis. Power is sensitive to the true variance value, true treatment difference and the percentage of second-stage sample size capping. With an initial target level of 80%, the study can be significantly over- or under-powered, depending on the true variance and/or true treatment difference and percentage of second-stage sample size capping.

3.7.4 With futility stopping at interim and critical value adjustment at the final analysis

Table A3.19 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of early stopping at interim if conditional power is less than a pre-specified level (10%, 20%, and 30%), and corresponding Type I error rate when conditional power is calculated under the observed treatment difference $\hat{\theta}$, with σ^2 set to $\hat{\sigma}_1^2$, where both futility stopping at interim is permitted and adjusting critical value at the final analysis.

The final required sample size rises with σ^2 due to the increased sample size needed to protect study power. In addition, the final required sample size decreases with increased cutoff for futility stopping due to the increased chance of a correct decision to accept the null and stop at the first-stage analysis.

The Type I error rate inflation observed in its counterpart (section 3.7.2: with futility stopping at interim and no critical value adjustment) is diminished. The Type I error rate is constant and controlled at the nominal level over true variance values.

Table A3.20 in Appendix A shows the mean and standard deviation of re-estimated final sample size, % of early stopping at interim if conditional power is less than a pre-specified level (10%, 20%, and 30%), and corresponding power when conditional power is calculated under the observed treatment difference $\hat{\theta}$, with σ^2 set to $\hat{\sigma}_1^2$, where futility stopping at interim is permitted and critical value is adjusted at the final analysis. Power is sensitive to the true treatment difference. When the true treatment difference is not very different from the original assumption, the design achieves the target power well. The study can be significantly under-powered depending on the true treatment difference.

3.7.5 Comparison of procedures

In this section, we will compare four procedures discussed in earlier sections in terms of final sample size, Type I error rate, and power.

Table 10 displays the values of final sample size for those designs under the null hypothesis. As expected, the final sample size rises with the true σ^2 for all four scenarios, there is no difference between average sample size with or without critical value adjustment. In addition, designs with possible futility stopping cause a drop in final sample size for all variance values when compared to the design without futility stopping due to the chance of a correct decision to accept the null and stop at the first-stage analysis. It appears that seeking to maintain conditional power at $\theta = \hat{\theta}$ inevitably increases the final sample size moderately.

Table 10: Final sample size at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome under the null hypothesis

w/o futility stopping w/o critical value adj.	571±249	687±273	814±249	867±220	917±176
w futility stopping w/o critical value adj.	554±424	578±462	605±515	611±543	598±572
w/o futility stopping w critical value adj.	571±249	687±273	814±249	867±220	917±176
w futility stopping w critical value adj.	554±424	578±462	605±515	611±543	598±572

Table 11a displays the values of Type I error rate for those designs. The Type I error rate was significantly inflated when there is no early stopping for futility at interim and no critical value adjustment at final analysis. Using the critical value adjustment derived in Section 3.4.3, the Type I error rate is controlled quite well. Allowing for early stopping for futility regardless of the critical value adjustment at the final analysis seems to be able to control Type I error as well, though early stopping without a critical value adjustment may cause too conservative a test (a test where the overall Type I error rate is below the nominal value).

Table 11a: Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome ($\sigma_1^2 = \sigma_2^2$)

w/o futility stopping w/o critical value adj.	0.03234	0.03152	0.02991	0.02926	0.02817
w futility stopping	0.02274	0.02077	0.01783	0.01652	0.01430

w/o critical value adj.					
w/o futility stopping w critical value adj.	0.02564	0.02544	0.02522	0.02540	0.02510
w futility stopping w critical value adj.	0.02563	0.02544	0.02534	0.02577	0.02539

Table 11b, Table 11c and Table 11d display the values of Type I error rate for those designs, where the variances are assumed to be not the same in each treatment group.

The findings are similar to the cases where the equal variance was assumed.

Table 11b: Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome ($\sigma_1^2 = 0.75 \times \sigma_2^2$)

	$\sigma_2^2 = 1.0$	$\sigma_2^2 = 1.2$	$\sigma_2^2 = 1.5$	$\sigma_2^2 = 1.7$	$\sigma_2^2 = 2.0$
w/o futility stopping w/o critical value adj.	0.0308	0.0308	0.0304	0.0303	0.0298
w futility stopping w/o critical value adj.	0.0289	0.0297	0.0284	0.0263	0.0257
w/o futility stopping w critical value adj.	0.0251	0.0247	0.0252	0.0256	0.0254
w futility stopping w critical value adj.	0.0255	0.0252	0.0251	0.0253	0.0249

Table 11c: Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome ($\sigma_1^2 = 0.5 \times \sigma_2^2$)

	$\sigma_2^2 = 1.0$	$\sigma_2^2 = 1.2$	$\sigma_2^2 = 1.5$	$\sigma_2^2 = 1.7$	$\sigma_2^2 = 2.0$
w/o futility stopping w/o critical value adj.	0.0299	0.0297	0.0304	0.0298	0.0290
w futility stopping w/o critical value	0.0281	0.0274	0.0281	0.0280	0.0260

adj.					
w/o futility stopping w critical value adj.	0.0257	0.0240	0.0251	0.0253	0.0252
w futility stopping w critical value adj.	0.0253	0.0254	0.0253	0.0253	0.0254

Table 11d: Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome ($\sigma_1^2 = 0.25 \times \sigma_2^2$)

	$\sigma_2^2 = 1.0$	$\sigma_2^2 = 1.2$	$\sigma_2^2 = 1.5$	$\sigma_2^2 = 1.7$	$\sigma_2^2 = 2.0$
w/o futility stopping w/o critical value adj.	0.0292	0.0295	0.0304	0.0294	0.0291
w futility stopping w/o critical value adj.	0.0297	0.0275	0.0273	0.0265	0.0261
w/o futility stopping w critical value adj.	0.0258	0.0253	0.0249	0.0250	0.050
w futility stopping w critical value adj.	0.0252	0.0253	0.0251	0.0248	0.0254

Finally we investigated the impact of capping the second stage sample size on Type I error rates. As expected, the Type I error rate was significantly inflated when there is no early stopping for futility at interim and no critical value adjustment at final analysis. Using the critical value adjustment derived in Section 3.4.3, the Type I error rate is controlled quite well. Allowing for early stopping for futility regardless of the critical value adjustment at the final analysis seems to be able to control Type I error as well, though early stopping without a critical value adjustment may cause too conservative a test (a test where the overall Type I error rate is below the nominal value).

Table 11e: Type I error rate at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome (without capping the stage 2 sample size)

	$\sigma^2 = 1.0$	$\sigma^2 = 1.2$	$\sigma^2 = 1.5$	$\sigma^2 = 1.7$	$\sigma^2 = 2.0$
w/o futility stopping w/o critical value adj.	0.0311	0.0306	0.0309	0.0299	0.0293
w futility stopping w/o critical value adj.	0.0299	0.0284	0.0273	0.0265	0.0253
w/o futility stopping w critical value adj.	0.0257	0.0253	0.0255	0.0254	0.0251
w futility stopping w critical value adj.	0.0252	0.0249	0.0251	0.0246	0.0246

Table 12 displays the values of power for those designs. As expected, all four scenarios are shown to be vulnerable to misspecification of the variance and treatment difference. When the true variance and/or treatment difference do not deviate significantly from the original assumptions, the procedures seem to result in a study that is adequately powered. While the true variance and/or treatment difference are different from the original assumptions, the procedure can not appreciably relieve the danger of under-powering a study, in large part due to the practical capping placed on overall sample size. The designs with first-stage futility stopping have power at slightly lower levels than their counterparts without futility stopping.

Table 12: Power at different true variances in two-stage SSR using the observed variance and treatment difference in a conditional power calculation in trials with a continuous outcome

	$\mu_C - \mu_T$	$\sigma^2 = 1.0$	$\sigma^2 = 1.2$	$\sigma^2 = 1.5$	$\sigma^2 = 1.7$	$\sigma^2 = 2.0$
w/o futility stopping w/o critical value adj.	0	0.94968	0.91012	0.80778	0.72038	0.59320
	0.02	0.92029	0.86005	0.72681	0.63342	0.50836
	0.05	0.83789	0.73776	0.57985	0.48774	0.38265
w futility stopping w/o critical value adj.	0	0.84557	0.73361	0.57141	0.47619	0.36268
	0.02	0.78207	0.65915	0.49321	0.40549	0.30460
	0.05	0.65545	0.52492	0.37379	0.29988	0.22178
w/o futility stopping w critical value adj.	0	0.94438	0.88866	0.76248	0.66855	0.54061
	0.02	0.90642	0.82568	0.67426	0.57777	0.45966
	0.05	0.80096	0.68272	0.51986	0.43520	0.33924
w futility stopping w critical value adj.	0	0.84818	0.73897	0.58455	0.49639	0.39106
	0.02	0.78520	0.66594	0.50990	0.42836	0.33407
	0.05	0.65814	0.53274	0.39136	0.32336	0.25041

In summary, in this type of procedure, it is necessary to protect the Type I error rate due to the dependence of the sample size on z_1 , the first-stage test statistic. The method that we derived to adjust the rejection region for the final test statistic seems to work well at dealing with this issue.

Additionally, the design with futility stopping once again shows the added benefit of yielding a reduced sample size but at the cost of decreased power.

The sample size resulting from this method is clearly sensitive to $\hat{\theta}$. Investigators must understand the ramifications of this and should exercise caution if adopting this approach. For example, if $\hat{\theta}$ at an interim analysis is smaller than expected, it will cause the test statistic to be small as well, thereby influencing the conditional power calculation both through the test statistic on which it is conditioned and the magnitude of the treatment difference anticipated in the second stage.

Chapter 4: BINOMIALLY DISTRIBUTED ENDPOINTS

4.1 Introduction and Background

Sample size estimation is a key step for a successful clinical research. However, it is well recognized that estimation of sample size in clinical trials requires knowledge of parameters that involve treatment effect and variability, which are usually uncertain to medical researchers. This uncertainty is especially pronounced for clinical trials with new classes of study therapies for unfamiliar diseases where natural history data are lacking. Misspecification of treatment effect or of variability results in either an over-sized study, which is neither economical to the researcher nor ethical to the patients, or an under-powered study, which may yield inconclusive results. Even with previous data available, one must exercise caution regarding possible differences between the trials in terms of patient population, disease severity, diagnostic criteria, medical procedures, and other study conditions. It is therefore desirable to re-estimate the sample size using interim data of the trial under study.

In non-inferiority trials with binary endpoints, in contrast to conventional superiority studies, the proper determination of the clinically acceptable absolute margin (δ), whether or not applied to a risk difference or relative risk, is a vital problem. It is commonly accepted that δ has to be stated in advance of the study and must not be a post hoc decision that occurred after the data have been analyzed. In ideal situations, where event rates for the control group are well established, a fixed δ for a relative risk is appropriate. I.e., the following null and alternative hypothesis would be appropriate to test:

$$H_0: \pi_C - \pi_T \geq \delta$$

$$H_a: \pi_C - \pi_T < \delta$$

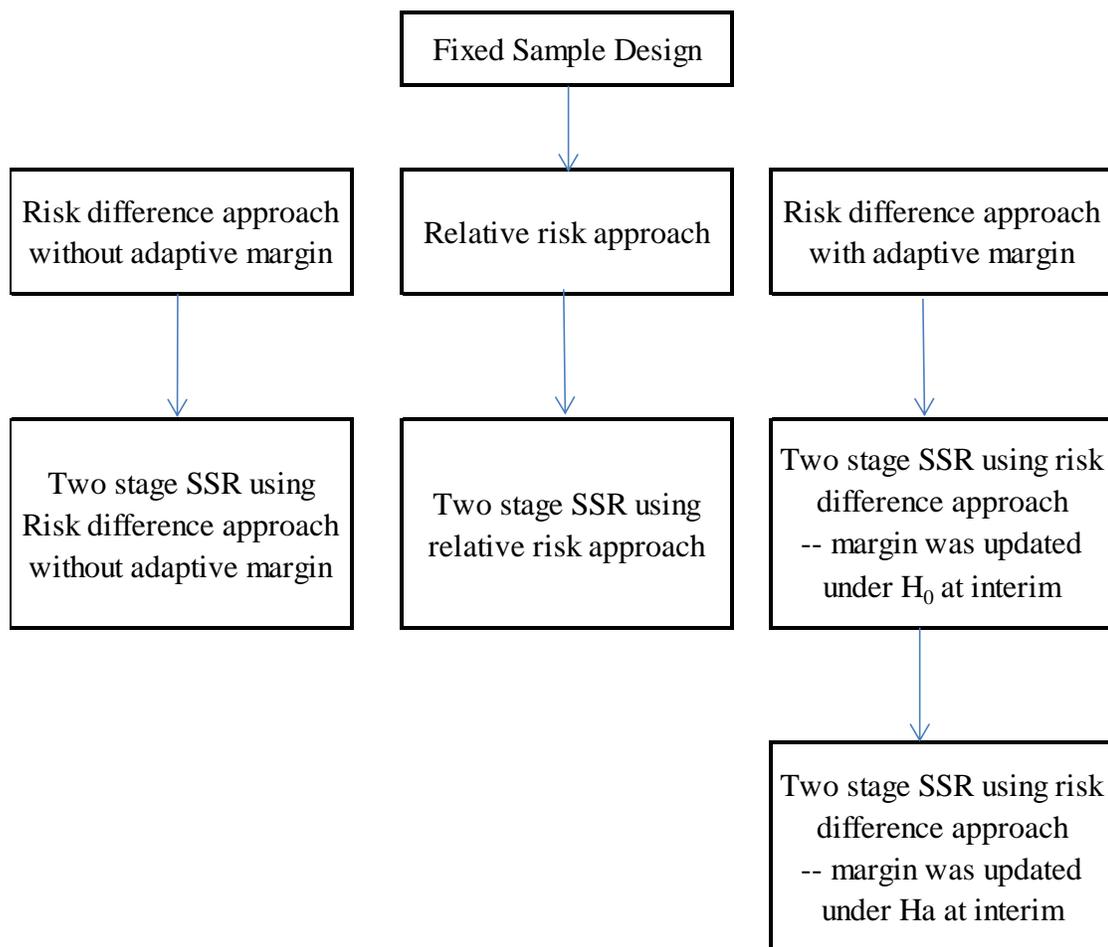
Where π_C and π_T are the true binomial event rates for the control and treatment groups, respectively (assume without loss of generality the outcome is a positive outcome such as survival or cancer remission) and where $\delta > 0$ is the pre-specified non-inferiority margin.

However, fixed δ s may not be justifiable in situations in which the expected event rates are difficult to estimate. For example, if the expected (positive) event rate for the control arm is assumed to be 90%, it may be reasonable to set the non-inferiority margin to be 10% as is often done in anti-infective trials where the outcome is infection cure; but if the actual control rate is only 70%, a margin of 10% may be less clinically applicable. Thus, revising the margin to reflect the revised estimate of the control group rate is desired (e.g., if a 10% margin is appropriate for a 90% control group event rate, then a 7.8% margin, or $70\% \times (10\%/90\%)$, might be appropriate for a 70% control group event rate).

The question we are trying to answer is how to adapt the non-inferiority margin and resulting sample size based on the interim estimate of the observed pooled (blinded) event rate in order to make sure the study is adequately powered to claim non-inferiority at the end of the trial, while maintaining the nominal Type I error. Our discussion will be focused on the statistical and clinical implications of changing the absolute margin and sample size based on the interim estimate of the pooled event rate, and then revising the margin again if necessary based on the estimate of the control group event rate at the end

of the study. Figure 2 outlines the progression of the approaches to sample size re-estimation in binomially distributed outcome that will be discussed in this chapter. All the procedures will be applied to both positive outcomes (the larger values of the outcome proportion indicate a more beneficial response to treatment, such as cancer remission or infection cure) and negative outcomes (the larger values of the outcome proportion indicate a less beneficial response to treatment, such as death or occurrence of major cardiovascular event, an outcome often used in cardiac device trials).

Figure 2: The progression of approaches to two-stage SSR in binomially distributed outcome



4.2 Computational Methods

Simulation studies were carried out to evaluate the performance of the procedure over a range of circumstances likely to be met in most clinical trials, including when the null hypothesis was true and when it was false. All programs for simulations were written in SAS version 9.3 (SAS Institute, 2004). Under each scenario, 10,000 and 100,000 trials were simulated for Type I error rate and power, respectively.

The normal distribution was used to approximate the binomial distribution for Type I error rate and power calculation. The validity of the test depends on having sample sizes that justify the usual arguments for asymptotic normality, and this assumption will carry over through the entire chapter.

We restrict our investigations to two tests, namely Blackwelder's test⁷ and Farrington and Manning's test⁶³, which we believe are commonly used in practice.

In addition, calculations were made for optimal allocation (i.e., sample sizes are equal in each treatment group), and the interim analysis conducted at $\frac{1}{2}$ the originally calculated sample size.

The other parameters required for the specification of model and test ran through all combinations of an extended range of values are shown in Table 13.

Here the assumed parameter values δ_1^a and π^a are used for the initial sample size calculation, and the actual values δ_1 (for assessment of Type I error) and π are used in the true model, from which the distribution of possible test outcomes is taken. Additionally, as a practical consideration, different values of δ_1^a and π^a are used for positive and negative outcomes. As for the positive outcomes, like the success rate in anti-infective trials, the success rate is usually around 80%-90%. On the contrary, for the negative outcomes, like MACE (major adverse cardiac event) in cardiac device trials, the event rate is usually in the neighborhood of 5%-10%.

Table 13: Parameter values used for the simulations of Type I error rate and power for non-inferiority tests with binary outcome

		Positive Outcome	Negative Outcome
Nominal significance level	α	2.5%	2.5%
Target power of the test	$1 - \beta$	80%	80%
Sample ratio (n_C/n_T)	θ	1	1
Non-inferiority margin	δ	0.1 for fixed margin 0.125 \times π for adaptive margin	0.05 for fixed margin 0.5 \times π for adaptive margin
Assumed treatment difference for purpose of designing the study	δ_1^a	0	0
Assumed overall event rate (both treatments pooled)	π^a	0.8	0.1
Actual treatment difference for assessment of Type I error	δ_1	0.1 for fixed margin 0.125 \times π for adaptive margin	0.05 for fixed margin 0.5 \times π for adaptive margin
Actual overall event rate	π	0.95, 0.9, ...0.6, 0.55	0.05, 0.10.4, 0.45

4.3 Fixed Sample Size Design

4.3.1 Risk difference approach without adaptive margin

Consider a non-inferiority clinical trial that compares the effect of two treatments on a disease based on a binary outcome (x), for which we use the generic term “event” ($x=1$) versus “non-event” ($x=0$). Denote π_T and π_C as the true event rates for treatment group and active control group, respectively. If n_i independent observations are drawn from

population i , $i = T, C$, then X_i is the number of observed events in group i , and is binomially distributed with parameters n_i and π_i , that is $X_i \sim \text{Bin}(n_i, \pi_i)$, $i = T, C$. For a positive outcome (the larger values of the outcome indicate a more beneficial response to treatment), the hypotheses to test are:

$$H_0: \pi_C - \pi_T \geq \delta$$

$$H_a: \pi_C - \pi_T < \delta$$

where $\delta > 0$ is the pre-specified non-inferiority margin.

The test statistic used to test the above hypotheses is $z = (\hat{\pi}_C - \hat{\pi}_T - \delta) / \widehat{SE}$, where $\hat{\pi}_i = x_i/n_i$ is the unrestricted maximum likelihood estimate (MLE) of π_i , $i = T, C$, and \widehat{SE} an estimate of the standard error of $\hat{\pi}_C - \hat{\pi}_T$. Under the null hypothesis z asymptotically follows a standard normal distribution. The null hypothesis can be rejected at level α if $z \leq \Phi^{-1}(1 - \alpha)$, where $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. Blackwelder⁷ suggested using the unrestricted MLE of π_i to compute the standard error, whereas Farrington and Manning⁶³ proposed the use of restricted MLE which is constrained by the null hypothesis.

At the planning stage we need to estimate the sample size for the trial. Conventionally, we do this by assuming values of π_T and π_C , respectively. The clinically meaningful non-inferiority margin, δ also needs to be specified at the design stage. The sample size depends on the true treatment difference $\delta_1 = \pi_C - \pi_T$, the non-inferiority margin (δ), the desired power ($1 - \beta$), and the desired (nominal) Type I error rate (α). Roebuck and Kuhn⁶⁴ described several non-inferiority tests for binary data and compared them in

an extensive simulation study. In this study, we will concentrate on the tests of Blackwelder and of Farrington and Manning and assume the treatment allocation ratio of

$$\frac{n_T}{n_C} = 1 \equiv \theta$$

For Blackwelder's test, uses the sample size formula is

$$n_T = \frac{(\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2}{(\pi_C - \pi_T - \delta)^2} \cdot (\pi_T(1 - \pi_T) + \pi_C(1 - \pi_C))$$

For Farrington and Manning's test, the corresponding formula is

$$n_T = \frac{(\Phi^{-1}(\alpha)\sqrt{V_0} + \Phi^{-1}(\beta)\sqrt{V_1})^2}{(\pi_C - \pi_T - \delta)^2}$$

$$V_0 = \tilde{\pi}_T(1 - \tilde{\pi}_T) + \tilde{\pi}_C(1 - \tilde{\pi}_C)$$

$$V_1 = \pi_T(1 - \pi_T) + \pi_C(1 - \pi_C)$$

Where V_0 and V_1 are variances under H_0 and H_a respectively. $\tilde{\pi}_T$ and $\tilde{\pi}_C$ are Farrington and Manning's large sample approximations of π_T and π_C .

Example

Assume an oncology clinical trial involves active control with a study objective of non-inferiority. The end point of the trial is cancer remission. At the planning stage, we assume that the expected remission rates for both treatment and control arms are 80%. Non-inferiority is to be assessed from a risk difference perspective. The non-inferiority margin of 10% was pre-specified and deemed clinically acceptable.

The hypotheses to be tested are

$$H_0: \pi_C - \pi_T \geq 10\%$$

$$H_a: \pi_C - \pi_T < 10\%$$

The trial would be declared as positive at the end of the study if the estimated remission rate in treatment group ($\hat{\pi}_T$) is greater than the remission rate in control group ($\hat{\pi}_C$) - 10%. In order to achieve a power of 80% with one-sided $\alpha = 0.025$ to demonstrate non-inferiority, a sample size of 255 patients per group would have been required using Farrington and Manning's method.

Table A4.1a and Table A4.1b in Appendix A shows the simulated Type I error rate for Blackwelder's and Farrington-Manning's tests in the fixed sample size design with risk difference approach for both positive (Table A4.1a) and negative (Table A4.1b) outcomes with a fixed non-inferiority margin of 0.1. The Type I error rate was not substantially affected when the true event rate was mis-specified at the design stage (as is expected due to the normal approximation of the tests, the Type I error rate was slightly different from the nominal level at the extreme end of outcome rates). Blackwelder's test and Farrington-Manning's test are very comparable, with some fluctuation due to the random nature of the total sample size.

Since power depends on the expected event rate, which usually represents the investigator's anticipated treatment effects, its credibility is often questionable for various reasons, such as incomplete knowledge regarding the disease and treatment under study. Obviously, the appropriateness of the initially estimated sample size depends on how close the assumed π_i are to the true π_i , where $i = T, C$. Misspecification of the true event rates at the design stage could lead to a poor estimate of the sample size necessary to attain the pre-specified power at a given one-sided Type I error rate α .

Table A4.2a and Table A4.2b present the simulated actual power (assuming the treatment difference is zero) over a range of possible but unknown true event rates for both positive (Table A4.2a) and negative (Table A4.2b) outcomes. The table suggests that over-estimate of the event rate can lead to an underpowered trial for positive outcome and the under-estimate of the event rate can lead to an underpowered trial for negative outcome. In addition to the under-powering problem, the design of a risk difference approach with a fixed margin raises another serious concern: if the true event rate is 60% as opposed to 80% as originally assumed at the design stage, is 10% absolute non-inferiority margin clinically acceptable?

4.3.2 Relative risk approach

Relative risk measures how much the risk is reduced or increased in the experimental group compared to the active control group.

For a positive outcome (the larger values of the outcome indicate a more beneficial response to treatment), the hypotheses to test in a non-inferiority setting are:

$$H_0: \pi_T/\pi_C \leq R$$

$$H_a: \pi_T/\pi_C > R$$

where $R < 1$ is the pre-specified non-inferiority margin.

The test statistic used to test the above hypotheses is $z = (\hat{\pi}_T - R \times \hat{\pi}_C) / \widehat{SE}$, where $\hat{\pi}_i = x_i/n_i$ is the unrestricted maximum likelihood estimate (MLE) of π_i , $i = T, C$, and \widehat{SE} an estimate of the standard error of $\hat{\pi}_T - R \times \hat{\pi}_C$. Under the null hypothesis z asymptotically follows a standard normal distribution. The null hypothesis can be

rejected at level α if $z \leq \Phi^{-1}(1 - \alpha)$, where $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution.

Farrington and Manning⁶³ proposed the use of restricted MLE of π_i to compute the standard error, which is $\widehat{SE} = \sqrt{((\tilde{\pi}_T(1 - \tilde{\pi}_T) + R^2\tilde{\pi}_C(1 - \tilde{\pi}_C))/n_0)}$, where $\tilde{\pi}_T$ and $\tilde{\pi}_C$ are Farrington and Manning's large sample approximations of π_T and π_C under the null hypothesis.

For the normal approximation test, the standard error is

$\widehat{SE} = \sqrt{((\hat{\pi}_T(1 - \hat{\pi}_T) + R^2(\hat{\pi}_C(1 - \hat{\pi}_C)))/n_0)}$, where $\hat{\pi}_T$ and $\hat{\pi}_C$ are the unrestricted maximum likelihood estimate (MLE) of estimate of π_T and π_C under the null hypothesis.

Example (continued)

Suppose the study is instead designed with a relative risk approach. Therefore, the trial would be considered as positive if the remission rate in the treatment group is within 0.875 of remission rate in the control group.

The hypotheses to be tested are

$$H_0: \pi_T/\pi_C \leq 0.875$$

$$H_a: \pi_T/\pi_C > 0.875$$

In order to achieve a power of 80% with one-sided $\alpha = 0.025$ to demonstrate non-inferiority, a sample size of 231 patients per group is required using Farrington and Manning's method.

Table A4.3a and Table A3.3b in Appendix A displays the values of Type I error rate for normal approximation and Farrington-Manning's tests in the fixed sample size design with relative risk approach for both positive (Table A4.3a) and negative (Table A 4.3b) outcomes. In cases with positive outcomes, when the true event rate was mis-specified at the design stage, the Type I error rate was close to the nominal level. The normal approximation test and Farrington-Manning's test are very comparable. In cases with negative outcomes, Type I error rate is somewhat inflated when the normal approximation test is used. The Type I error rate lies closer to the nominal value when Farrington-Manning's test is used.

For a design with a relative risk approach, the power at a given sample size is quite sensitive to the true control event rate π_c as shown in Table A4.4a and Table A4.4b. Thus, if the control event rate was mis-specified, the trial may be over or underpowered at a given sample size. For a negative outcome (Table A4.4b), if the true control event rate is less than the assumed event rate, the trial could be underpowered at a given sample size. This could be a serious problem for the trial, as the control event rate at the design stage for a negative outcome trial, if mis-specified, is often over-estimated due to temporal trends (e.g., control event rate estimates may have come from trials executed several years earlier, but medical practice has improved since then thereby reducing the true control event rate). For a positive outcome, when power is based on relative risk and the true control group event rate is under-estimated, the trial is underpowered as shown in Table A4.4a in Appendix A.

4.3.3 Risk difference approach with adaptive non-inferiority margin based on observed control group rate at the end of the study

In non-inferiority trials with a binary endpoint, in contrast to conventional superiority studies, the proper determination of the clinically acceptable absolute margin (δ) is a vital problem. It is commonly accepted that the non-inferiority margin has to be stated in advance of the study and must not be a post hoc decision, occurring after the data have been analyzed. In ideal situations, where event rates for the control group are well-established, a fixed δ is appropriate. However, fixed δ 's may not be justifiable in situations in which the expected event rates are difficult to estimate. Therefore, a test that allows the non-inferiority margin to vary with the underlying event rate is desired. The proposed test itself is a simple modification of the standard fixed-margin test:

For a positive outcome (the larger values of the outcome indicate a more beneficial response to treatment), the hypotheses to test from a risk difference approach are:

$$H_0: \pi_C - \pi_T \geq \delta$$

$$H_a: \pi_C - \pi_T < \delta$$

where $\delta = b\pi_C > 0$ is the pre-specified non-inferiority margin, where $b > 0$ and π_C is the assumed control group event rate.

The test statistic used to test the above hypotheses is $z = ((1 - b)\hat{\pi}_C - \hat{\pi}_T)/\widehat{SE}$, where $\hat{\pi}_i = x_i/n_i$ is the unrestricted MLE of π_i , $i = T, C$, and \widehat{SE} , for the normal approximation test, is an estimate of the standard error of $(1 - b)\hat{\pi}_C - \hat{\pi}_T$, which is

$$\widehat{SE} = \sqrt{((\hat{\pi}_T(1 - \hat{\pi}_T) + (1 - b)^2\hat{\pi}_C(1 - \hat{\pi}_C))/n_0)}. \text{ For the Farrington-Manning's test,}$$

$\widehat{SE} = \sqrt{((\tilde{\pi}_T(1 - \tilde{\pi}_T) + (1 - b)^2\tilde{\pi}_C(1 - \tilde{\pi}_C))/n_0}$, where $\tilde{\pi}_T$ and $\tilde{\pi}_C$ are Farrington and Manning's large sample approximations of π_T and π_C under the null hypothesis.

Below we investigate the Type I error rate when using this test statistic to assess non-inferiority for positive and negative binary outcomes.

Example (continued)

When the expected remission rate for the control arm was assumed to be 80%, it may be reasonable to set the non-inferiority margin to be 10%; but if the actual remission rate for the control arm is actually 60%, a margin of 10% may be less clinically applicable. Thus, revising the margin to reflect the revised estimate of the remission rate in the control group is desired. If a 10% margin is appropriate for an 80% remission rate in the control group, then a 7.5% margin, or $60\% \times (10\%/80\%)$, might be appropriate for a 60% remission rate in the control group.

The hypotheses to be tested are

$$H_0: \pi_C - \pi_T \geq \delta$$

$$H_a: \pi_C - \pi_T < \delta$$

Where $\delta = 10\%\pi_C$. Then with 255 patients/group as calculated at the beginning of the trial using the original assumption of $\pi^a = 0.8$, the power would only have been 42% rather than the envisioned 80%. If the true event rates are 52.5% and 60% for the treatment and active control groups, respectively, which is considerably lower than the 80% initially expected.

Table A4.5a and Table A4.5b show that the simulated Type I error rate for various values of true control group event rate at the design stage, but the non-inferiority margin was allowed to vary based on the final observed control group event rate using both Blackwelder's test and Farrington-Manning's tests. In cases with positive outcomes, when the event rate in the active control group was under-estimated at the design stage, the Type I error rate was slightly inflated when Farrington-Manning's test was used. There is no inflation observed if Blackwelder's test was used. In cases with negative outcomes, Type I error rate is considerably inflated when the normal approximation test is used. The Type I error rate lies lower than the nominal ones when Farrington-Manning's test is used.

Since power depends on the expected event rate, which usually represents the investigator's anticipated treatment effects, its credibility is often questionable for various reasons, such as incomplete knowledge regarding the disease and treatment under study. Obviously, the appropriateness of the initially estimated sample size depends on how close the assumed $\hat{\pi}_i$ are to the true π_i , where $i = T, C$. Misspecification of the event rates could lead to a poor estimate of the sample size necessary to attain the pre-specified power at a given Type I error rate α .

Table A4.6a and Table A4.6b in Appendix A describes the actual simulated power under a range of true event rates using both Blackwelder's and Farrington-Manning's test for both positive and negative outcomes. As expected, for a positive outcome (Table A4.6a) if the true event rate is lower than the original assumption, the study was under-powered;

on the other hand, if the true event rate is greater than the original assumption, the study was over-powered, many resources could have been saved if the event rate had not been mis-specified. The reverse relationship is true for a negative outcome (Table A4.6b).

4.3.4 Comparison of fixed designs

In this section, we will compare two fixed designs – relative risk and risk difference with adaptive margin. Both designs allow the margin to change depending on the final observed event rate in the control group.

Table 14 displays the values for simulated power for both designs with positive outcome, including situations when the true event rate is mis-specified at the design stage. Power for both designs is sensitive to the true event rate. If the event rate is mis-specified, both designs yield inappropriate powers to roughly the same extent. There is remarkably little difference between the normal approximation and Farrington-Manning’s method with respect to the relative risk approach, whereas the power is constantly greater using Farrington-Manning’s method than the normal approximation in risk difference with adaptive margin, even in scenarios where both approaches controlled the Type I error at the nominal rate.

Table 14: Power at different true event rates in fixed designs with binary positive outcomes

	Relative Risk $H_0: \pi_T/\pi_C \leq 0.875$	Risk Difference w/ Adaptive Margin $H_0: \pi_C - \pi_T \geq (0.125 \times \pi_C)$
$\pi_T = \pi_C = 0.95$	0.99 (0.99)	0.99 (0.99)
$\pi_T = \pi_C = 0.9$	0.99 (0.98)	0.99 (0.99)

$\pi_T = \pi_C = 0.85$	0.92 (0.91)	0.93 (0.95)
$\pi_T = \pi_C = 0.8$	0.81 (0.80)	0.82 (0.86)
$\pi_T = \pi_C = 0.75$	0.70 (0.70)	0.70 (0.74)
$\pi_T = \pi_C = 0.7$	0.59 (0.58)	0.59 (0.64)
$\pi_T = \pi_C = 0.65$	0.50 (0.49)	0.49(0.54)
$\pi_T = \pi_C = 0.6$	0.42 (0.42)	0.40 (0.46)
$\pi_T = \pi_C = 0.55$	0.35 (0.35)	0.34 (0.39)

Note: the power is calculated using normal approximation (with the power in parentheses calculated from Farrington-Manning's test)

Similar results were observed for negative outcome as well, as shown in Table 15 below.

Table 15: Power at different true event rates in fixed designs with binary negative outcomes

	Relative Risk $H_0: \frac{\pi_T}{\pi_C} \leq 1.5$	Risk Difference w/ Adaptive Margin $H_0: \pi_C - \pi_T \geq (-0.5 \times \pi_C)$
$\pi_T = \pi_C = 0.02$	0.19 (0.23)	0.13 (0.25)
$\pi_T = \pi_C = 0.05$	0.46 (0.50)	0.52 (0.67)
$\pi_T = \pi_C = 0.08$	0.68 (0.70)	0.73 (0.85)
$\pi_T = \pi_C = 0.1$	0.79 (0.80)	0.84 (0.92)
$\pi_T = \pi_C = 0.12$	0.86 (0.87)	0.91 (0.96)
$\pi_T = \pi_C = 0.15$	0.94 (0.94)	0.96 (0.99)
$\pi_T = \pi_C = 0.18$	0.97 (0.97)	0.99 (0.99)
$\pi_T = \pi_C = 0.2$	0.99 (0.99)	0.99 (0.99)

Note: the power is calculated using normal approximation (with the power in parentheses calculated from Farrington-Manning's test)

4.4 Design with Blinded Two-Stage SSR

In general planning a study with a novel endpoint is difficult because of very limited experience with the endpoint and therefore very limited prior knowledge of the size of the event rates under the considered treatments. This means sample size calculations are based on assumptions and the risk of ending up with an inadequately low or high sample size is large. An obvious solution to this problem is the mid-course adjustment of the sample size. The design can be described as the following three-step procedure following the Wittes and Brittain¹² approach discussed for a superiority trial and applying it to the non-inferiority setting.

Table 16: General Procedures for blinded two-stage SSR with binary outcomes

Step 1: Estimate initial sample size $n_0 = n(\alpha, 1 - \beta, \pi^a, \delta)$ as if it is a fixed sample design
Step 2: Simulate the pilot $n_1 = \pi n_0$ subjects (e.g., $\pi=0.5$) per arm based on original assumptions
Step 3: Estimate pooled event rate (blinded) based on pilot (interim) data → consider pilot estimate of $\hat{\pi}$ as assumption of true event rate π
Step 4: Re-estimate the additional sample size needed $n_2 = \max(n_0, \hat{N}) - n_1$ <ul style="list-style-type: none"> • Based on the pooled event rate only: $\hat{N} = n(\alpha, 1 - \beta, \hat{\pi}_1, \delta)$ • Based on both the pooled event rate and adaptive non-inferiority margin: $\hat{N} = n(\alpha, 1 - \beta, \hat{\pi}_1, \hat{\delta})$
Step 5: Obtain the remaining observations based on the re-estimated sample size
Step 6: Final analysis with all $n_1 + n_2$ samples

In the situation considered here, for the initial sample size calculation the significance level α , the target power $1 - \beta$, the non-inferiority margin δ , and the event rates π_T and π_C need to be specified. Then the sample size formula (shown in Section 4.3.1) is used to calculate the required sample size for Farrington and Manning's test. After treatment and follow-up of a certain proportion of the initially estimated number of patients (for example 50%), the interim analysis and sample size review are carried out. Specifically, the overall pooled event rate is estimated from the interim data by $\hat{\pi}_{1\bullet} = x_{1\bullet}/n_{1\bullet}$, where $x_{1\bullet} = x_{1T} + x_{1C}$ and $n_{1\bullet} = n_{1T} + n_{1C}$ denoting the sample sizes and number of successes at the interim analysis by n_{1T} and x_{1T} and n_{1C} and x_{1C} for the experimental and control groups, respectively. Then the treatment event rates are estimated in a blinded fashion as

$$\pi_{1T} = \hat{\pi}_{1\bullet} - \frac{1}{2}\delta_1^a \quad \text{and} \quad \pi_{1C} = \hat{\pi}_{1\bullet} + \frac{1}{2}\delta_1^a$$

where δ_1^a is the assumed treatment difference under H_a . Since it is assumed that the treatment event rates are equal in a non-inferiority clinical trial, then the estimates for the event rates in the two treatment groups are usually assumed to be the same and to be equal to the observed overall event rate ($\hat{\pi}_{1T} = \hat{\pi}_{1C} = \hat{\pi}_{1\bullet}$). We then use the Farrington and Manning's test to estimate the new updated required sample size as follows. For The blinded event rate estimates $\hat{\pi}_{1T}$ and $\hat{\pi}_{1C}$ are computed and, in addition, constrained estimates $\tilde{\pi}_{1T}$ and $\tilde{\pi}_{1C}$ of the event rates required for the denominator of the test statistic are obtained by feeding $\hat{\pi}_{1T}$ and $\hat{\pi}_{1C}$ into the solutions to the cubic equations given by Farrington and Manning⁶³. The update on the required sample size is then obtained by replacing π_i and $\tilde{\pi}_i$ by π_{1i} and $\tilde{\pi}_{1i}$ in the equation (shown in Section 4.3.1) respectively for $i = T, C$. In the computations presented here the sample sizes for the second stage

were chosen as follows: if the updated sample size estimate for the sample size is smaller than the originally planned number of patients, the study will recruit the total number of patients as originally planned; otherwise the required additional number of patients will be recruited. Denoting the sample sizes and number of successes in the second stage of the study by n_{2i} and x_{2i} , $i = T, C$, respectively, $\hat{\pi}_i$ in the test statistics given above are replaced by $\hat{\pi}_{\bullet i} = (x_{1i} + x_{2i}) / (n_{1i} + n_{2i})$, $i = T, C$ in the statistical assessment of non-inferiority based on the final dataset. The standard errors for the Blackwelder test are based on $\hat{\pi}_{\bullet i}$ whereas the standard errors for Farrington and Manning's test are computed as the restricted MLEs under the null hypothesis based on sample sizes of $n_{1T} + n_{2T}$ and $n_{1C} + n_{2C}$, and numbers of successes of $x_{1T} + x_{2T}$ and $x_{1C} + x_{2C}$ for the experimental and control groups, respectively, according to the formulae given in Farrington and Manning⁶³.

The sample sizes n_{2i} of the second stage are functions of the overall event rate $\hat{\pi}_{1\bullet}$ and therefore random variables.

As discussed in Table 16, we will discuss the sample size recalculation based on the interim estimate of the pooled (blinded) event rate only and then based on both the interim estimate of the pooled event rate and adaptive non-inferiority margin at the interim stage, with the focus on the statistical and clinical implications of changing the absolute non-inferiority margin and sample size based on the interim estimate of pooled event rate, and then revising the margin again if necessary based on the estimate of the event rate in the control group at the end of the study.

4.4.1 Two Stage SSR based on pooled event rate at interim without changing the non-inferiority margin using the risk difference approach

Designs allowing midcourse adjustments to the sample size have been discussed intensely over the past years for non-inferiority from a risk difference approach^{38,46,43}. They generally dealt with the sample size re-estimation based on the interim estimate of the pooled event rate, without changing the pre-specified non-inferiority margin. Specifically, the same pre-specified non-inferiority margin is used at the initial design stage, and at the sample size re-estimation at the interim and the final analyses. Before evaluating the approach in Table 16, we will review the general algorithm here.

Example (continued)

Suppose the study is instead designed to allow for two-stage sample size re-estimation with margin fixed at design stage, interim, and final analysis. After 50% of the patients have been treated and followed, without unmasking the treatment assignments, the pooled event is estimated from the first-stage samples, and the updated sample size is calculated. Suppose the estimated event rate is 70%, then the required sample size is 328. Since this estimate reflects a smaller-than-anticipated event rate, the study would have been under-powered at the original sample size. At the end of the second stage, the hypothesis test is performed using data from both stages. It has close to the required power to claim non-inferiority with 10% of margin.

Table A4.7 in Appendix A displays the Type I error rate of a two-stage sample size estimation based solely on the interim estimate of the pooled event rate without updating non-inferiority margin using the risk difference approach. For both positive and negative outcomes, Type I error rates are in close agreement with their nominal value, $\alpha = 0.025$ using both Blackwelder's and Farrington-Manning's tests. These results are in general agreement with those reported by other authors and support Gould's claim that re-estimation of sample size using an interim estimate of the pooled event rate "does not materially affect the Type I error rates"¹⁴.

Table A4.8 in Appendix A displays the mean and standard deviation of re-estimated final sample size and the power of two-stage sample size estimation without updating non-inferiority margin using the risk difference approach. As would be expected, the final sample size is dependent on the true event rate, and sample size re-estimation helps to maintain the desired power in both positive and negative outcomes. Powers come close to the nominal value of 80% for both Blackwelder and Farrington-Manning tests.

4.4.2 Two-stage SSR based on pooled event rate at interim using the relative risk approach

There is no existing literature that discusses the sample size re-estimation using the relative risk approach. In this section, we will discuss the two-stage sample size re-estimation based on pooled event rate at interim using the relative risk approach.

Example (continued)

Suppose the study is instead designed to allow for two-stage sample size re-estimation of

the study with relative risk approach. After 50% of the patients have been treated and followed, without unmasking the treatment assignments, the pooled event is estimated from the first-stage samples, then the event rate for the control group is estimated under the null hypothesis, and the updated sample size is calculated. Suppose the estimated event rate is 70%, then an additional 156 samples are needed. Since this estimate reflects a smaller-than-anticipated event rate, the study would have been under-powered at the original sample size.

Table A4.9 in Appendix A displays Type I error rate of two-stage sample size estimation using the relative risk approach. For a positive outcome, Type I error rates are in close agreement with their nominal value, $\alpha = 0.025$ using both normal approximation and Farrington-Manning's tests. However, in cases with a negative outcome, the Type I error rate is inflated when the normal approximation is used. But the inflation of Type I error rate is not likely to be attributed to the sample size re-estimation procedure. Table 17 shows the Type I error rates for both procedures with and without sample size re-estimation in relative risk setting with a negative outcome. It indicates that the Type I error rates are above the nominal level regardless of sample size re-estimation. The Farrington-Manning method, where the variance is estimated based on restricted MLEs of \hat{p}_T and \hat{p}_C under the null hypothesis is shown to be superior to the normal approximation where p_T and p_C were estimated using observed value at interim⁶³. Together with the observations we have, we recommend use of the Farrington-Manning's test for relative risk approach.

Table 17: Type I error rates at different true event rate in relative risk design with binary negative outcomes

	Without SRR		With SSR	
	Normal Approximation	Farrington-Manning	Normal Approximation	Farrington-Manning
$\pi_C = 0.02$	0.0379	0.0257	0.0290	0.0251
$\pi_C = 0.05$	0.0305	0.0258	0.0286	0.0249
$\pi_C = 0.08$	0.0292	0.0254	0.0287	0.0251
$\pi_C = 0.1$	0.0286	0.0254	0.0284	0.0254
$\pi_C = 0.12$	0.0284	0.0254	0.0284	0.0253
$\pi_C = 0.15$	0.0278	0.0252	0.0277	0.0250
$\pi_C = 0.18$	0.0274	0.0252	0.0273	0.0252
$\pi_C = 0.2$	0.0274	0.0251	0.0272	0.0250

Table A4.10 in Appendix A displays the mean and standard deviation of re-estimated final sample size and the power of two-stage sample size estimation using relative risk approach. Again, power is greatly stabilized due to the sample size re-estimation procedure. Powers come close to the nominal value of 0.8 for both normal approximation and Farrington-Manning tests.

4.4.3 Two-stage SSR based on pooled event rate at interim using the risk difference approach -- margin was updated under H_a at interim

The results in Section 4.4.1 show that the sample size re-estimation with fixed margin is very effective in eliminating both under-powering and over-powering in non-inferiority trials. However, the clinical concerns over the fixed margin remains when analyzing the

data from a risk difference approach. Is it clinically acceptable to use the pre-specified non-inferiority margin if the true event rate is considerably different from the original assumption?

To solve the problem, we propose a sample size re-estimation procedure for the hypotheses to be tested below:

$$H_0: \pi_C - \pi_T \geq \delta$$

$$H_a: \pi_C - \pi_T < \delta$$

where $\delta = b\pi_C > 0$ is the pre-specified non-inferiority margin.

The procedure allows the non-inferiority margin to vary with the underlying event rates both at interim sample size re-estimation and at the time of final hypothesis testing.

Specifically, the first-stage data were used to estimate the overall pooled event rate $\hat{\pi}_{1\bullet} = x_{1\bullet}/n_{1\bullet}$, where $x_{1\bullet} = x_{1T} + x_{1C}$ and $n_{1\bullet} = n_{1T} + n_{1C}$ denoting the sample sizes and number of successes at the interim analysis by n_{1T} and x_{1T} and n_{1C} and x_{1C} for the experimental and control groups, respectively. Then the by-treatment event rates are estimated in a blinded fashion by

$$\pi_{1T} = \hat{\pi}_{1\bullet} - \frac{1}{2}\delta_1^a \quad \text{and} \quad \pi_{1C} = \hat{\pi}_{1\bullet} + \frac{1}{2}\delta_1^a$$

where δ_1^a is the assumed treatment difference under H_a . Since it is assumed that the treatments are equal, then the estimates for the event rates in the two treatment groups are the same and equal to the observed overall event rate ($\hat{\pi}_{1T} = \hat{\pi}_{1C} = \hat{\pi}_{1\bullet}$). We then recalculate the non-inferiority margin using the estimated pooled event rate, that is

$\hat{\delta} = b \times \hat{\pi}_{1\bullet}$. The new updated required sample size is obtained by replacing the original π with estimated event rate ($\hat{\pi}_{1\bullet}$) and using the updated non-inferiority margin ($\hat{\delta}$). Finally, the non-inferiority margin for final analysis was based on the event rate in the active control group, that is $b \times \hat{\pi}_C$.

Example (continued)

Suppose the study is instead designed to allow for two-stage sample size re-estimation of the study with an adaptive risk difference approach. Half-way through recruitment, without unmasking the treatment assignments, the pooled event is estimated from the first-stage samples, then the event rate for the control group is estimated under the null hypothesis, the non-inferiority margin is updated accordingly, and the updated sample size is calculated using the estimated pooled event rate and updated non-inferiority margin. Suppose the estimated event rate is 70% at interim, then the required total sample size is 375 patients per group. Since this estimate reflects a smaller-than-anticipated event rate, the study would have been under-powered at the original sample size. At the end of the second stage, the hypothesis test is performed using the newly established non-inferiority margin, which is estimated using the observed event rate in the active control group at the final stage.

Table A4.11 in Appendix A displays Type I error rate of two-stage sample size estimation with adaptive non-inferiority margin using the risk difference approach. In cases with positive outcomes, when the event rate in the active control group was underestimated at the design stage, the Type I error rate was slightly inflated when Farrington-

Manning's test was used. There is no inflation observed if Blackwelder's test was used. In cases with negative outcomes, Type I error rate is considerably inflated when the normal approximation test is used. The Type I error rate lies lower than the nominal ones when Farrington-Manning's test is used. As in the relative risk approach, we believe the inflation of Type I error rate is not due to the sample size re-estimation procedure. Table 18 shows the Type I error rates for both procedures with and without sample size re-estimation in adaptive margin setting with a positive outcome. It indicates that the Type I error rates are almost identical in sample size re-estimation procedures compared to their counterpart, where there is no sample size re-estimation.

Table 18: Type I error rates at different true event rates in risk difference with adaptive margin design with binary positive outcomes

	Without SRR		With SSR	
	Normal Approximation	Farrington-Manning	Normal Approximation	Farrington-Manning
$\pi_C = 0.95$	0.02033	0.03327	0.02124	0.03369
$\pi_C = 0.9$	0.02453	0.02955	0.02563	0.02949
$\pi_C = 0.85$	0.02267	0.02753	0.02523	0.02634
$\pi_C = 0.8$	0.02627	0.02656	0.02434	0.02617
$\pi_C = 0.75$	0.02439	0.02512	0.02409	0.02532
$\pi_C = 0.7$	0.02499	0.02571	0.02329	0.02379
$\pi_C = 0.65$	0.02508	0.02493	0.02445	0.02461
$\pi_C = 0.6$	0.02567	0.02481	0.02458	0.02437

Table 19 shows the Type I error rates for both procedures with and without sample size re-estimation in an adaptive margin setting with a negative outcome. It indicates that the

Type I error rates are above the nominal level when using Blackwelder's test with or without sample size re-estimation.

Table 19: Type I error rates at different true event rate in risk difference with adaptive margin design with binary negative outcomes

	Without SRR		With SSR	
	Normal Approximation	Farrington-Manning	Normal Approximation	Farrington-Manning
$\pi_C = 0.02$	0.03916	0.02512	0.0286	0.01915
$\pi_C = 0.05$	0.03241	0.02138	0.02876	0.01959
$\pi_C = 0.08$	0.03096	0.02055	0.02873	0.0198
$\pi_C = 0.1$	0.03012	0.02054	0.02806	0.01956
$\pi_C = 0.12$	0.02915	0.02037	0.02819	0.01901
$\pi_C = 0.15$	0.02918	0.02028	0.02749	0.01926
$\pi_C = 0.18$	0.02877	0.02054	0.0277	0.01978
$\pi_C = 0.2$	0.02836	0.02066	0.02683	0.0193

Table A4.12 in Appendix A displays the mean and standard deviation of re-estimated final sample size and power of two-stage sample size estimation with adaptive non-inferiority margin using a risk difference approach. As would be expected, the final sample size is dependent on the true event rate, and sample size re-estimation helps to maintain the desired power in both positive and negative outcomes. Powers come close to the nominal value of 0.8 for both Blackwelder's and Farrington-Manning's tests.

4.4.4 Two-stage SSR based on pooled event rate at interim using the risk difference approach -- margin was updated under H_0 at interim

The approach described in this section is quite similar to the one in Section 4.4.3. The only difference is that after estimation of the overall pooled event rate at interim, the event rate in the control group is estimated under H_0

$$\pi_{1T} = \hat{\pi}_{1\bullet} - \frac{1}{2}\delta_1^0 \quad \text{and} \quad \pi_{1C} = \hat{\pi}_{1\bullet} + \frac{1}{2}\delta_1^0$$

where δ_1^0 is the assumed treatment difference under H_0 .

Results shown in Table A4.14 in Appendix A indicate that hardly any additional sample is needed when the margin at interim was updated under the null hypothesis. Therefore, the procedure does not have a clear effect on the power functions. In other words, the re-estimation procedure can not appreciably relieve the dangers of over or under powering a study when the event rate in control group is estimated under the null hypothesis at the interim.

4.4.5 Comparison of two-stage SSR designs

As discussed earlier, the sample size re-estimation with fixed margin is very effective in eliminating both under-powering and over-powering in non-inferiority trials. However, the concerns over the fix margin remains. Is it clinically acceptable to use the same non-inferiority margin if the true event rate is considerably different from the original assumption? Also, we show that the procedure with adaptive margin using the risk difference approach (margin is updated under the alternative hypothesis) has no effect on maintaining the desired power. So the comparison in this section will focus on the relative risk approach and procedure with adaptive margin using the risk difference

approach (margin is updated under the null hypothesis). Table 20 shows the formulas for the final test statistics for both approaches side-by-side.

Table 20: Final test statistics in relative risk approach and procedure with adaptive margin using the risk difference approach

	Relative Risk $H_0: \pi_T/\pi_C \leq R$	Risk Difference w/ Adaptive Margin $H_0: \pi_C - \pi_T \geq (b \times \pi_C)$
Final test statistics	$z = (\hat{\pi}_T - R \times \hat{\pi}_C)/\widehat{SE}$	$z = ((1 - b)\hat{\pi}_C - \hat{\pi}_T)/\widehat{SE}$,
Estimation of SE	$\widehat{SE} = \sqrt{((\hat{\pi}_T(1 - \hat{\pi}_T) + R^2\hat{\pi}_C(1 - \hat{\pi}_C))/n_0)}$	$\widehat{SE} = \sqrt{((\hat{\pi}_T(1 - \hat{\pi}_T) + (1 - b)^2\hat{\pi}_C(1 - \hat{\pi}_C))/n_0)}$

Table 21 displays the Type I error rate for both procedures in cases with positive outcome. Both procedures control the Type I error rate at the nominal level for most cases. We recommend the Farrington-Manning's test should always be used for relative risk approach as it controls Type I error for both positive and negative outcomes.

Table 21: Type I error rates at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis) – Positive Outcome

	Relative Risk		Risk Difference with Adaptive Margin	
	Normal Approximation	Farrington-Manning	Normal Approximation	Farrington-Manning
$\pi_C = 0.95$	0.0205	0.0262	0.02124	0.03369
$\pi_C = 0.9$	0.0218	0.0242	0.02563	0.02949
$\pi_C = 0.85$	0.0236	0.0241	0.02523	0.02634
$\pi_C = 0.8$	0.0240	0.0243	0.02434	0.02617

$\pi_c = 0.75$	0.0237	0.0236	0.02409	0.02532
$\pi_c = 0.7$	0.0247	0.0244	0.02329	0.02379
$\pi_c = 0.65$	0.0246	0.0243	0.02445	0.02461
$\pi_c = 0.6$	0.0253	0.0248	0.02458	0.02437

Table 22 displays the Type I error rate for both procedures in cases with negative outcome. We recommend the Farrington-Manning's test should always be used for relative risk approach as it controls Type I error for both positive and negative outcomes.

Table 22: Type I error rates at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis) – Negative Outcome

	Relative Risk		Risk Difference with Adaptive Margin	
	Normal Approximation	Farrington-Manning	Normal Approximation	Farrington-Manning
$\pi_c = 0.02$	0.02898	0.02513	0.0286	0.01915
$\pi_c = 0.05$	0.02859	0.02491	0.02876	0.01959
$\pi_c = 0.08$	0.02867	0.02507	0.02873	0.0198
$\pi_c = 0.1$	0.02837	0.02537	0.02806	0.01956
$\pi_c = 0.12$	0.02842	0.02532	0.02819	0.01901
$\pi_c = 0.15$	0.02768	0.02495	0.02749	0.01926
$\pi_c = 0.18$	0.02733	0.02519	0.0277	0.01978
$\pi_c = 0.2$	0.0272	0.02502	0.02683	0.0193

Table 23 displays the values for power for both procedures in cases with positive outcome. Both procedures greatly achieve the desired power.

Table 23: Power at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis) – Positive Outcome

	Relative Risk		Risk Difference with Adaptive Margin	
	Normal Approximation	Farrington-Manning	Normal Approximation	Farrington-Manning
$\pi_c = 0.95$	0.99996	0.99988	0.99998	0.99998
$\pi_c = 0.9$	0.98774	0.98342	0.99261	0.99302
$\pi_c = 0.85$	0.92234	0.9142	0.94402	0.94755
$\pi_c = 0.8$	0.8401	0.83102	0.85612	0.8606
$\pi_c = 0.75$	0.81282	0.80856	0.8023	0.80778
$\pi_c = 0.7$	0.80586	0.80332	0.79819	0.80374
$\pi_c = 0.65$	0.80038	0.79974	0.79646	0.80098
$\pi_c = 0.6$	0.80456	0.8049	0.7995	0.80242

Table 24 displays the power for both procedures in cases with negative outcome. Both procedures greatly achieve the desired power.

Table 24: Type I error rates at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis) – Negative Outcome

	Relative Risk		Risk Difference with Adaptive Margin	
	Normal Approximation	Farrington-Manning	Normal Approximation	Farrington-Manning

$\pi_c = 0.02$	0.79122	0.80644	0.84562	0.81089
$\pi_c = 0.05$	0.7865	0.80268	0.84528	0.81167
$\pi_c = 0.08$	0.7858	0.8021	0.84448	0.81185
$\pi_c = 0.1$	0.8034	0.81698	0.84286	0.81154
$\pi_c = 0.12$	0.86238	0.87146	0.8423	0.81233
$\pi_c = 0.15$	0.93394	0.93998	0.84686	0.81843
$\pi_c = 0.18$	0.97256	0.97536	0.88907	0.86859
$\pi_c = 0.2$	0.9859	0.9872	0.92357	0.9089

Table 25 displays the final sample size for both procedures in cases with both positive and negative outcomes. The final sample sizes are very close to for positive outcome. The relative risk approach show a slight advantage in the final sample size for negative outcome.

Table 25: Final sample size at different true event rate in relative risk approach and procedure with adaptive margin using the risk difference approach (margin is updated under the null hypothesis)

Positive Outcome			Negative Outcome		
	Relative Risk	Risk Diff w/ Adaptive Margin		Relative Risk	Risk Diff. w/ Adaptive Margin
$\pi_c = 0.95$	231	255	$\pi_c = 0.02$	5058	6103
$\pi_c = 0.9$	231	255	$\pi_c = 0.05$	1879	2207
$\pi_c = 0.85$	231	255	$\pi_c = 0.08$	1127	1309
$\pi_c = 0.8$	246	263	$\pi_c = 0.1$	914	1016
$\pi_c = 0.75$	304	295	$\pi_c = 0.12$	871	822

$\pi_c = 0.7$	387	328	$\pi_c = 0.15$	869	640
$\pi_c = 0.65$	483	350	$\pi_c = 0.18$	869	584
$\pi_c = 0.6$	597	374	$\pi_c = 0.2$	869	581

In summary, the methods with sample size re-estimation greatly achieve the desired power while controlling the Type I error.

Farrington-Manning's test should always be used for relative risk approach as the normal approximation method inflates Type I error significantly with or without sample size re-estimation.

There is remarkably little difference between the performance of the relative risk approach and procedures with adaptive margin using the risk difference approach (margin is updated under the null hypothesis), and neither dominates with respect to final power. The relative risk approach offer slight sample size reduction for cases with negative outcome. The findings from the comparisons imply that the procedure with adaptive non-inferiority margin using the risk difference approach (margin is updated under the null hypothesis) works at least as well as the relative risk procedures.

Chapter 5: SUMMARY AND DISCUSSION

5.1 Two-stage SSR in normally distributed outcome

We have presented several two-stage sample size re-estimation procedures for non-inferiority trials with a continuous outcome. The procedures include the sample size re-estimation based on the updated variance only, and those that are based on both updated variance and treatment difference in a conditional power calculation. The procedure also generalizes the traditional two-stage sample size re-estimation procedure by allowing for early stopping for futility at the interim analysis. We have derived the conditional power formulae for non-inferiority trials, expected Type I error rate, necessary second-stage sample size and critical value to adjust the rejection region for the final test statistic.

Figure 3 highlights the performance of each procedure discussed in terms of Type I error rate and power for non-inferiority trials with a continuous outcome.

Figure 3: Performance of each procedure discussed in Chapter 3

Section	Design		Table	Conclusion
3.3	Fixed Design	Type I Error	3.1	When the variance was under-estimated, the study was under-powered
		Power	3.2	When the variance or treatment difference was under-estimated, the study was under-powered

to maintain the required power,
following methods were proposed

Figure 3 (continued)

Section	Design		Table	Conclusion	
				Blinded	Unblinded
3.4	SSR Design -- Update variance only	Type I Error	3.3a 3.3b	Type I error was not affected. There is not much gain in conducting SSR unblinded.	
		Power	3.4a 3.4b	No noticeable difference between blinded and unblinded estimate. Study under-powered if treatment difference was mis-specified.	
3.6.1 3.7.1	SSR Design -- Update variance and treatment difference without critical value adjustment (no early stopping)	Type I Error	3.5 3.13	Type I error rate was inflated when there is no critical value adjustment	
		Power	3.6 3.14	Study was under-powered if the true treatment difference was under-estimated at the original design	Study was over-powered, the method calls for unnecessary sample size increase if the true variance or treatment difference was not far from original assumptions
3.6.2 3.7.2	SSR Design -- Update variance and treatment difference without critical value adjustment (early	Type I Error	3.7 3.15	Type I error rate was inflated when there is no critical value adjustment. allowing for early stopping for futility causes a small reduction in the type I error rate.	
		Power	3.8 3.16	Design that allows early stopping for futility have power at slightly lower levels than its counterpart without early futility stopping.	

To fix the inflated type I error,
following methods were proposed



Figure 3 (continued)

Method	Design		Table	Conclusion	
				using theta at design stage for CP calculation	using theta hat for CP calculation
3.6.3 3.7.3	SSR Design -- Update variance and treatment difference with critical value adjustment (no early stopping)	Type I Error	3.9 3.17	Type I error rate was controlled	
		Power	3.10 3.18	Desired power was not maintained when treatment difference was mis-specified	over-powered?
3.6.4 3.7.4	SSR Design -- Update variance and treatment difference with critical value adjustment (early	Type I Error	3.11 3.19	Type I error rate was controlled	
		Power	3.12 3.20	Power was reasonably maintained with reasonable sample size increase	

In summary, in the procedure where only the variance drives the re-estimation process, there is no accounting for the fact that the first-stage data will contribute to the final test statistic, therefore the Type I error rate was not inflated. However, in the procedure in which the estimated treatment difference was involved for sample size re-estimation, it is necessary to protect the Type I error rate due to the dependence of the sample size on z_1 , the first-stage test statistic. The method that we derived to adjust the rejection region for the final test statistic seems to work well at dealing with this issue.

The primary goal of sample size re-estimation is to protect study power through interim analysis. This goal is greatly achieved by the procedures discussed.

Additionally, the design with futility stopping shows the added benefit of reduced sample size.

5.2 Two-stage SSR in binomially distributed outcome

The chief purpose of blinded sample size re-estimation is to mitigate the effect of false assumptions about nuisance parameters on the power of a trial. For non-inferiority trials, the fixed-size versions of Blackwelder's test and Farrington and Manning's test for both risk difference and relative risk approaches can lead to seriously inappropriate sample sizes or power levels for the test.

We have presented several two-stage sample size re-estimation procedures for non-inferiority trials with a binary outcome. The procedures include the sample size re-estimation based on the risk difference with fixed margin, relative risk, and risk difference with adaptive margin.

Figure 4 highlights the performance of each procedure discussed in terms of Type I error rate and power.

Figure 4: Performance of each procedure discussed in Chapter 4

Section	Design		Table	Conclusion	
				Positive Outcome	Negative Outcome
4.3.1	Fixed Design -- Risk Difference Approach (Fixed margin and no IA)	Type I Error	4.1	not much difference between FM and Blackwelder, similar results for both positive and negative outcomes	
		Power	4.2	FM and Blackwelder are quite comparable. Power is sensitive to event rates, the trial can be easily under-powered or over-powered depends on the true event rate	

In order to fix the margin problem, following methods were proposed

Section	Design		Table	Conclusion	
				Positive Outcome	Negative Outcome
4.3.2	Fixed Design -- Relative Risk Approach (no IA)	Type I Error	4.3	Type I error rate from both methods is around the nominal level with some fluctuation	Type I error rate is significantly inflated using normal approximation. FM method is fine
		Power	4.4	When control event rate is under-estimated, the study is under powered. Both FM and normal approximation give similar results	
4.3.3	Fixed Design -- Risk Difference Approach (Adaptive margin based on \hat{p}_c at the end of trial and no IA)	Type I Error	4.5	Blackwelder is closer to nominal level; FM is a little bit inflated for some cases	Type I error rate is significantly inflated using normal approximation. FM method is fine
		Power	4.6	When control event rate is under-estimated, the study is under powered. Both FM and normal approximation give similar results	

To maintain the designed power, following methods were proposed

Figure 4 (continued)

Section	Design	Table	Conclusion		
			Positive Outcome	Negative Outcome	
4.4.1	SSR Design -- Risk Difference Approach with Fixed Margin (with blinded IA)	Type I Error	4.7	Type I error was well controlled for both FM and Blackwelder methods	
		Power	4.8	Desired power was maintained	
4.4.2	SSR Design -- Relative Risk Approach (with blinded IA)	Type I Error	4.9	Type I error was well controlled for both FM and Blackwelder methods	Similar observation as in design without IA. Type I error rate is significantly inflated using normal approximation. FM method is fine
		Power	4.10	Desired power was maintained	
4.4.3	SSR Design -- Risk Difference Approach (Estimated \hat{p}_c and update margin under H_a at IA, then update margin again at final based on \hat{p}_c)	Type I Error	4.11	Similar observation as in design without IA. Blackwelder is closer to nominal level; FM is a little bit inflated for some cases	Similar observation as in design without IA. FM controls type I error rate; Blackwelder inflates type I error
		Power	4.12	Desired power was maintained	
4.4.4	SSR Design -- Risk Difference Approach (Estimated \hat{p}_c and update margin under H_o at IA, then update margin again at final based on \hat{p}_c)	Type I Error	4.13	Type I error was well controlled for both FM and Blackwelder methods	Blackwelder controls type I error rate; FM inflates type I error
		Power	4.14	Desired power was NOT maintained	

As our investigations show, the proposed two-stage sample size re-estimation procedures are extremely effective in eliminating both under-powering and over-powering in non-inferiority trials. In addition, these decision rules are not affected by treatment differences because there is no un-blinding. Although the Type I error rate may be affected by changing the strategy before un-blinding, simulation studies suggest that the effect is unlikely to be material.

Farrington-Manning's test should always be used for relative risk approach as the normal approximation method inflates Type I error significantly with or without sample size re-estimation.

There is remarkably little difference between the performance of the relative risk approach and procedures with adaptive margin using the risk difference approach (margin is updated under the null hypothesis), and neither dominates with respect to final power. The relative risk approach offers slight sample size reduction for cases with negative outcome. The findings from the comparisons imply that the procedure with adaptive non-inferiority margin using the risk difference approach (margin is updated under the null hypothesis) works at least as well as the relative risk procedures.

We feel the proposed adaptive margin approach has a number of appealing properties. First, risk difference, which we believe is a commonly used effect measure in practice. Second, it eases the concern about the validity of the fixed non-inferiority margin if the observed event rate is smaller or larger than assumed. Third, an adaptive margin approach allows the experimenter to use realistic sample sizes.

5.3 Limitation and future directions

The thesis discussed the sample size re-estimation at the interim stage of the clinical trial mathematically, it is not necessarily advocating that it has to be done all the time. The necessity and implications of the procedure on Type I error rate and expected power need to be carefully scrutinized by means of simulations on a case-by-case basis.

The proposed adjusted critical value for continuous endpoint allows the critical value to be less than originally planned z_α , this might raise some concerns on inflating type I error rate.

As a practical consideration, the second stage sample size was restricted so that it could not exceed double the originally designed sample size. Future work should make sure capping does not related to Type I error rate.

Investigating the choice of an appropriate target for the conditional power is subject to future research. While it may seem natural to choose the sample size to fix this conditional power at $(1 - \beta)$ regardless of the value of the first-stage test statistic, it can be argued that it is desirable to fix the conditional power above $(1 - \beta)$ when the value of the first-stage test statistic is large, since this may reflect a noteworthy underlying treatment difference that you would want to ensure it is not missed.

In addition to the choice of conditional power cutoff, important future work should include the incorporation of smart futility stopping boundary for the first-stage analysis. This can save much time and many resources when little or no effect is present in a study.

The resources saved could then be allocated to other promising studies and further important clinical research.

The findings for binomially distributed endpoint in Chapter 4 suggest that the procedure with adaptive non-inferiority margin using the risk difference approach (margin is updated under the null hypothesis) works at least as well as the relative risk approach. The users must decide a priori whether relative risk or risk difference approach will be taken.

We did not constrain the sample size to any maximum sample size as we did for continuous endpoint. In practice, however, especially with event rates closer to 0 or 1, sample sizes may be bounded above by practical constraints like budget, availability of patients, or drug supply.

Chapter 6: CONCLUSION

Methods for re-estimating the sample size for clinical trials was developed in early 90s. Extensive research has been done since then in the context of the superiority trial. Much less literature exists about sample size recalculation procedures for non-inferiority trials. We focus our research on re-estimating the sample size for non-inferiority trials with a continuous or binary outcome.

For continuous endpoints, current approaches to SSR for non-inferiority trials focus on estimating the variance (blinded or unblinded) at the interim and updating the sample size based solely on this estimated variance. The SSR using both sample variance and the observed treatment difference at interim in conditional power calculation is used in superiority trials. We extend the methodology to non-inferiority trials, quantify the effect on Type I error rate, derive the conditional power formulae for non-inferiority trial, the second-stage sample size and propose controlling it by modifying the critical value and/or stopping the trial at the interim for futility.

For binary endpoints, current approaches to SSR for non-inferiority trials focus on estimating the event rates (blinded or unblinded) at the interim and updating the sample size solely on the estimated event rates at the interim with fixed non-inferiority margin. We propose a procedure that adapts both the non-inferiority margin, and sample size based on the underlying interim observed pooled (blinded) event rate, and re-updates non-inferiority margin again at the final analysis based on the observed estimate of event rate in the control group at the end of the study.

Simulation results show the proposed adaptive procedures for extending a study by adding sample size are effective ways to help stabilize study power while preserving the Type I error so that important questions of interest can be explored accurately and dependably. Combining sample size re-estimation methods with early stopping rules for continuous endpoint and adapting non-inferiority margin for binary endpoint could increase study flexibility, scope, and efficiency of non-inferiority trials.

The proposed methodologies can be used for designing an efficient two-stage non-inferiority trial with sample size re-estimation in active controlled non-inferiority clinical trials. However, sample size re-estimation should not substitute for careful planning of a trial. In addition, the implications of the procedure on Type I error rate and expected power need to be carefully scrutinized by means of simulations on a case-by-case basis as each study is unique.

While the proposed sample size re-estimation with adaptive margin in trials with binary outcome effectively preserves the intended power, consideration has to be given to evaluate the clinical relevance of the updated non-inferiority margin estimated for interim and final analyses. Furthermore, whatever the choice of effect measure is, the margin needs to be pre-specified in the study protocol.

Appendix A: TABLES

Table A3.1: Type I error rate at different common unknown variances with fixed design

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/ group$. Simulation runs=100,000

True $\mu_C - \mu_T$	True σ^2	Type I error Rate
0.2	1	0.02508
0.2	1.2	0.02508
0.2	1.5	0.02508
0.2	1.7	0.02508
0.2	2.0	0.02508

Table A3.2: Power at various common unknown variances and treatment difference with fixed design

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=10,000

True $\mu_C - \mu_T$	True σ^2	Power
0	1	0.798
	1.2	0.6492
	1.5	0.4645
	1.7	0.381
	2.0	0.2901
0.02	1	0.7149
	1.2	0.558
	1.5	0.393
	1.7	0.3184
	2.0	0.2417
0.05	1	0.558
	1.2	0.4186
	1.5	0.2901
	1.7	0.234
	2.0	0.1831
0.07	1	0.4451
	1.2	0.3308
	1.5	0.2275
	1.7	0.1881
	2.0	0.1504
0.1	1	0.2901
	1.2	0.2141
	1.5	0.1568
	1.7	0.1311
	2.0	0.1093

Table A3.3a: Type I error rate at different common unknown variances when only variance was updated at the interim – Blinded estimate of variance

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=100,000

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	Type I error Rate
0.2	1	406 \pm 19	0.02441
0.2	1.2	570 \pm 41	0.02557
0.2	1.5	888 \pm 64	0.02545
0.2	1.7	1139 \pm 82	0.02504
0.2	2.0	1575 \pm 113	0.02454

Table A3.3b: Type I error rate at different common unknown variances when only variance was updated at the interim – Unblinded estimate of variance

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=100,000

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	Type I error Rate
0.2	1	404 \pm 17	0.02445
0.2	1.2	566 \pm 41	0.02557
0.2	1.5	884 \pm 63	0.02542
0.2	1.7	1135 \pm 82	0.02539
0.2	2.0	1571 \pm 113	0.02438

Table A3.4a: Power at different common unknown variances when only variance was updated at the interim – Blinded estimate of variance

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/\text{group}$. Simulation runs=10,000

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size ± SD	Power
0	1	404±17	0.8107
	1.2	566±41	0.8033
	1.5	884±63	0.8052
	1.7	1135±81	0.8023
	2.0	1571±113	0.7991
0.02	1	404±17	0.7282
	1.2	566±41	0.7177
	1.5	884±63	0.7164
	1.7	1135±81	0.7163
	2.0	1571±113	0.7131
0.05	1	404±17	0.5662
	1.2	566±41	0.5633
	1.5	884±63	0.5595
	1.7	1135±81	0.5555
	2.0	1571±113	0.5563
0.07	1	404±17	0.4541
	1.2	566±41	0.4463
	1.5	884±63	0.4491
	1.7	1135±81	0.4416
	2.0	1571±113	0.4455
0.1	1	404±17	0.294
	1.2	566±41	0.2862
	1.5	884±63	0.2869
	1.7	1135±81	0.2802
	2.0	1571±113	0.2922

Table A3.4b: Power at different common unknown variances when only variance was updated at the interim – Unblinded estimate of variance

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/\text{group}$. Simulation runs=10,000

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size ± SD	Power
0	1	404±17	0.8109
	1.2	566±41	0.8034
	1.5	884±63	0.805
	1.7	1135±82	0.803
	2.0	1571±113	0.7981
0.02	1	404±17	0.7277
	1.2	566±41	0.7169
	1.5	884±63	0.7152
	1.7	1135±82	0.7175
	2.0	1571±113	0.7128
0.05	1	404±17	0.5656
	1.2	566±41	0.5632
	1.5	884±63	0.5588
	1.7	1135±82	0.5534
	2.0	1571±113	0.5566
0.07	1	404±17	0.455
	1.2	566±41	0.4466
	1.5	884±63	0.4499
	1.7	1135±82	0.441
	2.0	1571±113	0.4462
0.1	1	404±17	0.2939
	1.2	566±41	0.2853
	1.5	884±63	0.2873
	1.7	1135±82	0.2824
	2.0	1571±113	0.2921

Table A3.5a: Type I error rate at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment

(early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/ \text{group}$. Simulation runs=100,000

The second stage sample size was capped at **2×original fixed sample size**

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Type I error Rate
0.2	1	691 \pm 163	5%	0.03105
0.2	1.2	925 \pm 126	69%	0.03129
0.2	1.5	982 \pm 37	99%	0.02629
0.2	1.7	983 \pm 26	100%	0.02584
0.2	2.0	984 \pm 19	100%	0.02556

Table A3.5b: Type I error rate at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment

(early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/ \text{group}$. Simulation runs=100,000

The second stage sample size was capped at **5×original fixed sample size**

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Type I error Rate
0.2	1	695 \pm 170	0%	0.03107
0.2	1.2	1121 \pm 288	0%	0.03216
0.2	1.5	2018 \pm 267	61%	0.02838
0.2	1.7	2155 \pm 91	98%	0.02577
0.2	2.0	2162 \pm 57	100%	0.0253

Table A3.6: Power at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=10,000

The second stage sample size was capped at 2×original fixed sample size

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Power
0	1	444±93	0%	0.88701
	1.2	660±227	15%	0.90478
	1.5	929±166	87%	0.83568
	1.7	960±118	96%	0.74105
	2.0	974±80	98%	0.60209
0.02	1	460±105	0%	0.82369
	1.2	693±227	19%	0.84926
	1.5	941±148	89%	0.75462
	1.7	966±104	97%	0.65071
	2.0	976±71	99%	0.51484
0.05	1	488±122	0%	0.689
	1.2	743±222	26%	0.71773
	1.5	955±123	93%	0.60033
	1.7	972±87	98%	0.49929
	2.0	979±58	99%	0.38496
0.07	1	510±133	0%	0.57363
	1.2	774±215	31%	0.60456
	1.5	962±108	94%	0.48501
	1.7	975±76	98%	0.397
	2.0	980±51	99%	0.30483
0.1	1	547±146	1%	0.38992
	1.2	819±200	40%	0.4112
	1.5	970±88	96%	0.31886
	1.7	978±61	99%	0.2593
	2.0	982±41	100%	0.20072

Table A3.7: Type I error rate at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment

(allowing early stopping for futility) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / \text{group}$. Simulation runs=100,000

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Type I error Rate
0.2	1.0	10%	480 \pm 214	31%	0.02992
0.2		20%	389 \pm 198	48%	0.02841
0.2		30%	330 \pm 172	60%	0.02696
0.2	1.2	10%	510 \pm 366	55%	0.02718
0.2		20%	371 \pm 285	71%	0.02359
0.2		30%	299 \pm 217	81%	0.02043
0.2	1.5	10%	615 \pm 708	73%	0.01703
0.2		20%	399 \pm 499	85%	0.01262
0.2		30%	305 \pm 357	91%	0.00949
0.2	1.7	10%	737 \pm 1071	79%	0.0122
0.2		20%	449 \pm 739	89%	0.00824
0.2		30%	325 \pm 518	94%	0.00583
0.2	2.0	10%	1017 \pm 1908	84%	0.00819
0.2		20%	564 \pm 1292	92%	0.00509
0.2		30%	377 \pm 894	96%	0.00345

Table A3.8: Power at different common unknown variances when only variance was updated at the interim using CP without critical value adjustment (allowing early stopping for futility) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=10,000

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0	1	0.1	440 \pm 89	1%	0.88134
		0.2	432 \pm 85	2%	0.8695
		0.3	420 \pm 82	4%	0.85089
0.02		0.1	452 \pm 100	1%	0.81557
		0.2	439 \pm 96	4%	0.79889
		0.3	424 \pm 94	7%	0.77663
0.05		0.1	472 \pm 116	2%	0.67707
		0.2	450 \pm 114	6%	0.65696
		0.3	426 \pm 112	11%	0.63036
0.07		0.1	484 \pm 127	4%	0.55977
		0.2	455 \pm 126	9%	0.53907
		0.3	425 \pm 125	15%	0.51309
0.1		0.1	500 \pm 144	7%	0.3772
		0.2	457 \pm 146	15%	0.35999
		0.3	416 \pm 144	24%	0.33815
0	1.2	0.1	610 \pm 236	6%	0.85365
		0.2	553 \pm 235	14%	0.7872
		0.3	493 \pm 223	22%	0.71296
0.02		0.1	626 \pm 247	9%	0.78728
		0.2	556 \pm 248	18%	0.71262
		0.3	488 \pm 235	27%	0.63405
0.05		0.1	640 \pm 264	13%	0.64781
		0.2	550 \pm 267	25%	0.56999
		0.3	470 \pm 249	36%	0.49251
0.07		0.1	641 \pm 278	17%	0.5361
		0.2	538 \pm 279	30%	0.4639
		0.3	452 \pm 256	42%	0.39421
0.1		0.1	631 \pm 300	24%	0.35716
		0.2	511 \pm 294	39%	0.30217
		0.3	421 \pm 261	52%	0.25273

Table A3.8 (continued):

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0	1.5	0.1	741 \pm 345	24%	0.67537
		0.2	620 \pm 375	39%	0.55081
		0.3	523 \pm 370	52%	0.44498
0.02		0.1	720 \pm 357	28%	0.59477
		0.2	591 \pm 380	44%	0.47572
		0.3	494 \pm 367	57%	0.37849
0.05		0.1	679 \pm 374	35%	0.45648
		0.2	544 \pm 381	52%	0.35709
		0.3	448 \pm 356	64%	0.27676
0.07		0.1	646 \pm 382	40%	0.36143
		0.2	511 \pm 378	57%	0.2794
		0.3	418 \pm 345	69%	0.21412
0.1		0.1	592 \pm 389	48%	0.23031
		0.2	459 \pm 365	65%	0.17559
		0.3	374 \pm 322	76%	0.13309
0	1.7	0.1	678 \pm 378	36%	0.53352
		0.2	544 \pm 384	53%	0.40627
		0.3	448 \pm 360	65%	0.30877
0.02		0.1	649 \pm 385	40%	0.45621
		0.2	513 \pm 381	58%	0.34151
		0.3	420 \pm 349	69%	0.25559
0.05		0.1	601 \pm 390	47%	0.33797
		0.2	466 \pm 370	64%	0.24816
		0.3	381 \pm 329	75%	0.18252
0.07		0.1	567 \pm 390	52%	0.26265
		0.2	436 \pm 359	69%	0.19071
		0.3	425 \pm 125	15%	0.51309
0.1		0.1	517 \pm 385	59%	0.16661
		0.2	394 \pm 338	74%	0.11921
		0.3	323 \pm 285	83%	0.08605
0	2	0.1	579 \pm 391	50%	0.37154
		0.2	447 \pm 363	67%	0.26123
		0.3	365 \pm 318	77%	0.18575
0.02		0.1	552 \pm 389	54%	0.31077
		0.2	422 \pm 353	71%	0.21531
		0.3	345 \pm 304	80%	0.15227

Table A3.8 (continued)

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0.05	2	0.1	509 \pm 383	60%	0.22404
		0.2	387 \pm 334	75%	0.15333
		0.3	318 \pm 281	84%	0.10679
0.07		0.1	480 \pm 376	64%	0.17395
		0.2	364 \pm 320	78%	0.1178
		0.3	302 \pm 264	86%	0.08159
0.1		0.1	439 \pm 362	69%	0.11166
		0.2	334 \pm 296	82%	0.07452
		0.3	280 \pm 239	89%	0.05094

Table A3.9a: Type I error rate at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/ group$. Simulation runs=100,000

The second stage sample size was capped at **2×original fixed sample size**

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Type I error Rate
0.2	1	691±163	5%	0.0253
0.2	1.2	925±126	69%	0.02583
0.2	1.5	982±37	99%	0.02515
0.2	1.7	983±26	100%	0.02514
0.2	2.0	984±19	100%	0.02514

Table A3.9b: Type I error rate at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/ group$. Simulation runs=100,000

The second stage sample size was capped at **5×original fixed sample size**

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Type I error Rate
0.2	1	695±170	0%	0.0253
0.2	1.2	1121±288	0%	0.02573
0.2	1.5	2018±267	61%	0.02528
0.2	1.7	2155±91	98%	0.02468
0.2	2.0	2162±57	100%	0.02471

Table A3.10: Power at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/ \text{group}$. Simulation runs=10,000

The second stage sample size was capped at 2×original fixed sample size

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Power
0	1	444±93	0%	0.87327
	1.2	660±227	15%	0.8854
	1.5	929±166	87%	0.79469
	1.7	960±118	96%	0.6943
	2.0	974±80	98%	0.55636
0.02	1	460±105	0%	0.80062
	1.2	693±227	19%	0.81703
	1.5	941±148	89%	0.70843
	1.7	966±104	97%	0.60356
	2.0	976±71	99%	0.47429
0.05	1	488±122	0%	0.65116
	1.2	743±222	26%	0.66543
	1.5	955±123	93%	0.55255
	1.7	972±87	98%	0.45805
	2.0	979±58	99%	0.35086
0.07	1	510±133	0%	0.52866
	1.2	774±215	31%	0.5424
	1.5	962±108	94%	0.44233
	1.7	975±76	98%	0.36163
	2.0	980±51	99%	0.27821
0.1	1	547±146	1%	0.3447
	1.2	819±200	40%	0.3529
	1.5	970±88	96%	0.28727
	1.7	978±61	99%	0.23601
	2.0	982±41	100%	0.18384

Table A3.11: Type I error rate at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment

(allowing early stopping for futility) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=100,000

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Type I error Rate
0.2	1.0	10%	480 \pm 214	31%	0.02531
0.2		20%	389 \pm 198	48%	0.02536
0.2		30%	330 \pm 172	60%	0.02545
0.2	1.2	10%	510 \pm 366	55%	0.02584
0.2		20%	371 \pm 285	71%	0.02581
0.2		30%	299 \pm 217	81%	0.0258
0.2	1.5	10%	615 \pm 708	73%	0.02536
0.2		20%	399 \pm 499	85%	0.02538
0.2		30%	305 \pm 357	91%	0.025
0.2	1.7	10%	737 \pm 1071	79%	0.02537
0.2		20%	449 \pm 739	89%	0.02482
0.2		30%	325 \pm 518	94%	0.02489
0.2	2.0	10%	1017 \pm 1908	84%	0.02512
0.2		20%	564 \pm 1292	92%	0.02465
0.2		30%	377 \pm 894	96%	0.02464

Table A3.12: Power at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (allowing early stopping for futility) – Assuming $\mu_C - \mu_T = 0$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / \text{group}$. Simulation runs=10,000

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0	1	0.1	440 \pm 89	1%	0.86987
		0.2	432 \pm 85	2%	0.8608
		0.3	420 \pm 82	4%	0.84503
0.02		0.1	452 \pm 100	1%	0.79633
		0.2	439 \pm 96	4%	0.78481
		0.3	424 \pm 94	7%	0.76726
0.05		0.1	472 \pm 116	2%	0.64652
		0.2	450 \pm 114	6%	0.63473
		0.3	426 \pm 112	11%	0.61629
0.07		0.1	484 \pm 127	4%	0.52453
		0.2	455 \pm 126	9%	0.51407
		0.3	425 \pm 125	15%	0.4971
0.1		0.1	500 \pm 144	7%	0.34223
		0.2	457 \pm 146	15%	0.33511
		0.3	416 \pm 144	24%	0.32274
0	1.2	0.1	610 \pm 236	6%	0.84942
		0.2	553 \pm 235	14%	0.79046
		0.3	493 \pm 223	22%	0.72163
0.02		0.1	626 \pm 247	9%	0.77821
		0.2	556 \pm 248	18%	0.7155
		0.3	488 \pm 235	27%	0.64439
0.05		0.1	640 \pm 264	13%	0.63052
		0.2	550 \pm 267	25%	0.57104
		0.3	470 \pm 249	36%	0.50469
0.07		0.1	641 \pm 278	17%	0.5138
		0.2	538 \pm 279	30%	0.46341
		0.3	452 \pm 256	42%	0.40681
0.1		0.1	631 \pm 300	24%	0.33492
		0.2	511 \pm 294	39%	0.30181
		0.3	421 \pm 261	52%	0.26451

Table A3.12 (continued):

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0	1.5	0.1	741 \pm 345	24%	0.68429
		0.2	620 \pm 375	39%	0.57213
		0.3	523 \pm 370	52%	0.47478
0.02		0.1	720 \pm 357	28%	0.60744
		0.2	591 \pm 380	44%	0.503
		0.3	494 \pm 367	57%	0.41382
0.05		0.1	679 \pm 374	35%	0.47363
		0.2	544 \pm 381	52%	0.39161
		0.3	448 \pm 356	64%	0.31835
0.07		0.1	646 \pm 382	40%	0.38088
		0.2	511 \pm 378	57%	0.31498
		0.3	418 \pm 345	69%	0.25696
0.1		0.1	592 \pm 389	48%	0.251
		0.2	459 \pm 365	65%	0.21007
		0.3	374 \pm 322	76%	0.17345
0	1.7	0.1	678 \pm 378	36%	0.55443
		0.2	544 \pm 384	53%	0.43983
		0.3	448 \pm 360	65%	0.35044
0.02		0.1	649 \pm 385	40%	0.48237
		0.2	513 \pm 381	58%	0.38064
		0.3	420 \pm 349	69%	0.30153
0.05		0.1	601 \pm 390	47%	0.36776
		0.2	466 \pm 370	64%	0.29037
		0.3	381 \pm 329	75%	0.23057
0.07		0.1	567 \pm 390	52%	0.2928
		0.2	436 \pm 359	69%	0.23294
		0.3	425 \pm 125	15%	0.4971
0.1		0.1	517 \pm 385	59%	0.19579
		0.2	394 \pm 338	74%	0.15903
		0.3	323 \pm 285	83%	0.12956
0	2	0.1	579 \pm 391	50%	0.40665
		0.2	447 \pm 363	67%	0.30644
		0.3	365 \pm 318	77%	0.23769
0.02		0.1	552 \pm 389	54%	0.3483
		0.2	422 \pm 353	71%	0.26242
		0.3	345 \pm 304	80%	0.20462

Table A3.12 (continued)

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0.05	2	0.1	509 \pm 383	60%	0.2619
		0.2	387 \pm 334	75%	0.19998
		0.3	318 \pm 281	84%	0.15778
0.07		0.1	480 \pm 376	64%	0.21038
		0.2	364 \pm 320	78%	0.16223
		0.3	302 \pm 264	86%	0.12974
0.1		0.1	439 \pm 362	69%	0.14362
		0.2	334 \pm 296	82%	0.11394
		0.3	280 \pm 239	89%	0.09316

Table A3.13a: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Under

$$\hat{\mu}_C - \hat{\mu}_T$$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/ \text{group}$. Simulation runs=100,000

The second stage sample size was capped at **2×original fixed sample size**

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Type I error Rate
0.2	1	939 \pm 143	89%	0.03234
0.2	1.2	953 \pm 120	92%	0.03152
0.2	1.5	966 \pm 94	95%	0.02991
0.2	1.7	971 \pm 80	96%	0.02926
0.2	2.0	976 \pm 65	98%	0.02817

Table A3.13b: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Under

$$\hat{\sigma}_C^2 - \hat{\sigma}_T^2$$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/ \text{group}$. Simulation runs=100,000

The second stage sample size was capped at **5×original fixed sample size**

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Type I error Rate
0.2	1	1925 \pm 529	80%	0.03574
0.2	1.2	1982 \pm 464	84%	0.03443
0.2	1.5	2044 \pm 378	88%	0.03237
0.2	1.7	2073 \pm 328	91%	0.03137
0.2	2.0	2104 \pm 266	94%	0.02971

Table A3.14: Power at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (early stopping for futility NOT allowed) – Under $\hat{\sigma}_\square - \hat{\sigma}_\square$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=10,000

The second stage sample size was capped at 2×original fixed sample size

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Power
0	1	571±249	22%	0.94968
	1.2	687±273	40%	0.91012
	1.5	814±249	63%	0.80778
	1.7	867±220	74%	0.72038
	2.0	917±176	85%	0.5932
0.02	1	614±263	28%	0.92029
	1.2	725±271	47%	0.86005
	1.5	839±236	68%	0.72681
	1.7	885±206	78%	0.63342
	2.0	928±163	87%	0.50836
0.05	1	681±273	39%	0.83789
	1.2	780±260	56%	0.73776
	1.5	872±215	74%	0.57985
	1.7	908±184	82%	0.48774
	2.0	941±143	90%	0.38265
0.07	1	727±271	47%	0.74451
	1.2	814±247	63%	0.62591
	1.5	892±199	79%	0.46984
	1.7	922±169	85%	0.39044
	2.0	949±131	91%	0.30455
0.1	1	793±255	59%	0.54428
	1.2	859±222	72%	0.42974
	1.5	917±173	84%	0.31182
	1.7	939±146	89%	0.25848
	2.0	958±113	93%	0.20289

Table A3.15: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (allowing early stopping for futility) – Under $\hat{\mu}_C - \hat{\mu}_T$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393/\text{group}$. Simulation runs=100,000

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Type I error Rate
0.2	1.0	10%	415 \pm 513	77%	0.02274
0.2		20%	300 \pm 273	83%	0.02054
0.2		30%	258 \pm 176	87%	0.01896
0.2	1.2	10%	447 \pm 614	79%	0.02077
0.2		20%	308 \pm 316	85%	0.01853
0.2		30%	259 \pm 199	89%	0.01694
0.2	1.5	10%	503 \pm 785	81%	0.01783
0.2		20%	324 \pm 392	88%	0.01545
0.2		30%	263 \pm 239	91%	0.01388
0.2	1.7	10%	543 \pm 911	83%	0.01652
0.2		20%	336 \pm 448	89%	0.01419
0.2		30%	268 \pm 270	92%	0.01249
0.2	2.0	10%	611 \pm 1124	84%	0.0143
0.2		20%	355 \pm 540	90%	0.01206
0.2		30%	273 \pm 316	93%	0.01033

Table A3.16: Power at different common unknown variances when both variance and treatment difference were updated at the interim using CP without critical value adjustment (allowing early stopping for futility) – Under $\hat{\mu}_C - \hat{\mu}_T$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / \text{group}$. Simulation runs=10,000

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0	1	0.1	487 \pm 227	11%	0.84557
		0.2	450 \pm 206	15%	0.79949
		0.3	419 \pm 180	19%	0.75986
0.02		0.1	496 \pm 246	15%	0.78207
		0.2	452 \pm 224	21%	0.72736
		0.3	414 \pm 195	25%	0.68115
0.05		0.1	502 \pm 273	23%	0.65545
		0.2	445 \pm 247	30%	0.59171
		0.3	400 \pm 214	36%	0.53908
0.07		0.1	497 \pm 288	29%	0.54946
		0.2	434 \pm 258	37%	0.48545
		0.3	385 \pm 221	43%	0.43504
0.1		0.1	476 \pm 304	40%	0.37196
		0.2	407 \pm 267	49%	0.32067
		0.3	359 \pm 226	55%	0.28186
0	1.2	0.1	530 \pm 283	20%	0.73361
		0.2	471 \pm 263	27%	0.66384
		0.3	424 \pm 236	33%	0.60607
0.02		0.1	528 \pm 296	25%	0.65915
		0.2	464 \pm 274	33%	0.5856
		0.3	412 \pm 244	40%	0.52568
0.05		0.1	517 \pm 312	33%	0.52492
		0.2	445 \pm 286	43%	0.45393
		0.3	390 \pm 250	50%	0.39832
0.07		0.1	502 \pm 320	40%	0.42535
		0.2	427 \pm 289	49%	0.36223
		0.3	374 \pm 252	56%	0.3148
0.1		0.1	470 \pm 326	50%	0.27765
		0.2	396 \pm 288	59%	0.23297
		0.3	344 \pm 245	66%	0.19971

Table A3.16 (continued):

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0	1.5	0.1	551 \pm 331	33%	0.57141
		0.2	472 \pm 311	44%	0.48712
		0.3	412 \pm 281	51%	0.42199
0.02		0.1	537 \pm 338	38%	0.49321
		0.2	455 \pm 314	49%	0.41439
		0.3	396 \pm 281	56%	0.35508
0.05		0.1	510 \pm 344	46%	0.37379
		0.2	427 \pm 313	57%	0.30836
		0.3	370 \pm 276	64%	0.26005
0.07		0.1	487 \pm 345	51%	0.2944
		0.2	406 \pm 309	62%	0.24111
		0.3	351 \pm 268	69%	0.20206
0.1		0.1	451 \pm 341	59%	0.18877
		0.2	374 \pm 298	69%	0.15328
		0.3	324 \pm 254	75%	0.12842
0	1.7	0.1	545 \pm 350	41%	0.47619
		0.2	459 \pm 327	52%	0.39386
		0.3	398 \pm 295	60%	0.33193
0.02		0.1	527 \pm 353	46%	0.40549
		0.2	441 \pm 326	57%	0.33138
		0.3	381 \pm 290	64%	0.27682
0.05		0.1	495 \pm 354	53%	0.29988
		0.2	411 \pm 319	63%	0.24194
		0.3	354 \pm 279	71%	0.19964
0.07		0.1	472 \pm 351	57%	0.23453
		0.2	390 \pm 311	68%	0.18874
		0.3	385 \pm 221	43%	0.43504
0.1		0.1	436 \pm 343	64%	0.15176
		0.2	360 \pm 297	74%	0.12123
		0.3	312 \pm 253	80%	0.09958
0	2	0.1	523 \pm 364	50%	0.36268
		0.2	433 \pm 335	62%	0.289
		0.3	372 \pm 298	69%	0.23615
0.02		0.1	503 \pm 363	54%	0.3046
		0.2	414 \pm 329	65%	0.24036
		0.3	356 \pm 290	73%	0.19508

Table A3.16 (continued)

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0.05	2	0.1	471 \pm 358	60%	0.22178
		0.2	386 \pm 318	71%	0.17412
		0.3	333 \pm 276	77%	0.14066
0.07		0.1	448 \pm 353	64%	0.17378
		0.2	368 \pm 308	74%	0.13577
		0.3	318 \pm 265	80%	0.10961
0.1		0.1	415 \pm 341	69%	0.11399
		0.2	341 \pm 291	79%	0.08878
		0.3	298 \pm 247	84%	0.07132

Table A3.17a: Type I error rate at different common unknown variances when only variance was updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Under $\hat{\mu}_C - \hat{\mu}_T$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=100,000

The second stage sample size was capped at **2×original fixed sample size**

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Type I error Rate
0.2	1	939 \pm 143	89%	0.02564
0.2	1.2	953 \pm 120	92%	0.02544
0.2	1.5	966 \pm 94	95%	0.02522
0.2	1.7	971 \pm 80	96%	0.0254
0.2	2.0	976 \pm 65	98%	0.0251

Table A3.17b: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Under $\hat{\mu}_C - \hat{\mu}_T$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=100,000

The second stage sample size was capped at **5×original fixed sample size**

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Type I error Rate
0.2	1	1925 \pm 529	80%	0.02515
0.2	1.2	1982 \pm 464	84%	0.02494
0.2	1.5	2044 \pm 378	88%	0.02463
0.2	1.7	2073 \pm 328	91%	0.02512
0.2	2.0	2104 \pm 266	94%	0.02464

Table A3.18: Power at different common unknown variances when both variance and treatment difference were updated at the interim using CP with critical value adjustment (early stopping for futility NOT allowed) – Under $\hat{\mu}_C - \hat{\mu}_T$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / \text{group}$. Simulation runs=10,000

The second stage sample size was capped at $2 \times$ original fixed sample size

True $\mu_C - \mu_T$	True σ^2	Re-estimated Sample Size \pm SD	% of Capping	Power
0	1	571 \pm 249	22%	0.94438
	1.2	687 \pm 273	40%	0.88866
	1.5	814 \pm 249	63%	0.76248
	1.7	867 \pm 220	74%	0.66855
	2.0	917 \pm 176	85%	0.54061
0.02	1	614 \pm 263	28%	0.90642
	1.2	725 \pm 271	47%	0.82568
	1.5	839 \pm 236	68%	0.67426
	1.7	885 \pm 206	78%	0.57777
	2.0	928 \pm 163	87%	0.45966
0.05	1	681 \pm 273	39%	0.80096
	1.2	780 \pm 260	56%	0.68272
	1.5	872 \pm 215	74%	0.51986
	1.7	908 \pm 184	82%	0.4352
	2.0	941 \pm 143	90%	0.33924
0.07	1	727 \pm 271	47%	0.68909
	1.2	814 \pm 247	63%	0.56131
	1.5	892 \pm 199	79%	0.41353
	1.7	922 \pm 169	85%	0.34236
	2.0	949 \pm 131	91%	0.26819
0.1	1	793 \pm 255	59%	0.4736
	1.2	859 \pm 222	72%	0.36843
	1.5	917 \pm 173	84%	0.26635
	1.7	939 \pm 146	89%	0.22256
	2.0	958 \pm 113	93%	0.17729

Table A3.19: Type I error rate at different common unknown variances when both variance and treatment difference were updated at the interim using CP with critical value adjustment (allowing early stopping for futility) – Under $\hat{\mu}_C - \hat{\mu}_T$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / group$. Simulation runs=100,000

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Type I error Rate
0.2	1.0	10%	415 \pm 513	77%	0.02563
0.2		20%	300 \pm 273	83%	0.02556
0.2		30%	258 \pm 176	87%	0.02553
0.2	1.2	10%	447 \pm 614	79%	0.02544
0.2		20%	308 \pm 316	85%	0.02542
0.2		30%	259 \pm 199	89%	0.02508
0.2	1.5	10%	503 \pm 785	81%	0.02534
0.2		20%	324 \pm 392	88%	0.02488
0.2		30%	263 \pm 239	91%	0.02494
0.2	1.7	10%	543 \pm 911	83%	0.02577
0.2		20%	336 \pm 448	89%	0.02539
0.2		30%	268 \pm 270	92%	0.02532
0.2	2.0	10%	611 \pm 1124	84%	0.02539
0.2		20%	355 \pm 540	90%	0.02513
0.2		30%	273 \pm 316	93%	0.02503

Table A3.20: Power at different common unknown variances when both variance and treatment difference were updated at the interim using CP with critical value adjustment (allowing early stopping for futility) – Under $\hat{\mu}_C - \hat{\mu}_T$

Study design parameters: Assume $\mu_C - \mu_T = 0$, $\Delta = 0.2$, common within-group variance $\sigma^2 = 1$, initial design sample size was calculated with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size, $n = 393 / \text{group}$. Simulation runs=10,000

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0	1	10%	487 \pm 227	11%	0.84818
		20%	450 \pm 206	15%	0.80419
		30%	419 \pm 180	19%	0.76665
0.02		10%	496 \pm 246	15%	0.7852
		20%	452 \pm 224	21%	0.7334
		30%	414 \pm 195	25%	0.68996
0.05		10%	502 \pm 273	23%	0.65814
		20%	445 \pm 247	30%	0.59894
		30%	400 \pm 214	36%	0.54953
0.07		10%	497 \pm 288	29%	0.5514
		20%	434 \pm 258	37%	0.49385
		30%	385 \pm 221	43%	0.44759
0.1		10%	476 \pm 304	40%	0.37277
		20%	407 \pm 267	49%	0.33
		30%	359 \pm 226	55%	0.2956
0	1.2	10%	530 \pm 283	20%	0.73897
		20%	471 \pm 263	27%	0.6736
		30%	424 \pm 236	33%	0.61939
0.02		10%	528 \pm 296	25%	0.66594
		20%	464 \pm 274	33%	0.5971
		30%	412 \pm 244	40%	0.54142
0.05		10%	517 \pm 312	33%	0.53274
		20%	445 \pm 286	43%	0.46907
		30%	390 \pm 250	50%	0.41787
0.07		10%	502 \pm 320	40%	0.4341
		20%	427 \pm 289	49%	0.37897
		30%	374 \pm 252	56%	0.33567
0.1		10%	470 \pm 326	50%	0.28569
		20%	396 \pm 288	59%	0.24874
		30%	344 \pm 245	66%	0.21953

Table A3.20 (continued):

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0	1.5	10%	551 \pm 331	33%	0.58455
		20%	472 \pm 311	44%	0.50733
		30%	412 \pm 281	51%	0.44729
0.02		10%	537 \pm 338	38%	0.5099
		20%	455 \pm 314	49%	0.43851
		30%	396 \pm 281	56%	0.38374
0.05		10%	510 \pm 344	46%	0.39136
		20%	427 \pm 313	57%	0.33362
		30%	370 \pm 276	64%	0.28937
0.07		10%	487 \pm 345	51%	0.31233
		20%	406 \pm 309	62%	0.26629
		30%	351 \pm 268	69%	0.23113
0.1		10%	451 \pm 341	59%	0.20535
		20%	374 \pm 298	69%	0.17678
		30%	324 \pm 254	75%	0.15481
0	1.7	10%	545 \pm 350	41%	0.49639
		20%	459 \pm 327	52%	0.42184
		30%	398 \pm 295	60%	0.36465
0.02		10%	527 \pm 353	46%	0.42836
		20%	441 \pm 326	57%	0.3615
		30%	381 \pm 290	64%	0.31133
0.05		10%	495 \pm 354	53%	0.32336
		20%	411 \pm 319	63%	0.27248
		30%	354 \pm 279	71%	0.23396
0.07		10%	472 \pm 351	57%	0.25686
		20%	390 \pm 311	68%	0.21774
		30%	385 \pm 221	43%	0.44759
0.1		10%	436 \pm 343	64%	0.17232
		20%	360 \pm 297	74%	0.14798
		30%	312 \pm 253	80%	0.12938
0	2	10%	523 \pm 364	50%	0.39106
		20%	433 \pm 335	62%	0.3246
		30%	372 \pm 298	69%	0.27579
0.02		10%	503 \pm 363	54%	0.33407
		20%	414 \pm 329	65%	0.27672
		30%	356 \pm 290	73%	0.23534

Table A3.20 (continued)

True $\mu_C - \mu_T$	True σ^2	CP cutoff at interim to stop for futility	Re-estimated Sample Size \pm SD	% of Stopping	Power
0.05	2	10%	471 \pm 358	60%	0.25041
		20%	386 \pm 318	71%	0.20917
		30%	333 \pm 276	77%	0.17896
0.07		10%	448 \pm 353	64%	0.2005
		20%	368 \pm 308	74%	0.16853
		30%	318 \pm 265	80%	0.14539
0.1		10%	415 \pm 341	69%	0.13718
		20%	341 \pm 291	79%	0.11731
		30%	298 \pm 247	84%	0.10261

Table A4.1a: Type I error Rate for Blackwelder's and Farrington-Manning's tests in the fixed sample size design with Risk Difference Approach– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning's method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255 / group$. Simulation runs=100,000

True π_T	True π_C	Fixed Margin	Type I error Rate	
			Blackwelder	Farrington-Manning
0.85	0.95	0.1	0.02188	0.02513
0.8	0.9	0.1	0.02430	0.02523
0.75	0.85	0.1	0.02472	0.02459
0.7	0.8	0.1	0.02555	0.02471
0.65	0.75	0.1	0.02562	0.02464
0.6	0.7	0.1	0.02516	0.02432
0.55	0.65	0.1	0.02558	0.02412
0.5	0.6	0.1	0.02394	0.02343
0.45	0.55	0.1	0.02420	0.02420

Table A4.1b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the fixed sample size design with Risk Difference Approach – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581/ \text{group}$. Simulation runs=100,000

True π_T	True π_C	Fixed Margin	Type I error Rate	
			Blackwelder	Farrington-Manning
0.07	0.02	-0.05	0.01946	0.02628
0.1	0.05	-0.05	0.02412	0.02601
0.13	0.08	-0.05	0.02517	0.0263
0.15	0.1	-0.05	0.0257	0.02527
0.17	0.12	-0.05	0.02591	0.02595
0.2	0.15	-0.05	0.0262	0.02596
0.23	0.18	-0.05	0.02623	0.02586
0.25	0.2	-0.05	0.02623	0.02586
0.27	0.22	-0.05	0.0262	0.02593
0.3	0.25	-0.05	0.02648	0.02595

Table A4.2a: Power for Blackwelder’s and Farrington-Manning’s tests in the fixed sample size design with different true event rates– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255 / group$. Simulation runs=10,000

True π_T	True π_C	Fixed Margin	Power	
			Blackwelder	Farrington-Manning
0.95	0.95	0.1	0.99900	0.99804
0.9	0.9	0.1	0.96311	0.95788
0.85	0.85	0.1	0.88837	0.88419
0.8	0.8	0.1	0.80798	0.80431
0.75	0.75	0.1	0.7447	0.74346
0.7	0.7	0.1	0.69654	0.69654
0.65	0.65	0.1	0.6624	0.66367
0.6	0.6	0.1	0.63245	0.63632
0.55	0.55	0.1	0.62352	0.6236

Table A4.2b: Power for Blackwelder’s and Farrington-Manning’s tests in the fixed sample size design with different true event rates – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581/ \text{group}$. Simulation runs=10,000

True π_T	True π_C	Fixed Margin	Power	
			Blackwelder	Farrington-Manning
0.02	0.02	-0.05	0.99996	0.99987
0.05	0.05	-0.05	0.97382	0.96775
0.08	0.08	-0.05	0.87991	0.8717
0.1	0.1	-0.05	0.80811	0.80142
0.12	0.12	-0.05	0.74235	0.73795
0.15	0.15	-0.05	0.6629	0.6587
0.18	0.18	-0.05	0.59931	0.59646
0.2	0.2	-0.05	0.56652	0.56384
0.22	0.22	-0.05	0.53632	0.53417
0.25	0.25	-0.05	0.49986	0.49818

Table A4.3a: Type I error rate for Normal Approximation and Farrington-Manning's tests in the fixed sample size design with relative risk approach– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $R = 0.875$, initial design sample size was calculated using Farrington-Manning's method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 231 / group$. Simulation runs=100,000

True π_T	True π_C	Relative Margin	Type I error Rate	
			Normal Approximation	Farrington-Manning
0.83125	0.95	0.875	0.0211212	0.0266918
0.7875	0.9	0.875	0.0225148	0.0249028
0.74375	0.85	0.875	0.0240270	0.0240360
0.7	0.8	0.875	0.0239150	0.0239150
0.65625	0.75	0.875	0.0259372	0.0259002
0.6125	0.7	0.875	0.0257368	0.0252546
0.56875	0.65	0.875	0.0259010	0.0253230
0.525	0.6	0.875	0.0259706	0.0249414
0.48125	0.55	0.875	0.0258506	0.0250354

Table A4.3b: Type I error rate for Normal Approximation and Farrington-Manning's tests in the fixed sample size design with relative risk approach – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $R = 1.5$, initial design sample size was calculated using Farrington-Manning's method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 869 / group$. Simulation runs=100,000

True π_T	True π_C	Relative Margin	Type I error Rate	
			Normal Approximation	Farrington-Manning
0.03	0.02	1.5	0.0378532	0.0256764
0.075	0.05	1.5	0.0305326	0.0257794
0.12	0.08	1.5	0.0292308	0.0254316
0.15	0.1	1.5	0.0285756	0.0254332
0.18	0.12	1.5	0.028421	0.0253774
0.225	0.15	1.5	0.0278334	0.0252324
0.27	0.18	1.5	0.0274454	0.0252326
0.3	0.2	1.5	0.0273788	0.0250684
0.33	0.22	1.5	0.0273392	0.0250488
0.375	0.25	1.5	0.0270678	0.0250612

Table A4.4a: Power for Normal Approximation and Farrington-Manning's tests in the fixed sample size design with relative risk approach– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $R = 0.875$, initial design sample size was calculated using Farrington-Manning's method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 231/ \text{group}$. Simulation runs=10,000

True π_T	True π_C	Relative Margin (R)	Power	
			Normal Approximation	Farrington-Manning
0.95	0.95	0.875	0.9999362	0.9998424
0.9	0.9	0.875	0.9871088	0.9826822
0.85	0.85	0.875	0.920805	0.9119718
0.8	0.8	0.875	0.8140686	0.8034138
0.75	0.75	0.875	0.6981288	0.6895568
0.7	0.7	0.875	0.5914768	0.584794
0.65	0.65	0.875	0.4983854	0.4930402
0.6	0.6	0.875	0.4195334	0.4181272
0.55	0.55	0.875	0.3524624	0.3537142

Table A4.4b: Power for Normal Approximation and Farrington-Manning's tests in the fixed sample size design with relative risk approach – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $R = 1.5$, initial design sample size was calculated using Farrington Manning's method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 869 / group$. Simulation runs=10,000

True π_T	True π_C	Relative Margin (R)	Power	
			Normal Approximation	Farrington-Manning
0.02	0.02	1.5	0.192754	0.2279562
0.05	0.05	1.5	0.4627588	0.4995392
0.08	0.08	1.5	0.6800054	0.7034356
0.1	0.1	1.5	0.788216	0.8017598
0.12	0.12	1.5	0.864731	0.8736788
0.15	0.15	1.5	0.9352348	0.9408216
0.18	0.18	1.5	0.972305	0.97469
0.2	0.2	1.5	0.9851184	0.9863914
0.22	0.22	1.5	0.9923136	0.993049
0.25	0.25	1.5	0.9923136	0.993049

Table A4.5a: Type I error rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in Risk Difference Approach– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255 / group$. Simulation runs=100,000

True π_T	True π_C	Relative Margin	Type I error Rate		Margin			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.83125	0.95	$0.125 \times \hat{\pi}_C$	0.02033	0.03327	0.11874	0.11863	0.11765	0.12010
0.7875	0.9	$0.125 \times \hat{\pi}_C$	0.02453	0.02955	0.11249	0.11275	0.11078	0.11422
0.74375	0.85	$0.125 \times \hat{\pi}_C$	0.02267	0.02753	0.10624	0.10637	0.10441	0.10833
0.7	0.8	$0.125 \times \hat{\pi}_C$	0.02627	0.02656	0.09999	0.10000	0.09804	0.10196
0.65625	0.75	$0.125 \times \hat{\pi}_C$	0.02439	0.02512	0.09374	0.09363	0.09167	0.09608
0.6125	0.7	$0.125 \times \hat{\pi}_C$	0.02499	0.02571	0.08749	0.08775	0.08529	0.08971
0.56875	0.65	$0.125 \times \hat{\pi}_C$	0.02508	0.02493	0.08124	0.08137	0.07892	0.08382
0.525	0.6	$0.125 \times \hat{\pi}_C$	0.02567	0.02481	0.07499	0.07500	0.07255	0.07745
0.48125	0.55	$0.125 \times \hat{\pi}_C$	0.02542	0.02404	0.06874	0.06863	0.06618	0.07157

Table A4.5b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in Risk Difference Approach – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581/ group$. Simulation runs=100,000

True π_T	True π_C	Relative Margin	Type I error Rate		Margin			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.03	0.02	$-0.5 \times \hat{\pi}_C$	0.03418	0.02456	-0.0100	-0.0100	-0.0115	-0.0085
0.075	0.05	$-0.5 \times \hat{\pi}_C$	0.03035	0.02058	-0.0250	-0.0251	-0.0276	-0.0225
0.12	0.08	$-0.5 \times \hat{\pi}_C$	0.02932	0.02007	-0.0400	-0.0401	-0.0431	-0.0371
0.15	0.1	$-0.5 \times \hat{\pi}_C$	0.02854	0.01967	-0.0500	-0.0501	-0.0531	-0.0466
0.18	0.12	$-0.5 \times \hat{\pi}_C$	0.02858	0.02011	-0.0600	-0.0601	-0.0636	-0.0566
0.225	0.15	$-0.5 \times \hat{\pi}_C$	0.02813	0.02	-0.0750	-0.0752	-0.0787	-0.0711
0.27	0.18	$-0.5 \times \hat{\pi}_C$	0.02775	0.02031	-0.0900	-0.0902	-0.0942	-0.0857
0.3	0.2	$-0.5 \times \hat{\pi}_C$	0.02779	0.02051	-0.1000	-0.1002	-0.1042	-0.0957
0.33	0.22	$-0.5 \times \hat{\pi}_C$	0.02746	0.02078	-0.1100	-0.1102	-0.1142	-0.1057
0.375	0.25	$-0.5 \times \hat{\pi}_C$	0.02749	0.021	-0.1250	-0.1247	-0.1298	-0.1202

Table A4.6a: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in Risk Difference Approach– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255 / group$. Simulation runs=10,000

True π_T	True π_C	Relative Margin	Power		Margin			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.95	0.95	$0.125 \times \hat{\pi}_C$	1	1	0.11874	0.11863	0.11765	0.12010
0.9	0.9	$0.125 \times \hat{\pi}_C$	0.99048	0.99328	0.11249	0.11275	0.11078	0.11422
0.85	0.85	$0.125 \times \hat{\pi}_C$	0.93196	0.94805	0.10624	0.10637	0.10441	0.10833
0.8	0.8	$0.125 \times \hat{\pi}_C$	0.82351	0.85569	0.09999	0.10000	0.09804	0.10196
0.75	0.75	$0.125 \times \hat{\pi}_C$	0.69904	0.74451	0.09374	0.09363	0.09167	0.09608
0.7	0.7	$0.125 \times \hat{\pi}_C$	0.5867	0.6372	0.08749	0.08775	0.08529	0.08971
0.65	0.65	$0.125 \times \hat{\pi}_C$	0.48618	0.54115	0.08124	0.08137	0.07892	0.08382
0.6	0.6	$0.125 \times \hat{\pi}_C$	0.40469	0.4565	0.07499	0.07500	0.07255	0.07745
0.55	0.55	$0.125 \times \hat{\pi}_C$	0.33582	0.38626	0.06874	0.06863	0.06618	0.07157

Table A4.6b: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in Risk Difference Approach – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581/ group$. Simulation runs=10,000

True π_T	True π_C	Relative Margin	Power		Margin			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.02	0.02	$-0.5 \times \hat{\pi}_C$	0.1305	0.2456	-0.0100	-0.0103	-0.0120	-0.0077
0.05	0.05	$-0.5 \times \hat{\pi}_C$	0.517	0.6662	-0.0250	-0.0251	-0.0271	-0.0225
0.08	0.08	$-0.5 \times \hat{\pi}_C$	0.7344	0.8466	-0.0399	-0.0401	-0.0426	-0.0371
0.1	0.1	$-0.5 \times \hat{\pi}_C$	0.8351	0.9173	-0.0499	-0.0501	-0.0531	-0.0466
0.12	0.12	$-0.5 \times \hat{\pi}_C$	0.9065	0.9589	-0.0599	-0.0601	-0.0631	-0.0566
0.15	0.15	$-0.5 \times \hat{\pi}_C$	0.9638	0.9865	-0.0749	-0.0746	-0.0787	-0.0711
0.18	0.18	$-0.5 \times \hat{\pi}_C$	0.9878	0.9955	-0.0899	-0.0897	-0.0942	-0.0857
0.2	0.2	$-0.5 \times \hat{\pi}_C$	0.9943	0.9975	-0.0999	-0.0997	-0.1042	-0.0957
0.22	0.22	$-0.5 \times \hat{\pi}_C$	0.9967	0.9994	-0.1099	-0.1097	-0.1142	-0.1057
0.25	0.25	$-0.5 \times \hat{\pi}_C$	0.9994	0.9998	-0.1249	-0.1247	-0.1293	-0.1202

Table A4.7a: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with fixed margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255 / group$. Simulation runs=100,000

True π_T	True π_C	Fixed Margin	Type I error Rate		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.85	0.95	0.1	0.02198	0.0259	255	255	255	255
0.8	0.9	0.1	0.02419	0.0251	255	255	255	255
0.75	0.85	0.1	0.02458	0.02441	263	255	255	268
0.7	0.8	0.1	0.02546	0.02478	295	296	280	310
0.65	0.75	0.1	0.02561	0.02455	328	330	318	341
0.6	0.7	0.1	0.02574	0.02481	355	357	345	365
0.55	0.65	0.1	0.0257	0.02475	374	374	368	381
0.5	0.6	0.1	0.02598	0.02426	385	386	383	389
0.45	0.55	0.1	0.02698	0.02376	389	390	388	390

Table A4.7b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with fixed margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581/ \text{group}$. Simulation runs=100,000

True π_T	True π_C	Fixed Margin	Type I error Rate		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.07	0.02	-0.5	0.01969	0.02732	581	581	581	581
0.1	0.05	-0.5	0.02364	0.02607	581	581	581	581
0.13	0.08	-0.5	0.02376	0.02492	618	604	581	645
0.15	0.1	-0.5	0.02377	0.02434	698	700	653	739
0.17	0.12	-0.5	0.02507	0.02545	786	784	746	828
0.2	0.15	-0.5	0.02519	0.02522	911	913	871	954
0.23	0.18	-0.5	0.02548	0.02545	1025	1025	987	1068
0.25	0.2	-0.5	0.02559	0.02547	1096	1098	1056	1138
0.27	0.22	-0.5	0.02543	0.02525	1161	1160	1126	1202
0.3	0.25	-0.5	0.02541	0.02523	1250	1252	1218	1285

Table A4.8a: Power for Blackwelder’s and Farrington-Manning’s tests in the design with fixed margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255 / group$. Simulation runs=10,000

True π_T	True π_C	Fixed Margin	Power		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.95	0.95	0.1	0.9991	0.998	255	255	255	255
0.9	0.9	0.1	0.96283	0.9576	255	255	255	255
0.85	0.85	0.1	0.88612	0.88221	255	255	255	255
0.8	0.8	0.1	0.82064	0.8178	263	255	255	268
0.75	0.75	0.1	0.803	0.80305	295	296	280	310
0.7	0.7	0.1	0.79997	0.80109	328	330	318	341
0.65	0.65	0.1	0.79923	0.80121	350	351	341	360
0.6	0.6	0.1	0.7977	0.79959	374	374	368	381
0.55	0.55	0.1	0.79731	0.79788	385	386	383	389

Table A4.8b: Power for Blackwelder’s and Farrington-Manning’s tests in the design with fixed margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581 / group$. Simulation runs=10,000

True π_T	True π_C	Fixed Margin	Power		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.02	0.02	-0.05	0.99995	0.99985	581	581	581	581
0.05	0.05	-0.05	0.97359	0.96798	581	581	581	581
0.08	0.08	-0.05	0.88092	0.87355	582	581	581	581
0.1	0.1	-0.05	0.82713	0.8206	605	581	581	620
0.12	0.12	-0.05	0.80873	0.8042	676	676	629	716
0.15	0.15	-0.05	0.80549	0.80303	807	806	762	850
0.18	0.18	-0.05	0.80131	0.80013	931	934	885	974
0.2	0.2	-0.05	0.80081	0.80018	1007	1006	967	1050
0.22	0.22	-0.05	0.80012	0.79988	1078	1080	1037	1121
0.25	0.25	-0.05	0.80006	0.80007	1177	1176	1138	1218

Table A4.9a: Type I error Rate for Normal Approximation and Farrington-Manning's tests in the design with relative risk approach with sample size re-estimation at the interim– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $R = 0.875$, initial design sample size was calculated using Farrington-Manning's method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 231/ \text{group}$. Simulation runs=100,000

True π_T	True π_C	Relative Margin	Type I error Rate		Re-estimated Sample Size			
			Normal Approximation	Farrington-Manning	Mean	Median	Q1	Q3
0.83125	0.95	0.875	0.02046	0.0262	231	231	231	231
0.7875	0.9	0.875	0.02178	0.0242	231	231	231	231
0.74375	0.85	0.875	0.02356	0.0241	248	234	231	258
0.7	0.8	0.875	0.02404	0.02428	304	302	276	329
0.65625	0.75	0.875	0.02374	0.02362	381	380	343	411
0.6125	0.7	0.875	0.0247	0.0244	470	469	427	515
0.56875	0.65	0.875	0.02464	0.02426	574	574	524	626
0.525	0.6	0.875	0.02526	0.02476	695	684	626	758
0.48125	0.55	0.875	0.02582	0.02542	839	826	758	915

Table A4.9b: Type I error Rate for Normal Approximation and Farrington-Manning’s tests in the design with relative risk approach with sample size re-estimation at the interim– Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $R = 1.5$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 869 / group$. Simulation runs=100,000

True π_T	True π_C	Relative Margin	Type I error Rate		Re-estimated Sample Size			
			Normal Approximation	Farrington-Manning	Mean	Median	Q1	Q3
0.03	0.02	1.5	0.02898	0.02513	3964	3721	3263	4323
0.075	0.05	1.5	0.02859	0.02491	1476	1459	1327	1583
0.12	0.08	1.5	0.02867	0.02507	914	869	869	941
0.15	0.1	1.5	0.02837	0.02537	870	869	869	869
0.18	0.12	1.5	0.02842	0.02532	869	869	869	869
0.225	0.15	1.5	0.02768	0.02495	869	869	869	869
0.27	0.18	1.5	0.02733	0.02519	869	869	869	869
0.3	0.2	1.5	0.0272	0.02502	869	869	869	869
0.33	0.22	1.5	0.02717	0.02494	869	869	869	869
0.375	0.25	1.5	0.02691	0.02479	869	869	869	869

Table A4.10a: Power for Normal Approximation and Farrington-Manning's tests in the design with relative risk approach with sample size re-estimation at the interim– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $R = 0.875$, initial design sample size was calculated using Farrington-Manning's method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 231/ \text{group}$. Simulation runs=10,000

True π_T	True π_C	Relative Margin	Power		Re-estimated Sample Size			
			Normal Approximation	Farrington-Manning	Mean	Median	Q1	Q3
0.95	0.95	0.875	0.99996	0.99988	231	231	231	231
0.9	0.9	0.875	0.98774	0.98342	231	231	231	231
0.85	0.85	0.875	0.92234	0.9142	231	231	231	231
0.8	0.8	0.875	0.8401	0.83102	246	231	231	252
0.75	0.75	0.875	0.81282	0.80856	304	302	276	329
0.7	0.7	0.875	0.80586	0.80332	387	387	350	419
0.65	0.65	0.875	0.80038	0.79974	483	478	435	524
0.6	0.6	0.875	0.80456	0.8049	597	594	544	649
0.55	0.55	0.875	0.80082	0.80164	731	720	660	785

Table A4.10b: Power for Normal Approximation and Farrington-Manning's tests in the design with relative risk approach with sample size re-estimation at the interim– Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $R = 1.5$, initial design sample size was calculated using Farrington-Manning's method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 869 / group$. Simulation runs=10,000

True π_T	True π_C	Relative Margin	Power		Re-estimated Sample Size			
			Normal Approximation	Farrington-Manning	Mean	Median	Q1	Q3
0.02	0.02	1.5	0.79122	0.80644	5058	4843	4102	5502
0.05	0.05	1.5	0.7865	0.80268	1879	1857	1653	2057
0.08	0.08	1.5	0.7858	0.8021	1127	1104	1024	1216
0.1	0.1	1.5	0.8034	0.81698	914	869	869	941
0.12	0.12	1.5	0.86238	0.87146	871	869	869	869
0.15	0.15	1.5	0.93394	0.93998	869	869	869	869
0.18	0.18	1.5	0.97256	0.97536	869	869	869	869
0.2	0.2	1.5	0.9859	0.9872	869	869	869	869
0.22	0.22	1.5	0.99282	0.99354	869	869	869	869
0.25	0.25	1.5	0.9978	0.99804	869	869	869	869

Table A4.11a: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255/ group$. Simulation runs=100,000

Pooled event rates were calculated at the interim, $\hat{\pi}_C$ was then estimated under alternative hypothesis, the margin was re-set at the interim for sample size re-estimation ($\hat{\pi}_{1T} = \hat{\pi}_{1C} = \pi_{1\bullet}$).

True π_T	True π_C	Relative Margin	Type I error Rate		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.83125	0.95	$0.125 \times \hat{\pi}_C$	0.02124	0.03369	255	255	255	255
0.7875	0.9	$0.125 \times \hat{\pi}_C$	0.02563	0.02949	255	255	255	255
0.74375	0.85	$0.125 \times \hat{\pi}_C$	0.02523	0.02634	259	255	255	255
0.7	0.8	$0.125 \times \hat{\pi}_C$	0.02434	0.02617	298	293	264	325
0.65625	0.75	$0.125 \times \hat{\pi}_C$	0.02409	0.02532	374	371	338	407
0.6125	0.7	$0.125 \times \hat{\pi}_C$	0.02329	0.02379	464	461	422	503
0.56875	0.65	$0.125 \times \hat{\pi}_C$	0.02445	0.02461	547	547	494	594
0.525	0.6	$0.125 \times \hat{\pi}_C$	0.02458	0.02437	692	687	634	744
0.48125	0.55	$0.125 \times \hat{\pi}_C$	0.02469	0.02438	838	831	768	899

Table A4.11b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581 / group$. Simulation runs=100,000

Pooled event rates were calculated at the interim, $\hat{\pi}_C$ was then estimated under alternative hypothesis, the margin was re-set at the interim for sample size re-estimation ($\hat{\pi}_{1T} = \hat{\pi}_{1C} = \pi_{1\bullet}$).

True π_T	True π_C	Relative Margin	Type I error Rate		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.03	0.02	$-0.5 \times \hat{\pi}_C$	0.0286	0.01915	4742	4565	3736	5348
0.075	0.05	$-0.5 \times \hat{\pi}_C$	0.02876	0.01959	1725	1694	1512	1923
0.12	0.08	$-0.5 \times \hat{\pi}_C$	0.02873	0.0198	1016	1002	912	1109
0.15	0.1	$-0.5 \times \hat{\pi}_C$	0.02806	0.01956	783	769	711	849
0.18	0.12	$-0.5 \times \hat{\pi}_C$	0.02819	0.01901	640	624	581	680
0.225	0.15	$-0.5 \times \hat{\pi}_C$	0.02749	0.01926	582	581	581	581
0.27	0.18	$-0.5 \times \hat{\pi}_C$	0.0277	0.01978	581	581	581	581
0.3	0.2	$-0.5 \times \hat{\pi}_C$	0.02683	0.0193	581	581	581	581
0.33	0.22	$-0.5 \times \hat{\pi}_C$	0.02673	0.01941	581	581	581	581
0.375	0.25	$-0.5 \times \hat{\pi}_C$	0.02682	0.02018	581	581	581	581

Table A4.12a: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255 / group$. Simulation runs=10,000

Pooled event rates were calculated at the interim, $\hat{\pi}_C$ was then estimated under alternative hypothesis, the margin was re-set at the interim for sample size re-estimation ($\hat{\pi}_{1T} = \hat{\pi}_{1C} = \pi_{1\cdot}$).

True π_T	True π_C	Relative Margin	Power		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.95	0.95	$0.125 \times \hat{\pi}_C$	0.99998	0.99998	255	255	255	255
0.9	0.9	$0.125 \times \hat{\pi}_C$	0.99261	0.99302	255	255	255	255
0.85	0.85	$0.125 \times \hat{\pi}_C$	0.94402	0.94755	255	255	255	255
0.8	0.8	$0.125 \times \hat{\pi}_C$	0.85612	0.8606	258	255	255	255
0.75	0.75	$0.125 \times \hat{\pi}_C$	0.8023	0.80778	299	293	264	325
0.7	0.7	$0.125 \times \hat{\pi}_C$	0.79819	0.80374	379	379	344	415
0.65	0.65	$0.125 \times \hat{\pi}_C$	0.79646	0.80098	456	453	415	494
0.6	0.6	$0.125 \times \hat{\pi}_C$	0.7995	0.80242	592	584	538	644
0.55	0.55	$0.125 \times \hat{\pi}_C$	0.80051	0.80258	728	721	665	793

Table A4.12b: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581/ group$. Simulation runs=10,000

Pooled event rates were calculated at the interim, $\hat{\pi}_C$ was then estimated under alternative hypothesis, the margin was re-set at the interim for sample size re-estimation ($\hat{\pi}_{1T} = \hat{\pi}_{1C} = \pi_{1\cdot}$).

True π_T	True π_C	Relative Margin	Power		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.02	0.02	$-0.5 \times \hat{\pi}_C$	0.84562	0.81089	6103	5846	4565	7174
0.05	0.05	$-0.5 \times \hat{\pi}_C$	0.84528	0.81167	2207	2135	1860	2498
0.08	0.08	$-0.5 \times \hat{\pi}_C$	0.84448	0.81185	1309	1297	1157	1433
0.1	0.1	$-0.5 \times \hat{\pi}_C$	0.84286	0.81154	1016	1002	912	1109
0.12	0.12	$-0.5 \times \hat{\pi}_C$	0.8423	0.81233	822	807	745	895
0.15	0.15	$-0.5 \times \hat{\pi}_C$	0.84686	0.81843	640	624	581	680
0.18	0.18	$-0.5 \times \hat{\pi}_C$	0.88907	0.86859	584	581	581	581
0.2	0.2	$-0.5 \times \hat{\pi}_C$	0.92357	0.9089	581	581	581	581
0.22	0.22	$-0.5 \times \hat{\pi}_C$	0.94949	0.93897	581	581	581	581
0.25	0.25	$-0.5 \times \hat{\pi}_C$	0.97456	0.96945	581	581	581	581

Table A4.13a: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255/ group$. Simulation runs=100,000

Pooled event rates were calculated at the interim, $\hat{\pi}_C$ was then estimated under null hypothesis, the margin was re-set at the interim for sample size re-estimation ($\hat{\pi}_{1C} = \pi_{1\bullet} + \frac{\theta}{2}$).

True π_T	True π_C	Relative Margin	Type I error Rate		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.83125	0.95	$0.125 \times \hat{\pi}_C$	0.02124	0.03369	255	255	255	255
0.7875	0.9	$0.125 \times \hat{\pi}_C$	0.02564	0.02947	255	255	255	255
0.74375	0.85	$0.125 \times \hat{\pi}_C$	0.02615	0.02617	255	255	255	255
0.7	0.8	$0.125 \times \hat{\pi}_C$	0.02495	0.0253	255	255	255	255
0.65625	0.75	$0.125 \times \hat{\pi}_C$	0.02455	0.02598	255	255	255	255
0.6125	0.7	$0.125 \times \hat{\pi}_C$	0.02496	0.02502	255	255	255	255
0.56875	0.65	$0.125 \times \hat{\pi}_C$	0.02533	0.02532	255	255	255	255
0.525	0.6	$0.125 \times \hat{\pi}_C$	0.0254	0.02486	255	255	255	255
0.48125	0.55	$0.125 \times \hat{\pi}_C$	0.02537	0.02478	255	255	255	255

Table A4.13b: Type I error Rate for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581 / group$. Simulation runs=100,000

Pooled event rates were calculated at the interim, $\hat{\pi}_C$ was then estimated under null hypothesis, the margin was re-set at the interim for sample size re-estimation ($\hat{\pi}_{1C} = \pi_{1\bullet} + \frac{\theta}{2}$).

True π_T	True π_C	Relative Margin	Type I error Rate		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.03	0.02	$-0.5 \times \hat{\pi}_C$	0.03542	0.01993	655	663	647	667
0.075	0.05	$-0.5 \times \hat{\pi}_C$	0.02899	0.0185	581	581	581	581
0.12	0.08	$-0.5 \times \hat{\pi}_C$	0.02853	0.01826	581	581	581	581
0.15	0.1	$-0.5 \times \hat{\pi}_C$	0.0279	0.01878	581	581	581	581
0.18	0.12	$-0.5 \times \hat{\pi}_C$	0.02709	0.01871	581	581	581	581
0.225	0.15	$-0.5 \times \hat{\pi}_C$	0.02716	0.01872	581	581	581	581
0.27	0.18	$-0.5 \times \hat{\pi}_C$	0.02688	0.01921	581	581	581	581
0.3	0.2	$-0.5 \times \hat{\pi}_C$	0.02683	0.0193	581	581	581	581
0.33	0.22	$-0.5 \times \hat{\pi}_C$	0.02673	0.01941	581	581	581	581
0.375	0.25	$-0.5 \times \hat{\pi}_C$	0.02682	0.02018	581	581	581	581

Table A4.14a: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim– Positive Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.8$, $\delta = 0.1$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 255 / group$. Simulation runs=100,000

Pooled event rates were calculated at the interim, $\hat{\pi}_C$ was then estimated under null hypothesis, the margin was re-set at the interim for sample size re-estimation ($\hat{\pi}_{1C} = \pi_{1\bullet} + \frac{\theta}{2}$).

True π_T	True π_C	Relative Margin	Power		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.95	0.95	$0.125 \times \hat{\pi}_C$	1	1	255	255	255	255
0.9	0.9	$0.125 \times \hat{\pi}_C$	0.9921	0.9925	255	255	255	255
0.85	0.85	$0.125 \times \hat{\pi}_C$	0.9401	0.9433	255	255	255	255
0.8	0.8	$0.125 \times \hat{\pi}_C$	0.8484	0.8527	255	255	255	255
0.75	0.75	$0.125 \times \hat{\pi}_C$	0.7374	0.7431	255	255	255	255
0.7	0.7	$0.125 \times \hat{\pi}_C$	0.6307	0.6364	255	255	255	255
0.65	0.65	$0.125 \times \hat{\pi}_C$	0.5572	0.5613	255	255	255	255
0.6	0.6	$0.125 \times \hat{\pi}_C$	0.4565	0.4596	255	255	255	255
0.55	0.55	$0.125 \times \hat{\pi}_C$	0.3862	0.3882	255	255	255	255

Table A4.14b: Power for Blackwelder’s and Farrington-Manning’s tests in the design with adaptive non-inferiority margin in risk difference approach with sample size re-estimation at the interim – Negative Outcome

Study design parameters: Assume $\pi_T = \pi_C = 0.1$, $\delta = -0.05$, initial design sample size was calculated using Farrington-Manning’s method with $\alpha = 0.025$ and power $1 - \beta = 0.8$. Original planned sample size $n = 581 / group$. Simulation runs=100,000

Pooled event rates were calculated at the interim, $\hat{\pi}_C$ was then estimated under null hypothesis, the margin was re-set at the interim for sample size re-estimation ($\hat{\pi}_{1C} = \pi_{1\bullet} + \frac{\theta}{2}$).

True π_T	True π_C	Relative Margin	Power		Re-estimated Sample Size			
			Blackwelder	Farrington-Manning	Mean	Median	Q1	Q3
0.02	0.02	$-0.5 \times \hat{\pi}_C$	0.1391	0.1069	655.823	663	655	667
0.05	0.05	$-0.5 \times \hat{\pi}_C$	0.3246	0.2806	581	581	581	581
0.08	0.08	$-0.5 \times \hat{\pi}_C$	0.5093	0.4528	581	581	581	581
0.1	0.1	$-0.5 \times \hat{\pi}_C$	0.608	0.5609	581	581	581	581
0.12	0.12	$-0.5 \times \hat{\pi}_C$	0.7004	0.6578	581	581	581	581
0.15	0.15	$-0.5 \times \hat{\pi}_C$	0.8121	0.7771	581	581	581	581
0.18	0.18	$-0.5 \times \hat{\pi}_C$	0.8852	0.8681	581	581	581	581
0.2	0.2	$-0.5 \times \hat{\pi}_C$	0.9238	0.9081	581	581	581	581
0.22	0.22	$-0.5 \times \hat{\pi}_C$	0.9471	0.9386	581	581	581	581
0.25	0.25	$-0.5 \times \hat{\pi}_C$	0.9471	0.9386	581	581	581	581

Appendix B: PROGRAMS

Programs for continuous endpoints

```
/*The program was developed to simulate the 2-stage trial design with continuous
endpoints*/

/*
sigma0:    Originally assumed standard deviation based on previous trials
sigma:     Assumed true deviation for generating simulation data of two stage
theta:     Treatment difference specified according to previous trials (for power
calculation)
margin:    Pre-specified NI margin
alpha:     Propability of type I error specified for determining percentile of the
standard
           normal distribution used to calculate initial sample
beta:     Probability of type II error specified for determining percentile of the
standard
           normal distribution used to calculate initial sample
pi:        The proportion of subjects used in internal pilot from the originally planned
sample size
ulx:       Mean specified for treatment group (group 1) to simulate the data
u2x:       Mean specified for control group (group 2) to simulate the data
num_simu:  Number of simulations for calculating the Type I error rate and power
cp:        Cutoff of conditional power for stopping the trial for futility
*/

%macro twostage(sigma0=, sigma=, theta=, alpha=, beta=, margin=, pi=, ulx=, u2x=,
                num_simu=, cp=);
```

```

ods exclude all;

Data _null_;

/*calculate the total planned sample size per group*/
n_init_cal=2*(probit(1-&alpha)+probit(1-&beta))**2*&sigma0/((&theta-&margin)**2);

call symput('N_init', N_INIT_CAL);
Put "Initial sample size per group is " N_init_cal 6.3;

/*calculate the sample size per group of initial pilot*/
n_internal=&pi*n_init_cal;
n_internal=ceil(n_internal);

call symput('N_ips', N_INTERNAL);/*N-ips is the sample size of internal pilot study*/

Put "Initial sample size per group of internal pilot study is " N_internal 6.3;

Run;

data s1;
    retain seed_1 739213 seed_2 3126349;
do nra=1 to &num_simu;

/*generate simulation sample of internal pilot study*/
    do sample=1 to &n_ips;
        call rannor(seed_1,w1);
        call rannor(seed_2,w2);
        y1=&U1x+&sigma*w1;
        y2=&U2x+&sigma*w2;
        output;
    end;

```

```

        end;
run;

proc means data=s1;
    var y1 y2;
    class nra;
    output out=outmean1 mean=mean1 mean2 n=n1 n2 var=var1 var2;
run;

data stats;
    set outmean1;
    if nra^=.;
    meandif=mean1-mean2;
    var=((n1-1)*var1+(n2-1)*var2)/(n1+n2-2);
    z1=sqrt(n1/(2*var))*(meandif-&margin);

/*calculate the cp at the interim, if it is less than a certain cutoff, then end the
study*/
    n3=&n_init-&N_ips;
    cp=1-probnorm((probit(1-&alpha)*sqrt(2*(&n_ips+n3))-z1*sqrt(2*&n_ips)-n3*(0-
&margin)/var)/sqrt(2*n3));

    if cp<=&cp then n3=1;
        else if cp>&cp then do;
    do while(cp<1-&beta and n3<=2*&n_init);
        n3=n3+1;
        cp=1-probnorm((probit(1-&alpha)*sqrt(2*(&n_ips+n3))-z1*sqrt(2*&n_ips)-n3*(0-
&margin)/var)/sqrt(2*n3));
    end;
end;
end;

```

```

        if n3=1 then stop=1; else stop=0;
if n3>1 and cp<0.8 then cap=1; else cap=0;

keep nra n1 n2 n3 z1 cp mean1 mean2 meandif var1 var2 stop cap;
output;
run;

data s2;
    retain seed_1 729242 seed_2 489672;
    merge stats s1;
    by nra;
    output;

    if last.nra then do i=1 to n3;
        call rannor(seed_1,w1);
        call rannor(seed_2,w2);
        y1=&U1x+&sigma*w1;
        y2=&U2x+&sigma*w2;
        output;
    end;
run;

proc means data=s2;
    var y1 y2;
    class nra;
    output out=outmean mean=mean1 mean2 n=n1 n2 var=var1 var2;
run;

data outmeandif;

```

```

        set outmean;
        if nra^=.;
        meandif=mean1-mean2;
        keep meandif n1 n2 var1 var2;
run;

/*calculate the type 1 error*/
data typeI;
    merge outmeandif stats(keep=z1);

var_pool=((n1-1)*var1+(n2-1)*var2)/(n1+n2-2);

    /*calculate the test statistics*/
    t1=(meandif-&margin)*sqrt((n1)/(2*var_pool));

    lamda1=&n_ips/n1;
    lamda2=&N_init/n1;

    /*calculate the critical value*/
    t3=(probit(1-&alpha)*sqrt(&n_init)*sqrt(n1-&n_ips)-z1*sqrt(&n_ips)*(sqrt(n1-&n_ips)-
sqrt(&n_init-&n_ips)))/
        (sqrt(n1)*sqrt(&n_init-&n_ips));

    if t1>t3 then sig=1; else sig=0;
run;

proc means data=typei;
    var sig;
    output out=overall1 mean=sig;
run;

proc means data=typei;

```

```

        var n1;
        output out=overall12 mean=ss std=std;
run;

proc means data=stats;
    var stop;
    output out=overall13 mean=stop;
run;

proc means data=stats;
    var cap;
    output out=overall14 mean=cap;
run;

Data thesis;
    merge overall1 overall2 overall3 overall4;
run;

ods select all;
proc datasets library=work memtype=data;
delete s1 stats outmean outmean1 overall1 overall2 overall3 overall4 s2 outmeandif typei;
run;

%mend twostage;

%twostage(sigma0=1, sigma=1, theta=0, alpha=0.025, beta=0.2, margin=-0.2, pi=0.5,
            u1x=1.2, u2x=1.4, num_simu=100000, cp=0.1);

```

Programs for binary endpoints

```
/*The program was developed to simulate the 2-stage trial design for binary endpoints*/

/*
pt:      Originally assumed event rate for treatment group (group A)
pc:      Originally assumed event rate for control group (group B)
pc0:     Event rate specified for control group (group B) to simulate the data
theta:   Treatment difference specified according to previous trials (for power
calculation)
delta:   Pre-specified NI margin = delta*pc
alpha:   propability of type I error specified for determining percentile of the standard
         normal distribution used to calculate initial sample
beta:    Probability of type II error specified for determining percentile of the
         standard
         normal distribution used to calculate initial sample
pi:      The proportion of subjects used in internal pilot from the originally planned
         sample size
num_simu: Number of simulations for calculating the Type I error rate and power
*/

%macro bitwostage(pt=, pc=, pc0=, theta=, alpha=, beta=, delta=,
                 pi=, num_simu=);

ods exclude all;

data _null_;
/*calculate the total planning sample size per group using Farrington-Manning*/
a=1+1;
b=- (1+1+&pc+1*&pt+ (&delta*&pc) * (1+2) );
c= (&delta*&pc) **2+ (&delta*&pc) * (2*&pc+1+1)+&pc+1*&pt;
d=-&pc* (&delta*&pc) * (1+ (&delta*&pc) );
```

```

v=(b**3)/((3*a)**3)-(b*c/(6*(a**2)))+d/(2*a);
vtemp=abs(v)-v;
if vtemp eq 0 then vsign=1; else vsign=-1;
u=vsign*sqrt((b**2/(3*a)**2)-c/(3*a));
w=(1/3)*(3.141592654+arccos(v/u**3));

p1=2*u*cos(w)-b/(3*a);
p2=p1-(delta*pc);

zalpha=abs(probit(alpha));
zbeta=abs(probit(beta));

snum=zalpha*sqrt(p1*(1-p1)+p2*(1-p2)/1);
snum=snum+zbeta*sqrt(pt*(1-pt)+pc*(1-pc)/1);
sdenom=pc-pt-(delta*pc);
ssize=(snum/sdenom)**2;
n_init_cal=ceil(ssize);

call symput('N_init', N_INIT_CAL);

/*calculate the sample size per group of initial pilot*/
n_internal=&pi*n_init_cal;
n_internal=ceil(n_internal);

call symput('N_ips', N_INTERNAL);/*N_ips is the sample size of internal pilot study*/

Put "Initial sample size per group of internal pilot study is " N_internal 6.3;

Run;

/*simulate the first-stage samples*/
data s1;

```

```

do nra=1 to &num_simu;
    pt0=&pc0*(1-&delta);
    FT1 = RanBin( 739213 , &n_ips , pt0) ;
    FC1 = RanBin( 3126349 , &n_ips , &pc0 ) ;

    /*calculate the blinded event rate at the interim*/
    Percent=(FT1+FC1)/(2*&n_ips)*100;
    FT0i=&N_ips-FT1;
    FC0i=&N_ips-FC1;
    Output ;
End ;

run;
proc transpose data=s1 out=s1a;
    by nra;
run;

data s1i;
    set s1a(where=( _name_ in ("FT1" "FC1" "FT0i" "FC0i")));
    count=coll;
    if _name_="FT1" then do; group="A"; f=1; end;
    else if _name_="FC1" then do; group="B"; f=1; end;
    if _name_="FT0i" then do; group="A"; f=0; end;
    if _name_="FC0i" then do; group="B"; f=0; end;
    drop _name_ coll;
run;
proc sort data=s1i; by nra group;run;
proc freq data=s1i;
    tables f /list out=out1i;
    weight count;
    by nra group;
run;

```

```

proc sort data=outli; by nra; run;
proc transpose data=outli(where=(f=1)) out=out2i prefix=pcti;
  by nra;
  var percent;
run;

data stats;
  set s1;
  pt=percent/100+&theta/2;
  pc=percent/100-&theta/2;
  margin=&delta*pc;

/*calculate the additional sample size needed*/
a=1+1;
b=- (1+1+pc+1*pt+margin*(1+2));
c=margin**2+margin*(2*pc+1+1)+pc+1*pt;
d=-pc*margin*(1+margin);
v=(b**3)/((3*a)**3)-(b*c/(6*(a**2)))+d/(2*a);
vtemp=abs(v)-v;
if vtemp eq 0 then vsign=1; else vsign=-1;
u=vsign*sqrt((b**2/(3*a)**2)-c/(3*a));
w=(1/3)*(3.141592654+arccos(v/u**3));

p1=2*u*cos(w)-b/(3*a);
p2=p1-margin;

zalpha=abs(probit(&alpha));
zbeta=abs(probit(&beta));

snum=zalpha*sqrt(p1*(1-p1)+(1-&delta)**2*p2*(1-p2)/1);
snum=snum+zbeta*sqrt(pt*(1-pt)+(1-&delta)**2*pc*(1-pc)/1);

```

```

sdenom=pc-pt-margin;
ssize=(snum/sdenom)**2;
n_new=ceil(ssize);

if n_new<=&N_ips/&pi then do; n_new=&N_init; n3=&N_init-&N_ips;end;
if n_new>&N_ips/&pi then n3=n_new-&N_ips;
run;
data s2a;
    retain seed1 729242 seed2 489672;
    set stats;
    pt0=&pc0*(1-&delta);
    call RanBin(seed1 , n3 , pt0, FT2) ;
    call RanBin(seed2 , n3 , &pc0, FC2 ) ;
    FT=FT1+FT2;
    FC=FC1+FC2;
    FT0=&N_ips+N3-FT;
    FC0=&N_ips+N3-FC;
run;

proc transpose data=s2a out=try;
    by nra;
run;

data s2;
    set try(where=( _name_ in ("FT" "FC" "FT0" "FC0")));
    count=coll;
    if _name_="FT" then do; group="A"; f=1; end;
    else if _name_="FC" then do; group="B"; f=1; end;
    if _name_="FT0" then do; group="A"; f=0; end;
    if _name_="FC0" then do; group="B"; f=0; end;
    drop _name_ coll;
run;

```

```

/*normal approximation*/
proc sort data=s2; by nra group;run;
proc freq data=s2;
  tables f /list out=out3;
  weight count;
  by nra group;
run;
proc sort data=out3; by nra; run;
proc transpose data=out3(where=(f=1)) out=out4 prefix=pct;
  by nra;
  var percent;
run;

/*calculate the type 1 error rate*/
data typeI;
  merge stats(keep=nra margin n_new) out4(keep=nra pct1 pct2);
  marginf=&delta*pct2/100;

/*FM method -- self calculation*/
  a=1+1;
  b=-(1+1+pct2/100+1*pct1/100+marginf*(1+2));
  c=marginf**2+marginf*(2*pct2/100+1+1)+pct2/100+1*pct1/100;
  d=-pct2/100*marginf*(1+marginf);
  v=(b**3)/((3*a)**3)-(b*c/(6*(a**2)))+d/(2*a);
  vtemp=abs(v)-v;
  if vtemp eq 0 then vsign=1; else vsign=-1;
  u=vsign*sqrt((b**2/(3*a)**2)-c/(3*a));
  w=(1/3)*(3.141592654+arccos(v/u**3));

  p1=2*u*cos(w)-b/(3*a);

```

```

p2=p1-marginf;
se_fm=sqrt(p1*(1-p1)/n_new + ((1-&delta)**2)*(p2*(1-p2)/n_new ));

z_fm=(pct2/100-pct1/100-marginf)/se_fm;
if -z_fm<probit(&alpha) then sig_fm1=1; else sig_fm1=0;

/*normal approximation with unpooled phat*/
se_unpool=sqrt(pct1/100*(1-pct1/100)/n_new + (1-&delta)**2*(pct2/100*(1-
pct2/100)/n_new ));
z_unpool=(pct2/100-pct1/100-(&delta*pct2/100))/se_unpool;
if -z_unpool<probit(&alpha) then sig_unpool=1; else sig_unpool=0;

run;

/*type I error calculated usinf normal approximation*/
proc means data=typei;
var sig_unpool;
output out=overall11 mean=sig_unpool;
run;

/*type I error calculated usinf FM method*/
proc means data=typei;
var sig_fm1;
output out=overall15 mean=sig_fm1;
run;

/*descriptive statistics of final sample size*/
proc means data=typei;
var n_new;
output out=overall12 mean=ss std=stds median=meds q1=q1s q3=q3s;
run;

```

```

/*descriptive statistics of margin at interim*/
proc means data=typei;
    var margin;
    output out=overall13 mean=mgf std=stdmi median=medmi q1=q1mi q3=q3mi;
run;

/*descriptive statistics of margin at final*/
proc means data=typei;
    var marginf;
    output out=overall14 mean=mgf std=stdmf median=medmf q1=q1mf q3=q3mf;
run;

Data cmpap;
    merge overall11 overall12 overall13 overall14 overall15;
    drop _freq_ _type_;
run;

ods select all;
PROC DATASETS LIBRARY = work;
SAVE cmpap; QUIT; RUN;
%mend bitwostage;

%bitwostage(pt=0.8, pc=0.8, pc0=0.95, theta=0, alpha=0.025,
            beta=0.2, delta=0.125, pi=0.5, num_simu=1000000, order=1);

```

REFERENCES

- 1: Temple R, Ellenberg SS. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments: part 1: ethical and scientific issues. *Annals of Internal Medicine* 2000; 133:455–463.
- 2: Wang SJ, Hung HMJ, Tsong Y, Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 2001; 20:1903 –1912.
- 3: D’Agostino Sr, RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; 22:169 –186.
- 4: Denne JS. Sample size recalculation using conditional power. *Statistics in Medicine*. 2001; 20:2645-2660.
- 5: Posch M, Bauer P. Interim analysis and sample size reassessment. *Biometrics* 2000; 56:1170-1176.
- 6: Liu Q, Chi GYH. On sample size and inference for two-stage adaptive designs. *Biometrics* 2000; 57:172-177.
- 7: Blackwelder WC. Proving the null hypothesis. *Controlled Clinical Trials* 1982; 3:345–353.
- 8: Proschan MA, Hunsberger SA. Designed extension of studies based on conditional

power. *Biometrics* 1995; 51:1315-1324.

9: Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; 55:853-857.

10: Birkett MA, Day SJ. Internal pilot studies for estimating sample size. *Statistics in Medicine* 1994; 13:2455-2463.

11: International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH harmonized tripartite guideline: Statistical principles for clinical trials. *Statistics in Medicine* 1998; 18:1905-1942.

12: Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* 1990; 9:65 –72.

13: Wittes J, Schabenberger O, Zucker D, Brittain E, Proschan M. Internal pilot studies I: Type I error rate of the naïve t-test. *Statistics in Medicine* 1999; 18:3481-3491.

14: Gould AL. Interim analyses for monitoring clinical trials that do not affect the Type I error rate. *Statistics in Medicine* 1992; 11:55-66.

15: Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* 1945; 16:243-258.

16: Zucker DM, Wittes JT, Schabenberger O, Brittain E. Internal pilot studies II: comparison of various procedures. *Statistics in Medicine* 1999; 18:3493-3509.

17: Denne JS, Jennison C. Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine* 1999; 18:1575-1585.

- 18: Kieser M, Friede T. Recalculating the sample size in internal pilot study designs with control of the Type I error rate. *Statistics in Medicine* 2000; 19:901-911.
- 19: Gould AL. Planning and revising the sample size for a trial. *Statistics in Medicine* 1995; 14:1039-1051.
- 20: Gould AL, Shih WJ. Sample size recalculation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics – Theory and Methods* 1992; 21:2833-2853.
- 21: Gould AL, Shih WJ. Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* 1998; 17:89-100.
- 22: Friede T, Kieser M. On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine* 2002; 21:165-176.
- 23: Lan KKG, Wittes J. The B-value: a tool for monitoring data. *Biometrics* 1988; 44:579-585.
- 24: Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; 50:1029-1041. Correction in *Biometrics* 1996; 52:380.
- 25: Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; 17:1551-1562.
- 26: Shun Z, Yuan W, Brady WE, Hsu H. Type I error in sample size re-estimations based on observed treatment difference. *Statistics in Medicine* 2001; 20:497-513.

- 27: Muller H-H, Shafer H. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; 55:886–891.
- 28: Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; 22:971-993.
- 29: Wassmer G. Multi-stage adaptive test procedures based on Fisher's product criterion. *Biometrical Journal* 1999; 41:279-293.
- 30: Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; 97:236-244.
- 31: Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; 55:1286-1290.
- 32: Wassmer G, Eisebitt R, Coburger S. Flexible interim analyses in clinical trials using multistage adaptive test designs. *Drug Information Journal* 2001; 35: 1131-1146.
- 33: Shen Y, Fisher L. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* 1999; 55:190-197.
- 34: Denne, JS. Estimation following extension of a study on the basis of conditional power. *Journal of Biopharmaceutical Statistics* 2000; 10:131-144.
- 35: Dunnett C, Gent M. Significant testing to establish equivalence between treatments with special reference to data in the form of 2×2 tables, *Biometrics*, 1977; 33:593-602

- 36: Friede T, Kieser, M. Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics* 2004; 3:269-279.
- 37: Friede T, Kieser, M. Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal* 2006; 4, 537-555.
- 38: Shih WJ, Zhao P-L. Design for sample size reestimation with interim data for double-blind clinical trials with binary outcomes. *Statistics in Medicine* 1997; 16:1913–1923.
- 39: Friede T, Mitchell C, Muller-Velten G, M. Blinded sample size reestimation in non-inferiority trials with binary endpoints. *Biometrical Journal* 2007; 6:903–916.
- 40: Herson, J. and Wittes, J. The use of interim analysis for sample size adjustment. *Drug Information Journal* 1993; 27:753–760.
- 41: Friede T, Kieser, M. Blinded Sample size reestimation in non-inferiority trials with binary endpoints. *Biometrical Journal* 2007; 49:903–916.
- 42: Friede T, Kieser, M. Sample size adjustment in clinical trials for proving equivalence. *Drug Information Journal* 2001; 35:1401–1408.
- 43: Friede T, Kieser, M. Blinded sample size reassessment in non-inferiority and equivalence trials. *Statistics in Medicine* 2003; 22:995–1007.
- 44: Wang C, Keller D.S., Lan K.K.G. Sample size re-estimation for binary data via conditional power. *Joint Statistical Meeting, Biopharmaceutical Section* 2002; 3621-2626
- 45: Gould L. Sample size re-estimation: recent developments and practical considerations. *Statistics in Medicine* 2001; 20:2625–2643.

- 46: Phillips K. A new test of non-inferiority for anti-infective trials. *Statistics in Medicine* 2003; 22:201–212.
- 47: U.S. Food and Drug Administration. Point-to-Consider. Division of Anti-infective Drug Products, Clinical Development and Labeling of Anti-Infective Drug Products, 1992.
- 48: Seelbinder, B. M. On Stein's two-stage sampling scheme. *The Annals of Mathematical Statistics* 1953; 24:640–649.
- 49: Moshman, J. A method for selecting the size of the initial sample in Stein's two sample procedure. *The Annals of Mathematical Statistics* 1958; 29: 1271-1275.
- 50: Sandvik, L., Erikssen, J., Mowinckel, P. and Rodland, E. A. A method for determining the size of internal pilot studies. *Statistics in Medicine* 1996; 15:1587–1590.
- 51: Browne, R. H. On the use of a pilot sample for sample size determination. *Statistics in Medicine* 1995; 14:1933–1940.
- 52: Friede T., Kieser M., Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal* 2006; 48: 537-555
- 53: Singer, J. Letter to the editor: A method for determining the size of internal pilot studies. *Statistics in Medicine*; 1999; 18:1151–1153.
- 54: Ellenberg S.S., Temple R.; Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues, *Ann Intern Med* 2000; 133:455-463

- 55: Hung H.M.J., Wang S.-J., Tsong Y.; et al. Some fundamental issues with non-inferiority testing in active controlled trials, *Stat Med* 2003; 22:213-225
- 56: Siegel J.P.; Equivalence and non-inferiority trials, *Am Heart J* 2000; 139:S166-S170
- 57: Proschan MA. Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; 51:1315 –1324.
- 58: Herson J. Use of predictive power in mid-course correction. Presented at the Predictive Power Session (ST), Drug Information Association, Annual Meeting, 1998, Boston, U.S.A.
- 59: Denne J. Sample size recalculation using conditional power. *Statistics in Medicine* 2001; 20:2645-2660.
- 60: Lan KKG, Wittes J. The B-value: a tool for monitoring data. *Biometrics* 1988; 44:579-585.
- 61: Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; 50:1029-1041. Correction in *Biometrics* 1996; 52:380.
- 62: Fisher LD. Self-designing clinical trials. *Statistics in Medicine* 1998; 17:1551-1562.
- 63: Farrington C, 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 Medicine* 1990; 9:1447–1454.

64: Robruck P, Kuhn A. Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. *Statistics in Medicine* 1995; 14:1583–1594.

CURRICULUM VITAE

