An Evaluation of Flexible Summary Measures for the Comparison of Binary Outcomes in Non-Inferiority Trials

Erika Thommes

A thesis submitted in
partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2015

Committee:

Susanne May, Chair

Scott Emerson

Program Authorized to Offer Degree:
UW School of Public Health, Department of Biostatistics

University of Washington

Abstract

An Evaluation of Flexible Summary Measures for the Comparison
of Binary Outcomes in Non-Inferiority Trials

Erika Thommes

Chair of the Supervisory Committee:
Associate Professor Susanne May
Department of Biostatistics

In clinical trials, the comparison of binary outcomes between two independent treatment groups is most commonly measured by either relative or absolute differences between outcome rates. In the setting of a non-inferiority (NI) trial, an NI margin corresponding to one of these measures is defined to represent the maximum clinically meaningful limit by which an experimental intervention will be considered allowably inferior to a standard-of-care (SOC) treatment regimen. In the instance of extreme event rates, special consideration should be given to the intervention's allowable outcome rates as defined by the SOC rate and the NI margin in order to produce a meaningful assessment of the intervention's impact on public health.

We propose one method by which the definition of inferiority can be relaxed in the case of extremely rare failure events. Using the failure rate among subjects receiving the SOC, we introduce a clinically meaningful threshold at which the comparison of treatment groups will switch from a conservative relative comparison to a potentially more meaningful and interpretable absolute comparison. This threshold is to be defined at a failure rate at which study investigators feel comfortable enough with the rarity of events among subjects receiving the SOC such that they are willing to increase the allowable failure rate among

those receiving the intervention. This threshold is further defined in a manner that maintains a continuous margin by which NI can be judged for two binary outcomes throughout the parameter space.

Focusing on asymptotic methods, we compare statistical inference based on the Wald, score and likelihood ratio (LR) statistics under this proposal with that of a standard relative comparison of outcome rates. We illustrate the potential advantages of our proposal based on the relaxed assumption of inferiority for extremely low failure rates. We establish the type 1 error and coverage probability of our proposed method against relative and absolute comparisons. Finally, we compare the commonalities and differences in the statistical behavior of these three asymptotic methods under a fixed trial design versus a group sequential sampling design.

Using a one-sided significance level of 2.5%, our results indicate that the type 1 error rate under our threshold proposal is similar to that of a relative comparison when the observed SOC is above the threshold and to that of an absolute comparison when the observed SOC failure rate is at or below the threshold. Similarly, coverage probability remains stable at approximately 95% for the Wald, score and LR-based 95% confidence intervals under both fixed and sequential sampling designs, with slightly more variability in the latter. The potential for achieving non-inferiority when there truly is no difference in failure rates between subjects in the two treatment groups increased by up to 40% for a 7% threshold and up to 30% for a 5% threshold under the investigated scenarios. Marginal gains in power of up to 15% exist when detecting failure rates within the region of relaxed inferiority, but the most promising gains in power (up to 40% in our simulations) are observed in the region deemed NI by both the relative and absolute NI margins. We conclude that our threshold proposal increases the probability of detecting meaningfully non-inferior interventions without significantly increasing type 1 error or decreasing coverage probability.

Future work relating to this research might further investigate a trial design in which more than one threshold and corresponding rule for comparison of treatments can be invoked for a series of rare outcome rates. Or, one might alternatively consider the statistical behavior when switching from a relative to an absolute comparison in the case of extremely frequent outcomes. Overall, this work is a starting point by which the statistical comparison of two treatment arms can become more flexible in order to adhere most closely to a meaningful scientific comparison.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I thank Susanne May for her unwavering encouragement and assistance throughout the completion of this thesis. I thank Scott Emerson for his insightful commentary on this thesis (and beyond). I cannot express enough gratitude towards Gitana Garofalo for her commitment to my success in this program. Lastly, I extend a huge thank you to my classmates for their inspiration, collaboration, and friendship.

DEDICATION

To Mom and Dad, Brian, Vince, Stephanie, Greg, Penelope, Mark, Meghan, Emma, Craig, Maribel, Leilah, Ellen, Gary, Maria, Kelly, Anna, Emily, Marge, Jane, Kiana, Ryan, and especially to Art.

Chapter 1
INTRODUCTION

1.1 Background

Clinical trials are investigations in human subjects intended to discover or verify the efficacy and safety of an experimental treatment in promoting the health of a population. The U.S. Food and Drug Administration (FDA) is responsible for protecting and promoting public health through its regulation and monitoring of clinical studies. FDA regulations for adequate and well-controlled clinical trials describe four types of concurrent controls that can be used in order to obtain a valid quantitative assessment of a treatment's effect: placebo, dose-comparison, no treatment, and active treatment. The placebo, dose-comparison, and no treatment types are variations of a superiority trial in which the objective of the study is to show that a new treatment is superior to a control. The active treatment type is typically used in a trial for which the goal of the study is to show that a new treatment is not unacceptably inferior to an active control. This latter type of trial is often referred to as a non-inferiority (NI) trial and an active control is another name for a standard-of-care (SOC) treatment regimen for which clinical benefit has already been shown [1].

The need for NI trials typically arises when an experimental intervention (EXP) is believed to have slightly higher, no different, or slightly lower efficacy than an SOC, but at the same time has some notable advantage to the SOC. Examples of such advantages might be easier treatment administration, lower toxicities, a more preferable safety profile, a lower cost to treat, a greater resource abundancy, or lesser biological resistance. As advancements

in medical research and changes in medical practices continue to be made, the demand for NI trials and the need for rigorous scientific and regulatory evidence to assess the efficacy of an EXP relative to an SOC will continue to grow.

The unique statistical and scientific considerations that set NI trials apart from other clinical trials are well documented. One important element of NI trials that does not exist amongst superiority trials is the NI margin, or the maximum clinically meaningful limit by which the EXP is allowed to be inferior to an SOC. After consideration of all clinical data from the NI trial, results that with high confidence exclude an unacceptable degradation of the SOC effect would be interpreted as evidence of a treatment that provides benefit beyond no treatment. The NI margin is determined from the estimated advantage of SOC over no treatment as derived from previous randomized clinical trials. For example, a common choice is that the NI margin would correspond to having high confidence that a comparison between EXP and no treatment would retain at least half the effect that was previously estimated for SOC versus no treatment. We do not here address the many difficult issues that must be considered in choosing a NI margin or in designing a NI trial, but the interested reader is referred to other significant research in this area [1, 2, 3, 4, 5, 6, 7].

In this research we are primarily interested in how an otherwise acceptable statistical analysis plan and NI margin might be modified in the presence of observed low event rates on the SOC arm. When outcome rates are particularly low, sample size and power calculations can potentially return invalid approximations, even more so when deviations from the hypothesized SOC outcome rate are expected [8]. Therefore, it may be of particular interest to investigate a method by which we can increase the power to detect a meaningful NI between two low outcome rates while still maintaining clinical relevance and statistical validity. To the best of our knowledge, research regarding our proposal (as described in the following sections) has not yet been addressed.

1.2 Setting

Consider an uncensored outcome of interest defined as whether or not a subject experiences some undesirable event such as death or onset of disease by the end of a study's observation period. This failure event is a binary outcome for which the two most common summary measures for the comparison of two groups are a risk difference (RD) and relative risk (RR). The NI margin is traditionally defined according to which of these two scales study investigators believe to be most appropriate for the trial setting. Among the many aspects that investigators will consider in determining the NI margin, the expected failure rate among subjects receiving the SOC, and the implications of the choice of scale in regards to which event rates are deemed non-inferior among subjects receiving the EXP, are of utmost importance. We are particularly interested in the setting in which prior estimates of treatment effect comparing SOC to no treatment may no longer be relevant due to changes in ancillary care or in the population, and low event rates are expected. As such, we are more concerned in modifying a fixed margin approach to demonstrating non-inferiority, rather than a synthesis method approach [1].

To illustrate the disparities between the RD (i.e. absolute difference) and RR (i.e. relative difference) scales under a fixed margin approach to comparing event rates, consider Table 1.1 in which p_{soc} and p_{exp} represent failure rates corresponding to the SOC and EXP treatment arms, respectively, $\Delta = p_{exp} - p_{soc}$ is the measure for RD and $\theta = p_{exp}/p_{soc}$ is the measure for RR. Examples A and B list an identical range of possible failure rates for the SOC treatment arm beginning at 30% and decreasing to 10%. The two examples differ in that Example A provides EXP failure rates and corresponding relative risks according to a constant risk difference of 7.5% and Example B provides EXP failure rates and corresponding risk differences according to a constant relative risk of 1.5.

Table 1.1: Summary Measures for Comparing Various Outcome Rates

Example A				Example B				Example C			
Δ	θ	p_{soc}	p_{exp}	Δ	θ	p_{soc}	p_{exp}	Δ	θ	p_{soc}	p_{exp}
7.5%	1.25	30%	37.5%	15%	1.5	30%	45%	2.5%	1.5	5%	7.5%
7.5%	1.3	25%	32.5%	12.5%	1.5	25%	37.5%	2%	1.5	4%	6%
7.5%	1.375	20%	27.5%	10%	1.5	20%	30%	1.5%	1.5	3%	4.5%
7.5%	1.5	15%	22.5%	7.5%	1.5	15%	22.5%	1%	1.5	2%	3%
7.5%	1.75	10%	17.5%	5%	1.5	10%	15%	0.5%	1.5	1%	1.5%

Under the constant 7.5% RD of Example A, a 30% failure rate among subjects receiving the SOC corresponds to a 37.5% failure rate among subjects receiving EXP and a RR of 1.25. A 25% failure rate among subjects receiving the SOC corresponds to a 32.5% failure rate among subjects receiving the EXP, but this time to a RR of 1.3. In fact, for a constant risk difference, as the SOC failure rate decreases, the corresponding relative risk between the two treatments increases.

Under the constant 1.5 RR of Example B, a 10% failure rate among subjects receiving the SOC corresponds to a 15% failure rate among subjects receiving EXP and a 5% RD. A 15% failure rate among subjects receiving the SOC corresponds to a 22.5% failure rate among subjects receiving the EXP and a 7.5% RD. Hence, for a constant relative risk, as the SOC failure rate decreases, the corresponding risk differences between the two treatments also decreases.

Two immediate conclusions can be drawn from the results of Examples A and B. First, for $p_{soc} = 15\%$, an absolute difference of 7.5% and a relative risk of 1.5 both equate to $p_{exp} = 22.5\%$. Second, whether or not p_{soc} is above or below that threshold at which the two summary measures produce an equal p_{exp} determines which of the two summary measures corresponds to a higher p_{exp}. For instance, when $p_{soc} = 20\%$, a risk difference of 7.5% equates to $p_{exp} = 27.5\%$, which is 2.5% *lower* than the $p_{exp} = 30\%$ defined by the relative risk of

1.5. Conversely, when $p_{soc} = 10\%$, a risk difference of 7.5% equates to $p_{exp} = 17.5\%$, which is 2.5% *higher* than the $p_{exp} = 15\%$ defined by the relative risk of 1.5.

As described above, understanding the interplay between the SOC failure rate and relative versus absolute differences in the EXP failure rate is crucial to understanding the implications that a choice of scale has in regards to the NI of an EXP relative to an SOC. In the design of an NI trial, if investigators believe that SOC event rates will be low and they would like to be conservative in their tolerance for NI, then generally speaking, the relative risk might be a more reasonable scale to work on than the risk difference. If it is believed that event rates will be high, then the risk difference would be the more reasonable scale on which to make conservative inference.

Now consider Example C, which provides extremely rare failure rates for the SOC arm, and assume that investigators have set $\theta = 1.5$ as a relative NI margin. By setting the NI margin at 1.5, $p_{soc} = 4\%$ allows for p_{exp} to be up to 6%; $p_{soc} = 2\%$ allows for p_{exp} to be up to 3%. In these instances of extremely rare event rates, the value of the more conservative relative NI margin depreciates due to the insignificance of the corresponding absolute differences between failure rates. That is, for a 2% failure rate among subjects receiving the SOC, investigators may want to reconsider claiming inferiority of an EXP for failure rates above 3%, as this truly equates to a minimal 1% risk difference.

1.3 Proposal

Given that our setting of interest is one in which we believe the targeted population has changed from previous RCT and extremely low event rates are likely, Example C is just one of many plausible situations in which modification of the traditional approaches towards non-inferiority may be appropriate. In other words, to judge non-inferiority by the preservation of some relative effect of an SOC against no treatment (e.g. 50%) might not be meaningful if failure rates are extremely low or the target population is different than those from previously

conducted RCT.

We emphasize that the general framework of our following proposal be only applied under an appropriate trial setting in which the sample population adequately reflects the underlying target population of interest. It is known that a major benefit to the relative risk summary measure lies in its ability to maintain meaningful inference in the presence of a contaminated sample population, where we define contamination to mean excess enrollment of subjects who are particularly susceptible or particularly nonsusceptible to treatment effects. For example, if a study were to over enroll from a group of subjects that is nonsusceptible to treatment, then although failure rates will decrease in both treatment groups, the relative risk would remain constant. The risk difference, on the other hand, would underestimate the treatment effects in the target population. Thus, the appropriateness of our proposal weighs heavily on its assumption that our sample population is the target population and therefore the risk difference is a meaningful measure of treatment effect.

In this thesis, we propose the introduction of a clinically meaningful threshold, based on the failure rate among subjects receiving the SOC, at which the comparison of treatment groups will switch from a conservative relative comparison to a potentially more meaningful and interpretable absolute comparison. This threshold is to be defined for a failure rate at which investigators feel comfortable enough with the rarity of events among subjects receiving the SOC such that they are willing to relax the allowable failure rate among those receiving the EXP. This threshold must also be defined in a manner that produces a continuous margin by which NI can be judged in the parameter space for comparing two binary outcomes (i.e. in a manner such that we can determine whether or not an EXP is non-inferior to an SOC for all potential pairs of failure outcomes).

1.4 Objectives

This thesis aims to describe the statistical inference based on and the statistical behavior of an analysis of two failure rates in which the choice of an absolute or relative comparison for NI is based on the observed failure rate within a group of subjects receiving an SOC treatment regime. Chapter 2 describes a clinical trial setting in which this type of analysis might be appropriately applied. Chapter 3 details the statistical methodology and inference relating to the comparison of two binary outcomes in an NI setting along with various trial design considerations. In particular, we provide a framework for inference based on the Wald, score, and likelihood ratio (LR) statistics under fixed and sequential sampling schemes and various outcome scenarios. We use type 1 error, power and coverage probability to summarize the statistical behavior of these statistics. Chapter 4 showcases the simulation results under these settings and Chapter 5 speaks to the implications, limitations and potential for future work in this research.

Chapter 2

MOTIVATION

2.1 Innovative Treatments in Pneumonia

Innovative Treatments in Pneumonia (ITIP) is a Malawi-based study comprised of two clinical trials for which enrollment of children is expected to begin by the end of this year. Sponsored by the Program for Appropriate Technology in Health (PATH), the study aims to build evidence regarding the most appropriate length of treatment with amoxicillin dispersible tablets (DT), the first-line treatment recommendation for childhood pneumonia by the World Health Organization (WHO). At issue is the lack of definitive diagnosis of bacterial pneumonia that is sensitive to amoxicillin. Hence, a hypothesis considered in this research is that variation from the standard recommendation might pose no safety issue, because the empiric treatment with amoxicillin is not truly effective in the target population. A NI design is chosen over a superiority study because of the burden of proof required to change from the WHO recommendation. Clinical equipoise suggests that we would not demand proven superiority to continue the empiric use of amoxicillin, but if we can establish that we do not do markedly worse by delaying treatment until further complications arise, we might modify the practice.

ITIP-1, the trial of particular interest for this research, is a double-blind, 1:1 randomized NI trial with the objective to assess the effectiveness of no antibiotic treatment (placebo) versus empirical treatment with 3 days of oral amoxicillin DT (SOC). The primary endpoint will be the proportion of children who fail treatment (develop worsening signs and symptoms), where failure is tentatively defined as any of the following by day 4 of follow up: WHO danger signs, oxygen saturation by pulse oximetry less than 90%, chest-indrawing,

documented axillary temperature greater than or equal to 38°C in the absence of diagnosed co-infection with fever symptoms (e.g. malaria), vomiting within 30 minutes of two or more doses of study product, change in antibiotics, hospitalization due to pneumonia if not initially admitted, prolonged hospitalization or readmission due to pneumonia if initially admitted, or death. The study population will include 2,000 HIV-I seronegative children 2 to 59 months of age who present to Kamuzu Central Hospital in Lilongwe, Malawi with fast breathing pneumonia [9].

2.1.1 Scientific Rationale

Evidence that amoxicillin DT is an effective regimen for the treatment of bacterial pneumonia is widely accepted: The WHO's Integrated Management of Childhood Illness (IMCI) guidelines identify children with a cough or difficulty breathing who demonstrate fast breathing as having pneumonia, for which the WHO-recommended treatment regimen is antiobiotic amoxicillin DT. A Cochrane review determined that, among children 5 years of age and younger with fast-breathing pneumonia, a 3-day course of oral antibiotics is statistically nonsignificantly different than a 5-day course [10]. Furthermore, treatment regimens in which the oral amoxicillin DT is given twice per day is a feasible and comparably effective alternative to the three doses per day used in children with acute respiratory infections.

Although evidence of the effectiveness of amoxicillin DT for the treatment of bacterial pneumonia is clear, not all instances of fast-breathing are bacterial pneumonia, and there is a great need to not unnecessarily treat those children with similar symptoms but other illnesses that would not benefit from the antibiotics. An elimination of unnecessary antibiotic treatment of children would reduce the rate of adverse events associated with its unnecessary treatment and reduce issues relating to emerging antibiotic resistance. This is especially true in the resource-limited and malaria-endemic region of Malawi.

Thus, the scientific justification for the ITIP-1 clinical trial lies in the issue of proper

antibiotic treatment administration. The trial is not concerned with the notion of placebo being as effective or more effective than amoxicillin DT among children who truly would benefit from the treatment, but rather that the treatment might be being given to those who would not benefit. As such, policymakers and stakeholders expect low failure event rates among those randomized to the amoxicillin DT treatment arm and, in general, want to be conservative in the proportion of allowable failure events among those randomized to placebo.

Due in part to this unique resource-limited setting, we propose that if failure events are rare enough, it might be of interest to replace the more stringent relative NI margin with a slightly less conservative absolute NI margin. In other words, if the event rate is so low in the enrolled population as to demand a high number needed to treat (NNT) in order to prevent one complication, then the no-treatment strategy is within an acceptable NI margin. This allows for not only the potential for improved resource allocation in Malawi, but also a more meaningful assessment of the SOC's current impact on public health.

2.1.2 Statistical Rationale

Because study investigators believe there to be heterogeneity regarding the appropriateness of amoxicillin DT treatment for children with fast-breathing, there is also some uncertainty regarding the expected failure event rate among those children receiving it. This uncertainty further warranted investigators to choose a relative comparison of treatment groups so that the more conservative relative NI margin is maintained over a series of low failure rates.

The ITIP-1 trial is planned with a sample size of 2,000 subjects (1,000 per treatment arm). On the following page, Table 2.1 illustrates the power of the study to detect no difference between the two treatment groups for various failure rates given a lower one-sided alternative hypothesis of equal failure rates, a type 1 error rate of 2.5%, a relative NI margin of 1.5, and study designs with 1, 2, or 4 pre-planned analyses.

A sequential study is one in which pre-planned analyses are conducted prior to a clinical trials achievement of maximum enrollment to determine whether it should be terminated early for reasons such as sufficient evidence of an intervention being harmful or beneficial. Since repeated analyses will inflate the type I error of a study, various group sequential methods have been developed to adjust the significance level for each analysis. As will be discussed later, two such examples are the O'Brien-Fleming and Pocock group sequential methods [11, 12]. In Table 2.1, study designs with 2 or 4 pre-planned analyses assume O'Brien-Fleming boundaries for early stopping due to NI and Pocock boundaries for early stopping due to inferiority. We note that the introduction of interim analyses without increasing the maximal sample size of a trial will result in a loss of power. (See Appendix C for example code.)

Table 2.1: Power to Detect Alternative of Equal Failure Rates[1]

	Number of Analyses		
Failure Rate	1	2	4
3%	50.2%	48.3%	46.1%
4%	62.6%	60.3%	57.8%
5%	72.7%	70.3%	67.7%
6%	80.6%	78.2%	75.7%
7%	86.6%	84.4%	82.0%
8%	90.9%	89.0%	86.8%
9%	94.0%	92.3%	90.5%
10%	96.1%	94.7%	93.2%

[1]N = 2000; Relative NI Margin = 1.5

As shown, restrictions in sample size lead to low power for detecting no difference between treatment groups in the instances with extremely rare outcome rates. For example, if the trial plans two analysis times (one interim and one final analysis), the power to declare NI

in the presence of a 4% failure rate in both groups is 60.3%. We argue that, especially due to a restricted sample size, by switching from a relative to an absolute comparison for extremely rare outcomes, we are increasing our probability of deciding for NI in settings that no treatment would be clinically preferred.

Chapter 3

STATISTICAL METHODS

3.1 Notation and Asymptotic Theory

We define x_{ij} to be an indicator that the ith subject on the jth treatment arm (j=0 for placebo, j=1 for SOC) recorded some failure event of interest. Note that in a typical NI setting, the j=0 treatment arm would correspond to the SOC and j=1 arm to an EXP. The random variable X_{ij} is then said to be distributed according to a Bernoulli distribution B(1, p_j) where p_j is the unknown true probability of a failure event for a subject in the jth treatment arm. For n_j independent subjects, $X_j = \sum_{i=1}^{n_j} X_{ij}$ is distributed according to a Binomial distribution B(n_j, p_j).

When sample size allows, the use of asymptotic methods is generally most common for the comparison of two independent binary outcomes. According to the Central Limit Theorem, as sample size becomes large the Binomial distribution can be approximated by the Normal distribution, and in that case, the distribution of an observed proportion $\hat{p}_j = X_j/n_j$ is approximately Normal with mean p_j and variance $p_j(1-p_j)/n_j$. The standard error for each treatment group's estimated proportion is therefore based on the mean-variance relationship in which the variance is a function of the mean p_j. Notationally,

$$\sqrt{n}[\hat{p_0} - p_0] \xrightarrow{D} N(0, p_0(1-p_0)) \quad \text{and} \quad \sqrt{n}[\hat{p_1} - p_1] \xrightarrow{D} N(0, p_1(1-p_1)) \tag{3.1}$$

where $\xrightarrow{D}$ indicates convergence in distribution. For inference based on the absolute difference between proportions, we define $\Delta = p_0$ - p_1 as the summary measure of interest for comparing

treatment effects between arms and $\Delta_0 > 0$ as the risk difference NI margin. The null and alternative hypotheses to be considered are

$$H_0 : \Delta = p_0 - p_1 \geq \Delta_0 \qquad vs. \qquad H_A : \Delta = p_0 - p_1 < \Delta_0. \tag{3.2}$$

For two independent approximately normal random variables $\hat{p}_0$ and $\hat{p}_1$, the expected value and standard error of $\hat{\Delta} = \hat{p}_0 - \hat{p}_1$ are defined as

$$\mathbf{E}(\hat{\Delta}) = p0 - p1 \tag{3.3}$$

$$SE(\hat{\Delta}) = \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1}. \tag{3.4}$$

Thus, we obtain an asymptotic distribution of

$$\sqrt{n}[\hat{\Delta} - \Delta] \xrightarrow{D} N\left(0, \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1}\right). \tag{3.5}$$

For statistical inference based on the ratio of proportions, we define $\theta = p_0/p_1$ as the relative summary measure of interest for comparing treatment effects between treatment arms. The NI margin $\theta_0 > 1$ is defined as the factor by which the increase in the percentage of failure events in the EXP group is not believed to be unacceptably worse than that in the SOC group. The null and alternative hypotheses to be considered are

$$H_0 : \theta = \frac{p_0}{p_1} \geq \theta_0 \qquad vs. \qquad H_A : \theta = \frac{p_0}{p_1} < \theta_0. \tag{3.6}$$

Various methods can be applied to obtain the distribution of $\hat{\theta}$. Using a logarithmic transformation and applying the delta method (see Appendix A for derivations), we obtain an expected value and standard error of

$$\mathbf{E}(\log(\hat{\theta})) = \log(p_0) - \log(p_1) \tag{3.7}$$

$$SE(\log(\hat{\theta})) = \frac{1-p_1}{p_1 n_1} + \frac{1-p_0}{p_0 n_0}. \tag{3.8}$$

The estimated asymptotic distribution of $\hat{\theta}$ is therefore

$$\log(\hat{\theta}) \dot{\sim} N(\log(p_0) - \log(p_1), \frac{1-p_1}{p_1 n_1} + \frac{1-p_0}{p_0 n_0}).$$

Based on the above theory, $\hat{\Delta}$ and $\hat{\theta}$ are the summary measures of immediate concern to us. Our interest in the individual estimates of p_0 and p_1 lies in the calculation of the variances of these summary measures. As we discuss in further detail in the next section, maximum likelihood estimates for p_0 and p_1 can be calculated two ways: (1) $\hat{p}_0$ and $\hat{p}_1$ represent the unrestricted maximum likelihood estimates (MLEs), or sample estimates, of p_0 and p_1; and (2) $\tilde{p}_0$ and $\tilde{p}_1$ represent the MLEs restricted to the null hypothesis. That is, $\tilde{p}_0 = \theta\tilde{p}_1$ if $\hat{p}_1 > T$ or $\tilde{p}_0 = \Delta + \tilde{p}_1$ if $\hat{p}_1 \leq T$.

3.1.1 Proposal

In this thesis, we are interested in defining some threshold (T) such that if $\hat{p}_1$ is greater than T, treatment groups are compared using θ, and if $\hat{p}_1$ is less than or equal to T, using Δ. In order to maintain a continuous NI margin, Δ_0 is constrained to equal $T(\theta_0 - 1)$. Conceptually, this means that we first define an appropriate relative NI margin and then decide for which

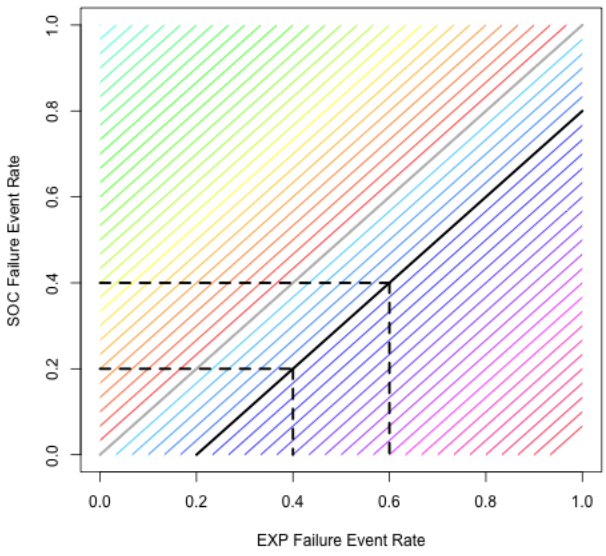
SOC failure rate we are comfortable with beginning to allow a slightly relaxed inferiority by switching to the absolute NI margin that corresponds to that rate.

3.2 Parameter Space Considerations

3.2.1 Difference of Proportions

Figure 3.1 shows all possible values of failure rates among subjects in the two treatment arms (where EXP would represent the placebo group for ITIP-1), with each contour representing an equal risk difference (RD). The thick black line represents an RD of 20% for $\Delta = p_0 - p_1$. As shown, $\Delta = 20\%$ might correspond to 40% and 60% failure rates in the SOC and EXP groups, respectively. It might also correspond to 20% and 40% failure rates in the two treatment arms.

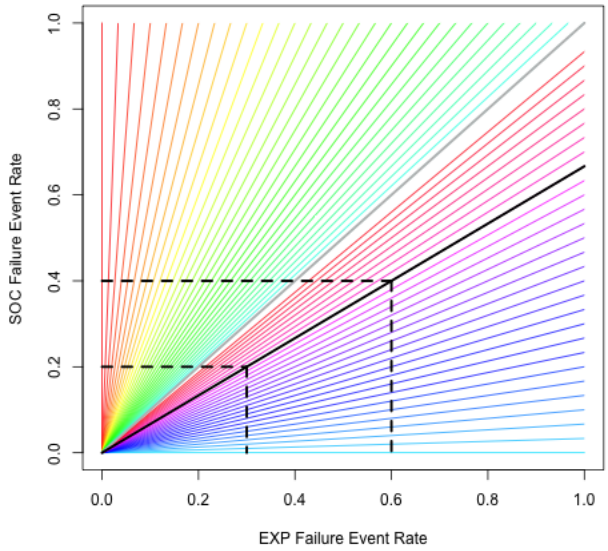
Figure 3.1: Risk Difference Parameter Space



3.2.2 Ratio of Proportions

Figure 3.2 again shows all possible values of failure rates among subjects in the two treatment groups, but this time with each contour representing an equal relative risk (RR). In this figure, the thick black line represents an RR of 1.5 for $\theta = p_0/p_1$. As shown, this factor of 1.5 might correspond to 40% and 60% failure rates in the SOC and EXP groups, respectively, or also to 20% and 30% failure rates.

Figure 3.2: Relative Risk Parameter Space



3.2.3 The Role of Nuisance Parameters

A nuisance parameter is any parameter that is not of immediate interest, but that must still be accounted for in the analysis of the target parameter of interest. As we have discussed, common targets of interest when assessing treatment effects in clinical trials include the difference or ratio of event rates. In the above figures, we illustrated that neither a 20% RD

nor a 1.5 RR discriminate between the two treatment arms' individual failure rates. The individual failure rates do, however, play an important role in determining the variance of these estimates, and are therefore considered nuisance parameters.

In order to account for nuisance parameters p_0 and p_1, focus can typically be directed at just one of the proportions, say p_1, since p_0 must then equal θp_1 or $\Delta + p_1$. In the previous chapter, we alluded to the mean-variance relationship inherent to binary data. Since the variance of the target parameters (RD or RR) will depend on the individual estimates, MLEs restricted to the null hypothesis must take into account all possible values of p_0 while maximizing the likelihood of p_1 in order to determine $\tilde{p}_0$ and $\tilde{p}_1$. While this task was previously considered computationally arduous and still to this day does not come standard in statistical software packages, computers are now able to overcome this rather easily.

The following sections discuss various asymptotic methods that each uniquely handle the issue of nuisance parameters. The Wald statistic bases its estimate of the target parameter's variance on the observed sample estimates and the score statistic on the restricted MLEs. The LR test statistic bases its null model on the restricted MLEs and alternative model on the observed sample estimates.

3.3 Wald-Based Inference

Proposed by Wald in 1943 [13], the Wald test is still the most widely used method for statistical inference today due to its computational simplicity and ease of interpretation.

3.3.1 Risk Difference

The Wald statistic estimates the standard error of the estimated difference in proportions using the unrestricted MLEs to obtain

$$\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}. \tag{3.9}$$

The standardized Wald test statistic corresponding to a difference in proportions is therefore

$$Z_{W_{RD}} = \frac{\hat{p}_0 - \hat{p}_1 - \Delta_0}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}}, \tag{3.10}$$

where Δ_0 represents the NI margin. The inversion of the Wald test to form a two-sided (1-α)% confidence interval for the true difference of proportions is a closed-form solution which takes the form

$$\hat{p}_0 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} \tag{3.11}$$

where α is defined as the type 1 error. When applying this method, a test that rejects the null hypothesis of $\Delta \geq \Delta_0$ in favor of NI when $Z_{W_{RD}} < z_{\alpha/2}$ is equivalent to a test that rejects if the upper bound of the (1-α)% excludes Δ_0.

3.3.2 Relative Risk

First proposed by Katz et al. in 1978 [14], relative risks are generally analyzed using a logarithmic transformation with the delta method applied to determine the asymptotic standard error of the estimated difference in log proportions (see Appendix A). Utilizing the observed sample estimates, the estimated standard error is therefore

$$\sqrt{\frac{1-\hat{p}_0}{\hat{p}_0 n_0} + \frac{1-\hat{p}_1}{\hat{p}_1 n_1}} \tag{3.12}$$

which leads to a Wald-like test statistic of

$$Z_{W_{RR}} = \frac{\log(\frac{\hat{p}_0}{\hat{p}_1}) - \log(\theta_0)}{\sqrt{\frac{1-\hat{p}_0}{\hat{p}_0 n_0} + \frac{1-\hat{p}_1}{\hat{p}_1 n_1}}}. \tag{3.13}$$

where θ_0 represents the NI margin. A (1-α)% confidence interval for $\log(\theta)$ can be calculated as

$$\log(\hat{p}_0) - \log(\hat{p}_1) \pm z_{\alpha/2}\sqrt{\frac{1-\hat{p}_0}{\hat{p}_0 n_0} + \frac{1-\hat{p}_1}{\hat{p}_1 n_1}} \tag{3.14}$$

which corresponds to a (1-α)% confidence interval for θ of

$$\left(\frac{\hat{p}_0}{\hat{p}_1}\exp\left(z_{\alpha/2}\sqrt{\frac{1-\hat{p}_0}{\hat{p}_0 n_0} + \frac{1-\hat{p}_1}{\hat{p}_1 n_1}}\right),\ \frac{\hat{p}_0}{\hat{p}_1}\exp\left(-z_{\alpha/2}\sqrt{\frac{1-\hat{p}_0}{\hat{p}_0 n_0} + \frac{1-\hat{p}_1}{\hat{p}_1 n_1}}\right)\right). \tag{3.15}$$

Note that since the confidence interval is symmetric about the estimate for log-relative risk, it will not be symmetric about the untransformed relative risk.

3.4 Score-Based Inference

In a score test, the standard error is calculated using restricted MLEs computed under the null hypothesis. That is, we estimate restricted MLEs such that $\tilde{p}_0$ - $\tilde{p}_1 = \Delta_0$ or $\tilde{p}_0/\tilde{p}_1 = \theta_0$.

3.4.1 Risk Difference

The score test statistic based on a risk difference was first discussed by Mee in 1984 and Miettinen and Nurminen in 1985 [15, 16]. Its confidence bounds were first proposed by Nurminen in 1986 and have been further discussed by many others, including Agresti in 2011 [17, 18].. The score test statistic is defined as

$$Z_{S_{RD}} = \frac{\hat{p}_0 - \hat{p}_1 - \Delta_0}{\sqrt{\frac{\tilde{p}_0(1-\tilde{p}_0)}{n_0} + \frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1}}} \tag{3.16}$$

where closed-form solutions exist for restricted MLEs $\tilde{p}_0$ and $\tilde{p}_1$, but an iterative process must be used when inverting $Z_{S_{RD}}$ to form a (1-α)% confidence interval (see Appendix B). The closed form solution to $\tilde{p}_0$ and $\tilde{p}_1$ is as follows:

$$\tilde{p}_1 = 2p\cos(a) - \frac{L_2}{3L_3} \quad \text{and} \quad \tilde{p}_0 = \tilde{p}_1 + \Delta_0 \tag{3.17}$$

where

$$a = \frac{1}{3}\left[\pi + \cos^{-1}\left(\frac{q}{p^3}\right)\right]$$
$$p = sign(q)\sqrt{\frac{L_2^2}{9L_3^2} - \frac{L_1}{3L_3}}$$
$$q = \frac{L_2^3}{27L_3^3} - \frac{L_1 L_2}{6L_3^2} + \frac{L_0}{2L_3}$$

and

$$L_3 = N$$
$$L_2 = (n_0 + 2n_1)\Delta_0 - N - X$$
$$L_1 = [n_1\Delta_0 - N - 2x_1]\Delta_0 + X$$
$$L_0 = x_1\Delta_0(1 - \Delta_0)$$

with $N = n_0 + n_1$ and $X = x_0 + x_1$. The alternative to this solution is an iterative procedure subject to the constraint $0 \leq \tilde{p}_1 + \Delta_0 \leq 1$.

3.4.2 Relative Risk

The score test statistic based on a relative risk was first discussed by Koopman in 1984 [19]. The test statistic is defined as

$$Z_{S_{RR}} = \frac{\hat{p}_0 - \theta_0\hat{p}_1}{\sqrt{\frac{\tilde{p}_0(1-\tilde{p}_0)}{n_0} + \frac{\theta_0^2\tilde{p}_1(1-\tilde{p}_1)}{n_1}}} \tag{3.18}$$

where, again, closed-form solutions exist for $\tilde{p}_0$ and $\tilde{p}_1$:

$$\tilde{p}_1 = \frac{-B - \sqrt{(B^2 - 4AC)}}{2A} \quad \text{and} \quad \tilde{p}_0 = \tilde{p}_1\theta_0 \tag{3.19}$$

where

$$\begin{aligned} A &= N\theta_0 \\ B &= -[n_0\theta_0 + X_0 + n_1 + X_1\theta_0] \\ C &= X \end{aligned}$$

and N and X are defined the same as before. Observe that this method eliminates both the need for a logarithmic transformation and a Delta Method approximation of the variance. Once again, an iterative process is used to invert $Z_{S_{RR}}$ to form a (1-α)% confidence interval (see Appendix B).

3.5 Likelihood Ratio-Based Inference

The likelihood ratio (LR) is defined as

$$\lambda_{LR} = \frac{L(\theta_0|x)}{L(\hat{\theta}|x)} \tag{3.20}$$

where x is the observed data and L corresponds to the likelihood function for the two independent binomial variables X_0 and X_1:

$$L = \binom{n_0}{x_0} p_0^{x_0}(1-p_0)^{n_0-x_0} \cdot \binom{n_1}{x_1} p_1^{x_1}(1-p_1)^{n_1-x_1}. \tag{3.21}$$

Applying Wilks Theorem, the LR test statistic takes the form

$$\begin{aligned}\chi^2_{LR} &= -2log(\lambda_{LR}) \\ &= -2log\left(\frac{\tilde{p}_1^{n_1\hat{p}_1}(1-\tilde{p}_1)^{n_1(1-\hat{p}_1)}\tilde{p}_0^{n_0\hat{p}_0}(1-\tilde{p}_0)^{n_0(1-\hat{p}_0)}}{\hat{p}_1^{n_1\hat{p}_1}(1-\hat{p}_1)^{n_1(1-\hat{p}_1)}\hat{p}_0^{n_0\hat{p}_0}(1-\hat{p}_0)^{n_0(1-\hat{p}_0)}}\right)\end{aligned} \tag{3.22}$$

where the numerator is restricted to the null hypothesis and the denominator to the observed data [20, 21]. According to the theorem, as the sample size becomes large, the test statistic χ^2_{LR} will have a χ^2 distribution with 1 degree of freedom under the null hypothesis. The LR z test statistic is defined as

$$Z_{LR} = sign(a)\sqrt{\chi^2_{LR}} \tag{3.23}$$

where

$$sign(a) = \begin{cases} +1, & \text{if } \hat{\Delta} > \Delta_0 \text{ or } \hat{\theta} > \theta_0. \\ -1, & \text{otherwise.} \end{cases} \tag{3.24}$$

The calculations for $\tilde{p}_0$ and $\tilde{p}_1$ will depend on whether or not treatment groups are compared using the RD or RR and have the same closed form solutions as given for the score statistic. We again use an iterative approach to invert the test statistic and form a (1-α)% confidence interval (see Appendix B).

3.6 Rejection Criteria

Using the NI margin approach, the conclusion of non-inferiority of an EXP relative to an SOC can be achieved through either formal hypothesis testing or confidence interval interpretation. The following sections describe the parameter spaces for the comparison of two binary proportions as they relate to non-inferiority and the two approaches that can be used for making inference.

3.6.1 Non-Inferiority Parameter Space

Figures 3.3 and 3.4 depict traditional risk difference and relative risk sample spaces for $\Delta_0 = 30\%$ and $\theta_0 = 2.0$, respectively. The thick black lines represent constant sample estimates $\hat{\Delta} = 18\%$ and $\hat{\theta} = 1.5$. The blue and red lines correspond to hypothetical confidence intervals for those estimates within the parameter spaces. The dark gray regions of the graphs represent the null hypothesis: inferiority of the EXP in relation to the SOC.

Figure 3.3: Example of Risk Difference Non-Inferiority

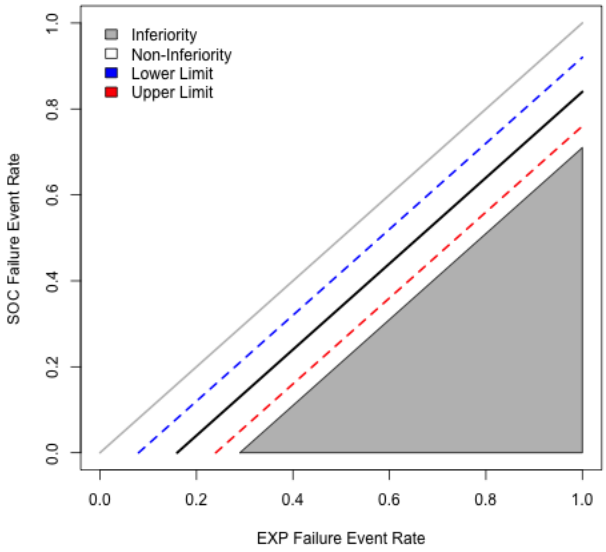
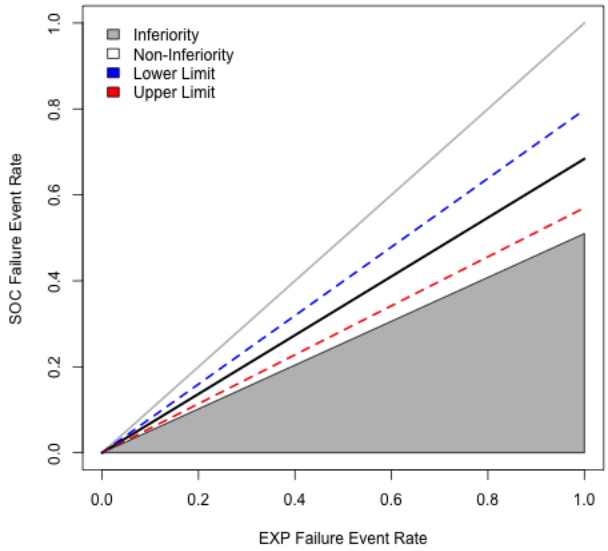
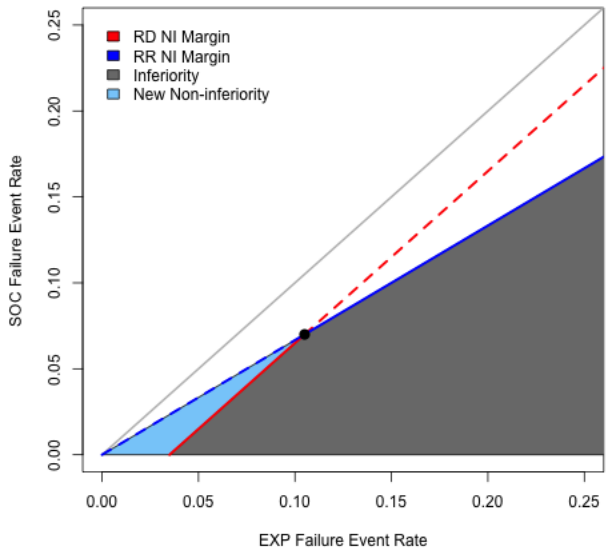


Figure 3.4: Example of Relative Risk Non-Inferiority



Emphasizing rare outcome rates, Figure 3.5 depicts inferiority under our proposed method for inference with $\theta_0 = 1.5$, $T = 7\%$ and $\Delta_0 = 3.5\%$. Under this proposal, the dark gray area represents the region of inferiority by both absolute and relative terms, while the blue area represents failure rates corresponding to the redefined definition of non-inferiority as determined by the switch from a relative to an absolute NI margin.

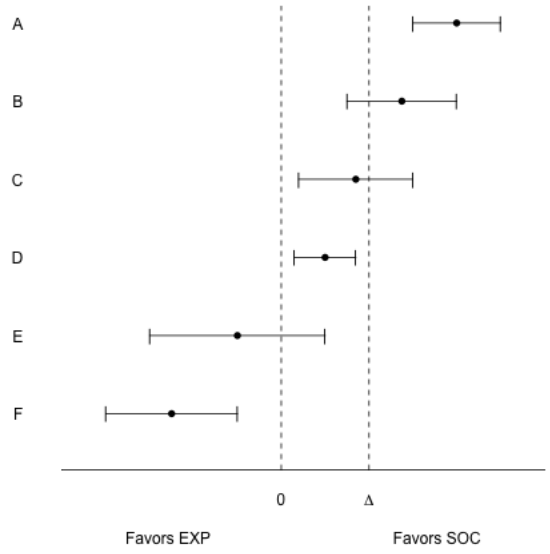
Figure 3.5: Example of Proposal for Non-Inferiority using 7% Threshold



3.6.2 Confidence Interval Approach

Using the confidence interval approach to make inference, NI is achieved if and only if the upper confidence limit excludes the NI margin. Figure 3.6 depicts various confidence interval interpretations for an NI trial given an observed estimate, its confidence interval, and a general NI margin Δ. Each of its scenarios can be thought of as having drawn a perpendicular line through the parameter estimates (thick black lines) from Figures 3.3 and 3.4. As shown, the confidence interval in scenario A is restricted to values greater than Δ and therefore is the only scenario for which we can conclude that the EXP is indeed inferior to the SOC. In scenario B, the observed estimate is suggestive of an inferior EXP, but its confidence interval supports both inferior and non-inferior treatment effects; we fail to conclude non-inferiority. Scenario C is suggestive of a non-inferior EXP, but as with scenario B its confidence interval fails to rule out either hypothesis; we fail to conclude non-inferiority.

Figure 3.6: Various Confidence Interval Interpretations



It is not until scenarios D through F that we obtain confidence intervals for which the upper limit excludes Δ in the correct direction for achieving NI, and we can therefore conclude with sufficient evidence that the EXP is indeed non-inferior to the SOC. More precisely, scenario D slightly favors the SOC since its confidence limits are above zero (and below Δ); scenario E is indicative of equivalent treatments since its confidence limits include zero; and scenario F favors the EXP since its confidence limits are below zero. Note that Scenario F is one that we do not believe to be true for the ITIP-1 trial since it compares placebo to SOC, but it could potentially represent the truth in a more typical NI trial.

3.6.3 Hypothesis Testing Approach

Using the formal hypothesis testing approach to make inference, the general test statistic, z_{test}, will be distributed as a standard normal variable under the null hypothesis. A value of

$z_{test} < z_{\alpha/2}$ is considered sufficient evidence against the null hypothesis at the one-sided $\alpha/2$ level. This research utilizes the hypothesis testing approach to determine whether or not to reject the null hypothesis and the 95% confidence interval approach to investigate coverage probability.

3.7 Study Design Considerations

3.7.1 Handling Extreme Outcomes

Table 3.1 lists the techniques used to determine confidence intervals when handling extreme failure rates under each of the three asymptotic inference methods. Notationally, X_0 and X_1 represent the number of failure events to have occurred in the EXP (i.e. placebo in the case of the ITIP-1 trial) and SOC treatment arms, respectively, and a dash is indicative of no adjustment to the data or confidence intervals.

Table 3.1: Techniques for Handling Extremely Rare Outcomes

Risk Difference							
		Wald		Score		Likelihood Ratio	
		Lower	Upper	Lower	Upper	Lower	Upper
$X_0 = 0$ *and* $X_1 = 0$		-1	1	-1	1	-1	1
$X_0 = 0$ *and* $X_1 \neq 0$		–	–	–	–	–	–
$X_0 \neq 0$ *and* $X_1 = 0$		–	–	–	–	–	–
Relative Risk							
		Wald		Score		Likelihood Ratio	
		Lower	Upper	Lower	Upper	Lower	Upper
$X_0 = 0$ *and* $X_1 = 0$		0	∞	0	∞	0	∞
$X_0 = 0$ *and* $X_1 \neq 0$		Sub $X_0 = \frac{1}{2}$*		–	–	–	–
$X_0 \neq 0$ *and* $X_1 = 0$		Sub $X_1 = \frac{1}{2}$*		–	–	–	–

*Subsitute 1/2 for the 0 value, and calculate CI accordingly.

3.7.2 Group Sequential Testing

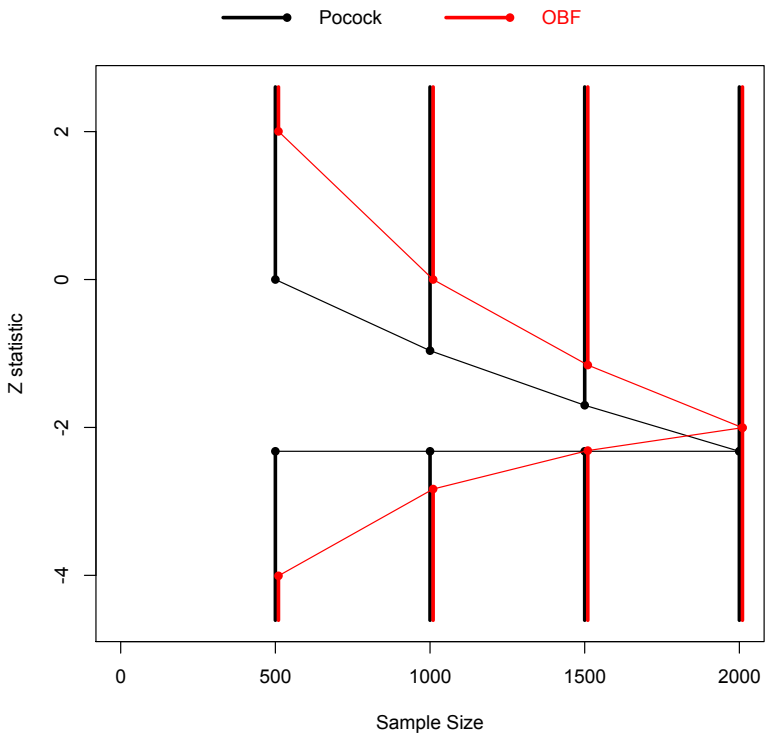
There are five well-documented reasons for terminating a clinical trial prior to achievement of maximum enrollment. These reasons are (1) the trial may already show sufficient evidence of an intervention's harmful effects; (2) the trial may already show sufficient evidence of an intervention's beneficial effects; (3) the trial may already indicate that a statistically significant result will probably not be achieved by the end of the study period; (4) severe data quality issues or markedly low participant recruitment rates which cannot be corrected will make it unlikely to achieve the targeted results; and (5) the question to be addressed has already been answered elsewhere or has lost its importance. Thus, for ethical reasons, the use of interim analyses in group sequential testing of data prior to the end of a clinical trial is widely accepted [22, 23, 24].

Since it can often be infeasible to perform an analysis after each subject's data has been collected, it is common to plan for a maximum number of analyses to be carried out at typically equivalent sample size increments until the termination of the study. Without adjusting the significance level to be used at each interim analysis, this use of repeated testing introduces a bias in that the probability of a significant test result by chance alone is greater than the selected overall significance level α. Therefore, statistical methods must be implemented to control type 1 error inflation. Two of the most common such designs are the OBrien-Fleming and the Pocock group sequential stopping boundaries. For both methods, a test statistic Z_i is calculated at analysis time i and compared to modified c_i critical values (i.e. the stopping boundaries). If Z_i falls outside the stopping boundaries, the trial should be terminated; if Z_i falls inside the stopping boundaries, the trial continues. The calculation of c_i depends on the group sequential stopping method chosen.

Figure 3.7 compares O'Brien-Fleming and Pocock stopping boundaries by the standardized normal statistic for up to four interim analyses using the RCTdesign statistical software package (see Appendix C for example R code). The vertical lines correspond to termination

at each of the 4 analysis times: if a test statistic occurs on the vertical lines the trial would terminate (the horizontal lines are displayed only to highlight boundary relationships). In a superiority trial, depending on the definition of the test statistic, the upper boundary might correspond to an intervention's futility and the lower boundary to its efficacy. In an NI trial, the upper boundary corresponds to an intervention's inferiority and the lower boundary to its non-inferiority. As shown by Figure 3.7, if a study reaches full enrollment (i.e. sample size of 2000), the OBrien-Fleming stopping boundary has more power to detect NI (i.e. its critical value is less than that of the Pocock boundary).

Figure 3.7: Sequential Monitoring of a Trial with O'Brien-Fleming and Pocock Boundaries



While both methods achieve the desired overall significance level, it is shown that the O'Brien Fleming boundary is more conservative than the Pocock boundary at earlier analyses. Due to the sensitivity of the ITIP-1 trial's intervention group in that it is a placebo being tested against an SOC, the simulations for this research make use of a hybrid design in which a Pocock boundary for early stopping for inferiority and an O'Brien Fleming boundary for early stopping for non-inferiority. Moreover, simulations were designed such that up to four analyses be carried out at equivalent sample size increments until the termination of the study, and at an overall one-sided significance level of 2.5%. Specifically for our proposed design, the determination of whether inference should be based on the absolute versus relative scale was based on the observed SOC failure rate at the most recent analysis time. Finally, it is important to acknowledge that the hypothesis testing performed in simulations accounts for the bias from multiple testing procedures by use of group sequential stopping boundaries. However, due to the more cumbersome implementation of group sequential methods applied to score- and LR-based CI, the corresponding confidence intervals are based on a fixed sample design.

Chapter 4

RESULTS

The following sections are organized according to the various scientific concerns of interest regarding our setting and proposal. Since we want to ensure that we do not approve a strategy that we would regard as truly inferior, we first make sure the type 1 error is constrained over the region of inferiority. It is sufficient to consider the behavior along the boundary of non-inferiority, with some attention paid to the sharp inflection point. We also want to examine how often we will accept a treatment in the perhaps controversial region that is now regarded as non-inferior. Furthermore, we want to examine how the power to adopt a non-inferior treatment might have improved more in the interior of the non-inferior region.

Table 4.1 outlines the principal scenarios used for assessing these concerns through fixed and sequential trial designs. Scenario A focuses on the interchange between type 1 error and power according to our relaxed inferiority assumptions, Scenario B on equal failure rates between the two groups in the region surrounding the threshold, and Scenario C on sample size considerations in order to detect outcomes within the relaxed inferiority region. Scenarios D through F investigate these same concepts under a smaller threshold (5% rather than 7%). Additional results corresponding to a 3% threshold for a fixed study design are available in Appendix D.

Table 4.1: Simulated Parameter Definitions

Part I: $\theta_0 = 1.5$, T=7% and Δ_0=3.5%							
Scenario	p_1	p_0	θ	Δ	Truth	n_1	n_0
A1	5.0%	6.5%	1.30	1.5%	θ=NI; Δ=NI	1000	1000
A2	5.0%	7.5%	1.50	2.5%	θ=I; Δ=NI	1000	1000
A3	5.0%	8.5%	1.70	3.5%	θ=I; Δ=I	1000	1000
A4	7.0%	10.5%	1.50	3.5%	θ=I; Δ=I	1000	1000
A5	9.0%	13.5%	1.50	4.5%	θ=I; Δ=I	1000	1000
A6	9.0%	12.5%	1.39	3.5%	θ=NI; Δ=I	1000	1000
B1	5.0%	5.0%	1.0	0.0%	θ=NI; Δ=NI	1000	1000
B2	7.0%	7.0%	1.0	0.0%	θ=NI; Δ=NI	1000	1000
B3	9.0%	9.0%	1.0	0.0%	θ=NI; Δ=NI	1000	1000
C1	5.0%	8.0%	1.6	3.0%	θ=I; Δ=NI	1000	1000
C2	5.0%	8.0%	1.6	3.0%	θ=I; Δ=NI	5000	5000
C3	5.0%	8.0%	1.6	3.0%	θ=I; Δ=NI	10000	10000
Part II: $\theta_0 = 1.5$, T=5% and Δ_0=2.5%							
Scenario	p_1	p_0	θ	Δ	Truth	n_1	n_0
D1	4.0%	5.0%	1.25	1.0%	θ=NI; Δ=NI	1000	1000
D2	4.0%	6.0%	1.50	2.0%	θ=I; Δ=NI	1000	1000
D3	4.0%	6.5%	1.63	2.5%	θ=I; Δ=I	1000	1000
D4	5.0%	7.5%	1.50	2.5%	θ=I; Δ=I	1000	1000
D5	6.0%	9.0%	1.50	3.0%	θ=I; Δ=I	1000	1000
D6	6.0%	8.5%	1.42	2.5%	θ=NI; Δ=I	1000	1000
E1	4.0%	4.0%	1.0	0.0%	θ=NI; Δ=NI	1000	1000
E2	5.0%	5.0%	1.0	0.0%	θ=NI; Δ=NI	1000	1000
E3	6.0%	6.0%	1.0	0.0%	θ=NI; Δ=NI	1000	1000
F1	4.0%	6.3%	1.58	2.3%	θ=I; Δ=NI	1000	1000
F2	4.0%	6.3%	1.58	2.3%	θ=I; Δ=NI	5000	5000
F3	4.0%	6.3%	1.58	2.3%	θ=I; Δ=NI	10000	10000

N simulations = 5000; NI=non-inferior; I=inferior

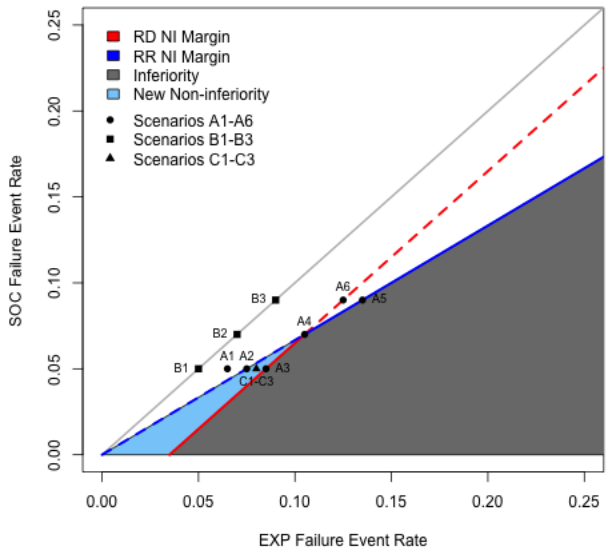
The organization of Table 4.1 is a reflection of similar trial outcomes that might be of interest among study investigators deciding whether or not to implement our proposed method for analysis. Since it may also be of interest to judge our proposal according to the statistical behaviors of highest concern within each of the various scenarios, the following list and references to Figure 4.1 (next page) are supplemental guides for navigating the results. We note that Figure 4.1 is a replicate of Figure 3.5 with the addition of Scenarios A through C superimposed on the plot.

- Control of type I error (i.e. outcomes along the solid blue and solid red lines): Scenarios A3-A5, D3-D5

- Power to detect original NI region (i.e. the region above the solid and dashed blue lines): Scenarios A1, A6, B1-B3, D1, D6, E1-E3

- Power to detect new NI region (i.e. the light blue region): Scenarios A2, C1-C3, D2, F1-F3

Figure 4.1: Scenarios A through C using 7% Threshold



4.1 Scenario A: Interchange Between Type 1 Error and Power

Rejection percentages and coverage probabilities for Scenario A are provided in Tables 4.2 and 4.3 on the following pages. Scenario A1 gives an example of the increase in power to detect an alternative that is NI by both the relative and absolute difference between p_0 and p_1. Given a 6.5% failure rate in the EXP group and a 5% failure rate in the SOC group, the power to detect non-inferiority increases from approximately 12% to approximately 47% under a fixed design and approximately 11% to 44% in the sequential trial design when comparing the relative risk margin to our proposal.

Scenario A2 is representative of a situation in which the relative difference between failure events is considered inferior, but that under our threshold proposal, the absolute difference in considered non-inferior. Moreover, the relative difference is equal to the NI margin and therefore its type 1 error is at its highest under the null hypothesis. The Risk Difference

column, in this scenario, can be interpreted as the ideal behavior of the proposed Threshold Design column since p_1 is less than T and inference should ideally be judged on absolute terms. From Scenario A2 in Tables 4.2 and 4.3, we observe that our proposal's relaxed inferiority assumptions increase the probability of achieving non-inferiority by approximately 13%.

Table 4.2: Scenario A Fixed Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
A1	12.5	12.7	12.3	48.6	46.5	47.5	48.6	46.5	47.5
A2	2.62	2.72	2.52	16.0	15.1	15.4	16.0	15.0	15.3
A3	0.42	0.42	0.40	2.50	2.34	2.42	2.46	2.30	2.36
A4	2.82	2.82	2.78	2.66	2.54	2.60	2.88	2.88	2.84
A5	2.48	2.48	2.48	0.44	0.38	0.42	2.48	2.48	2.48
A6	8.42	8.48	8.38	2.42	2.36	2.38	8.42	8.48	8.38
Coverage Probability (%)									
A1	95.6	95.4	95.4	95.4	95.4	95.3	95.3	95.3	95.2
A2	95.5	95.4	95.4	95.4	95.3	95.3	95.3	95.2	95.2
A3	95.8	95.2	95.4	95.3	95.3	95.3	95.2	95.1	95.1
A4	95.1	95.1	94.9	95.2	95.2	95.2	95.0	94.9	95.0
A5	95.4	95.3	95.1	95.3	95.3	95.3	95.6	95.5	95.4
A6	95.4	95.3	95.3	95.5	95.5	95.5	95.5	95.4	95.4

Scenarios A3 - A5 represent the situation in which p_0 is inferior by both relative and absolute differences, and therefore also for our hybrid threshold design. We expect type 1 error to be at its highest for the RD in Scenario A3, both RD and RR in Scenario A4, and RR in Scenario A5. As shown in the tables, type 1 error is approximately maintained across these scenarios, although it is slightly above the nominal 2.5% under our hybrid design. In

scenario A6, for which p_1 indicates inference should be kept on relative terms, we maintain an equal probability of rejection (approximately 8.5%) by both the Relative Risk and Threshold Design methods. These trends hold true for both a fixed and sequential trial design.

Table 4.3: Scenario A Sequential Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
A1	11.7	11.8	11.4	44.6	43.4	43.9	44.6	43.4	43.8
A2	2.60	2.64	2.58	15.0	14.3	14.6	15.0	14.2	14.6
A3	0.44	0.44	0.40	2.70	2.46	2.54	2.64	2.40	2.48
A4	2.84	2.86	2.78	2.76	2.56	2.68	3.00	3.02	2.92
A5	2.74	2.76	2.70	0.48	0.46	0.46	2.74	2.76	2.70
A6	8.34	8.44	8.14	2.78	2.66	2.70	8.34	8.44	8.14
Coverage Probability (%)									
A1	94.8	94.2	94.1	94.1	94.0	93.7	94.2	94.0	95.0
A2	95.7	95.3	95.1	93.7	93.8	93.6	93.8	93.9	95.6
A3	96.9	96.5	96.3	95.2	95.2	95.1	95.2	95.2	94.9
A4	95.6	95.4	95.2	95.3	95.2	95.1	95.1	94.9	95.0
A5	95.5	95.5	95.0	96.8	96.8	96.8	95.7	95.6	95.5
A6	94.4	94.2	94.0	95.3	95.3	95.3	94.5	94.3	95.7

In terms of 95% confidence interval coverage probability, we again see little difference between the behavior of the Wald, score, and LR statistics under the fixed and sequential trial designs for scenarios A1 - A6. For a fixed design, the coverage probabilties of our threshold method (between 94.9% - 95.6%) match well with the relative risk and risk difference confidence intervals (between 94.9% - 95.8% and 95.2% - 95.5%, respectively). Under the sequential design, the coverage probabilties of our threshold method (between 93.8% - 95.7%) again match well with the relative risk and risk difference confidence intervals (be-

tween 94.0% - 96.9% and 93.6% - 96.8%, respectively), but we do see more variability in this instance.

4.2 Scenario B: Equal Failure Rates Surrounding the Threshold

Scenario B represents a situation of no difference in failure rates (hence, non-inferiority) between treatment groups, for rates surrounding the threshold $T = 7\%$ and a sample size contrained to 2,000 subjects. Thus, scenario B especially pertains to the ITIP-1 trial settings in that if there truly is no difference between placebo and SOC treatment regimens, the power to detect this alternative can be drastically increased under the threshold design. As shown in Table 4.4, under a fixed trial there appears to be no significant difference between the Wald, score and LR methods in their ability to detect the equal failure rates between treatment groups nor in their corresponding coverage probabilities (which are slightly below 95%). Most notably, we again see that our proposal's relaxed inferiority assumptions increase the probability of achieving non-inferiority: by up to 40% when $p_1 < T$ (Scenario B1) and up to 15% when $p_1 = T$ (Scenario B2). When $p_1 > T$ (Scenario B3), the traditional relative comparison and our threshold proposal behave similarly, as desired.

Table 4.4: Scenario B Fixed Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
B1	54.3	54.5	53.9	94.5	93.7	94.1	94.5	93.7	94.1
B2	68.7	69.0	68.5	86.2	85.7	85.9	84.1	83.5	83.6
B3	80.7	80.9	80.7	78.2	77.6	77.9	80.8	81.0	80.7
Coverage Probability (%)									
B1	94.7	94.6	94.5	94.5	94.6	94.5	94.5	94.6	94.5
B2	94.9	94.7	94.7	94.7	94.7	94.7	94.8	94.7	94.7
B3	94.8	94.7	94.7	94.7	94.7	94.7	94.8	94.7	94.7

As shown in Table 4.5, we see similar statistical behavior in the sequential design as was seen in the fixed design. Slight decreases in rejection percentages (up to approximately 4%) and coverage probabilities (up to approximately 3%) exist between the sequential and fixed analyses, but the overall gain in power to detect non-inferiority remains similar (up to 40% in Scenario B1; 17% in Scenario B2). Nevertheless, the reduced coverage probabilities (as low as 91.1%) would affect the credibility of ones interpretation of the 95% confidence intervals.

Table 4.5: Scenario B Sequential Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
B1	49.3	49.7	48.9	91.4	90.9	91.1	91.4	90.9	91.1
B2	63.4	63.7	63.2	82.0	81.4	81.6	80.9	80.4	80.0
B3	76.2	76.3	76.0	73.9	73.5	73.8	77.0	77.1	76.3
Coverage Probability (%)									
B1	94.4	93.9	93.6	91.1	92.3	91.1	91.7	92.4	93.0
B2	94.0	93.6	93.5	92.5	92.8	92.4	93.0	92.9	94.9
B3	93.6	92.9	92.8	92.9	93.2	93.0	93.2	92.8	95.4

4.3 Scenario C: Detecting Outcomes in the Relaxed Inferiority Region

Scenario C is representative of a situation in which the true failure rates for the two treatment groups lie in a region that would be deemed inferior according to the relative difference, but non-inferior under our threshold proposal. It is important to acknowledge that if we were to detect an outcome in this narrow region of the parameter space for two binary outcomes, it is likely that the 95% confidence interval will not only exclude the NI margin, but also exclude zero (i.e. Scenario D in Figure 3.6). So while we may have sufficient statistical evidence of non-inferiority, we will also have sufficient evidence of some form of inferiority,

and study investigators will need to discuss the unique implications of such results. Each of the following three sub-scenarios correspond to increasing sample sizes of 1000, 5000 and 10,000 per treatment arm, respectively.

From Tables 4.6 and 4.7, we see the expected trends of decreased type 1 error (under the Relative Risk column) and increased power (under the Risk Difference and Threshold Design columns) as sample sizes increase.

Table 4.6: Scenario C Fixed Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
C1	0.86	0.90	0.86	7.04	6.44	6.68	6.94	6.36	6.60
C2	0.32	0.32	0.32	18.4	18.1	18.2	18.4	18.1	18.2
C3	0.08	0.08	0.08	29.4	29.1	29.2	29.4	29.1	29.2
Coverage Probability (%)									
C1	94.6	94.3	94.3	94.4	94.4	94.4	94.3	94.3	94.3
C2	94.9	94.9	95.0	94.7	94.7	94.7	94.7	94.7	94.7
C3	94.9	94.9	94.9	95.1	95.1	95.1	95.1	95.1	95.1

The results also indicate that while the probability of achieving non-inferiority under our threshold proposal compared to the traditional relative comparison does slightly increase for values in this region, large sample sizes are needed in order to obtain meaningful increases in the power to detect these differences. So, in the case of the ITIP-1 trial settings (Scenario C1), the threshold design yields an increase in power to detect non-inferiority of approximately 6%; if a much larger trial were to be conducted (Scenario C3), up to 28% increases in power could be obtained for this particular region of the parameter space. It is important to note that while these increases in power exist, the power to detect this alternative is still very low and therefore not particularly meaningful. As indicated by Scenario A1, the most meaningful gains in power will occur for outcomes deemed NI by the relative (and the absolute) margin, but analyzed on the absolute scale.

Table 4.7: Scenario C Sequential Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
C1	0.96	1.00	0.94	6.56	6.08	6.30	6.38	5.96	6.16
C2	0.42	0.42	0.42	16.9	16.6	16.7	16.9	16.6	16.7
C3	0.10	0.10	0.10	27.6	27.3	27.4	27.6	27.3	27.4
Coverage Probability (%)									
C1	96.7	96.4	96.2	94.6	94.4	94.2	94.7	94.5	95.4
C2	96.5	96.5	96.3	93.8	93.7	93.8	93.8	93.7	95.5
C3	95.9	95.9	95.8	92.9	92.8	92.8	92.9	92.8	94.5

Once again, 95% confidence interval coverage probability remains fairly consistent across all methods of inference, with more variability in the sequential design than the fixed design: coverage probability ranges between 94.3% and 95.1% for a fixed design and 92.8% and 96.7% for a sequential design.

4.4 Scenarios D - F

Scenarios D through F follow the same structure and interpretations as Scenarios A through C, with the major difference being that the threshold has been changed from 7% to 5%. As shown, the same general trends in type 1 error, power and coverage probability are upheld.

Rejection percentages and coverage probabilities for Scenario D, provided in Tables 4.8 and 4.9, indicate an increase from approximately 14% to 35% in power to detect non-inferiority after relaxing inferiority assumptions when the true failure rates are non-inferior by both relative and absolute terms (Scenario D1). When the true failure rates are only non-inferior according to their absolute difference, we see an increase from approximately 2.5% to approximately 6.5%-7% in power to detect the NI alternative (Scenario D2). Recalling from before that we expect type 1 error to be highest at approximately 2.5% for the RD in Scenario D3, RD and RR in Scenario D4, and RR in Scenario D5, Scenarios D3 - D5 indicate an approximately maintained type 1 error under the Wald, score, and LR methods. Scenario D6 indicates a maintained power (approximately 5.5% for a fixed design and 10% for a sequential design) between the Relative Risk and Threshold Design columns when failure rates indicate to make relative comparisons.

Table 4.8: Scenario D Fixed Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
D1	14.8	14.9	14.6	38.2	36.4	37.1	37.4	35.7	36.3
D2	2.84	2.84	2.74	7.32	7.02	7.22	6.66	6.42	6.52
D3	0.96	0.96	0.92	2.80	2.50	2.64	2.20	2.04	2.08
D4	2.62	2.72	2.52	2.54	2.44	2.50	2.68	2.78	2.60
D5	2.68	2.68	2.60	0.96	0.88	0.90	2.68	2.68	2.60
D6	5.50	5.56	5.30	2.78	2.56	2.66	5.50	5.56	5.30
Coverage Probability (%)									
D1	95.0	94.8	94.6	94.7	94.6	94.6	94.5	94.2	94.3
D2	94.9	94.8	94.7	94.9	95.0	94.9	94.3	94.3	94.4
D3	94.8	94.7	94.9	95.0	95.2	95.0	94.4	94.4	94.6
D4	95.5	95.4	95.4	95.4	95.3	95.3	95.4	95.1	95.3
D5	95.2	95.1	95.1	95.1	95.1	95.1	95.6	95.5	95.6
D6	95.3	95.2	95.1	94.8	94.9	94.9	95.5	95.4	95.4

In accordance with the previous results, the 95% confidence interval coverage probability of our threshold method matches well with the relative risk and risk difference confidence intervals for a fixed design (between 94.3% and 95.6% overall), but we again see more variability for a sequential design (between 92.0% and 96.9% overall). Moreover, there seems to be a slight trend of decreased coverage probability in sequential designs as we move from the relative risk intervals to the threshold design intervals (up to 2.4%).

Table 4.9: Scenario D Sequential Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
D1	14.0	14.2	13.8	35.4	33.7	34.6	34.7	33.0	33.7
D2	2.62	2.64	2.58	8.08	7.36	7.70	7.40	6.82	7.00
D3	1.04	1.12	1.02	3.02	2.68	2.88	2.60	2.36	2.50
D4	2.60	2.64	2.58	2.80	2.54	2.62	2.70	2.74	2.66
D5	2.84	2.86	2.78	0.46	0.46	0.46	2.84	2.86	2.78
D6	9.98	10.1	9.82	2.66	2.58	2.62	10.0	10.1	9.82
Coverage Probability (%)									
D1	95.1	94.3	93.7	93.5	93.6	93.5	93.4	93.1	95.3
D2	95.8	95.6	95.1	93.3	93.3	93.1	93.4	93.2	94.4
D3	96.9	96.3	95.9	95.0	94.9	94.7	94.9	94.7	94.1
D4	95.7	95.3	95.1	95.2	95.1	95.1	95.4	95.0	95.2
D5	95.6	95.4	95.2	96.0	95.8	95.7	94.9	94.8	94.8
D6	92.5	92.2	92.3	95.3	95.3	95.2	92.2	92.0	94.2

Scenario E again pertains to the ITIP-1 trial settings in that if there truly is no difference between treatment regimens, the power to detect this alternative is increased under the threshold design. As shown for a fixed design in Table 4.10, the probability of achieving non-inferiority increases by up to 30% when $p_1 < T$ (Scenario E1) and up to 13% when $p_1 = T$ (Scenario E2). When $p_1 > T$ (Scenario E3), the traditional relative comparison and our threshold proposal behave similarly. For the sequential design in Table 4.11, the probability of achieving non-inferiority increases by up to 36% when $p_1 < T$ (Scenario E1) and up to 17% when $p_1 = T$ (Scenario E2). Again, when $p_1 > T$ (Scenario E3), the traditional relative comparison and our threshold proposal behave similarly.

Table 4.10: Scenario E Fixed Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
E1	59.8	60.7	59.6	89.0	88.4	88.7	89.0	88.3	88.7
E2	68.1	68.3	67.6	83.4	82.5	82.9	81.3	80.7	80.9
E3	74.7	74.9	74.7	77.3	76.6	76.9	76.4	76.4	63.0
Coverage Probability (%)									
E1	95.0	94.8	94.7	94.7	94.8	94.7	94.8	94.8	94.7
E2	94.7	94.6	94.5	94.5	94.6	94.5	94.6	94.6	94.5
E3	95.0	94.9	94.9	94.9	94.9	94.9	95.0	94.9	94.9

Table 4.11: Scenario E Sequential Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
E1	41.0	41.5	40.8	76.9	75.9	76.4	76.8	75.8	76.2
E2	49.3	49.7	48.9	68.1	66.7	67.4	65.6	64.7	64.3
E3	57.0	57.2	56.8	60.3	59.3	59.6	59.3	59.3	58.1
Coverage Probability (%)									
E1	94.6	93.8	93.5	92.3	92.8	92.4	92.8	92.8	94.7
E2	94.4	93.9	93.6	93.1	93.4	92.9	93.3	93.3	95.0
E3	94.2	93.4	93.2	93.1	93.3	92.9	93.4	93.1	95.0

Scenario F investigates the situation in which the two failure rates lie in the new non-inferior region under our threshold proposal. As indicated by Tables 4.12 and 4.13, we again see that very large samples are necessary in order to obtain significant power in this region. As was the case with Scenario C, we again warn of the same potential interpretative issues that may arise from detecting outcomes in this region.

Table 4.12: Scenario F Fixed Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
F1	1.14	1.18	1.10	3.92	3.62	3.82	3.00	2.86	2.96
F2	0.44	0.44	0.44	6.46	6.26	6.38	6.46	6.26	6.38
F3	0.28	0.28	0.28	9.34	9.18	9.26	9.34	9.18	9.26
Coverage Probability (%)									
F1	94.9	94.8	94.8	95.1	95.1	95.1	94.4	94.4	94.4
F2	95.3	95.2	95.3	95.3	95.3	95.3	95.3	95.3	95.3
F3	94.7	94.7	94.6	95.2	95.1	95.2	95.2	95.1	95.2

Table 4.13: Scenario F Sequential Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
F1	1.46	1.46	1.44	4.12	3.70	3.84	3.48	3.16	3.26
F2	0.44	0.46	0.44	6.42	6.24	6.38	6.40	6.24	6.38
F3	0.38	0.38	0.38	8.64	8.54	8.58	8.64	8.54	8.58
Coverage Probability (%)									
F1	96.3	95.6	95.2	94.5	94.3	94.3	94.5	94.1	94.5
F2	96.8	96.7	96.6	94.2	94.1	94.2	94.2	94.1	95.4
F3	96.3	96.3	96.2	94.4	94.4	94.4	94.5	94.4	95.7

4.5 Coverage Probability for Additional Equal Failure Rate Scenarios

The following graphs compare coverage probability for equal failure event rates in the two treatment groups, with values ranging from 4% to 20% in 0.5% increments and using 1000 simulations per failure rate.

4.5.1 Fixed Analysis

Under a fixed trial design, there is no distinguishable difference in coverage probability between the Wald, score, and LR methods when comparing relative risk or risk difference intervals with those of our threshold proposal. All three methods produce an approximate confidence interval coverage surrounding the nominal 95% (anywhere between 93.9% and 96.5%). As was the case for Scenarios B and E, coverage probability has a tendency to be slightly less than the nominal 95% when the observed equivalent failure rates are below the threshold.

Figure 4.2: Fixed Trial RD Coverage Probability for Equal Failure Rates

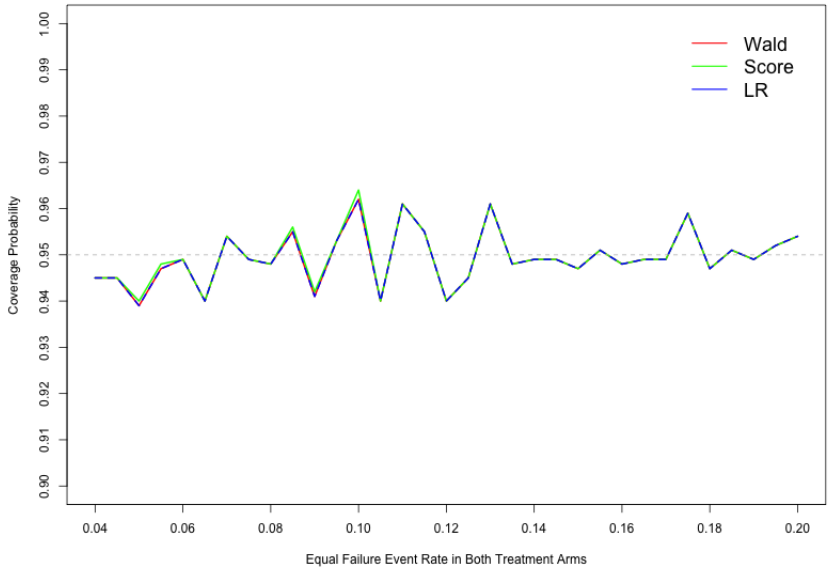


Figure 4.4: Fixed Trial Proposal's Coverage Probability for Equal Failure Rates

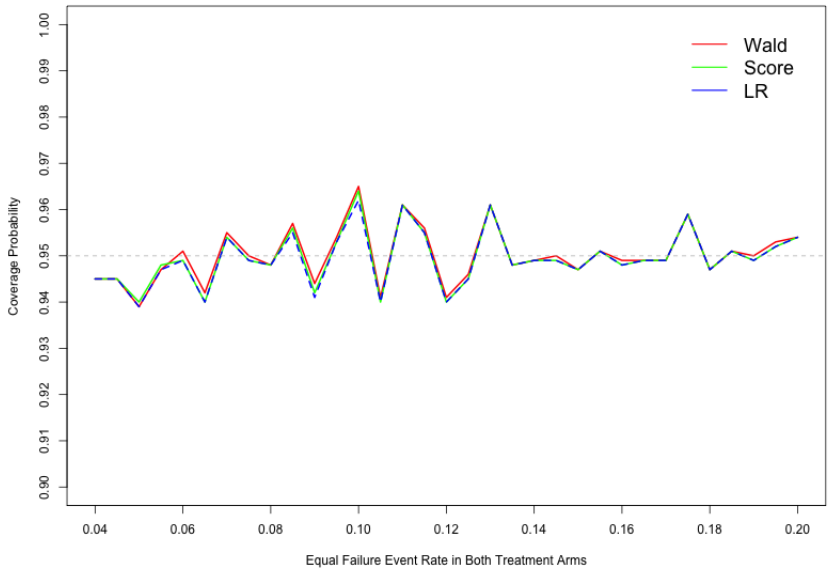
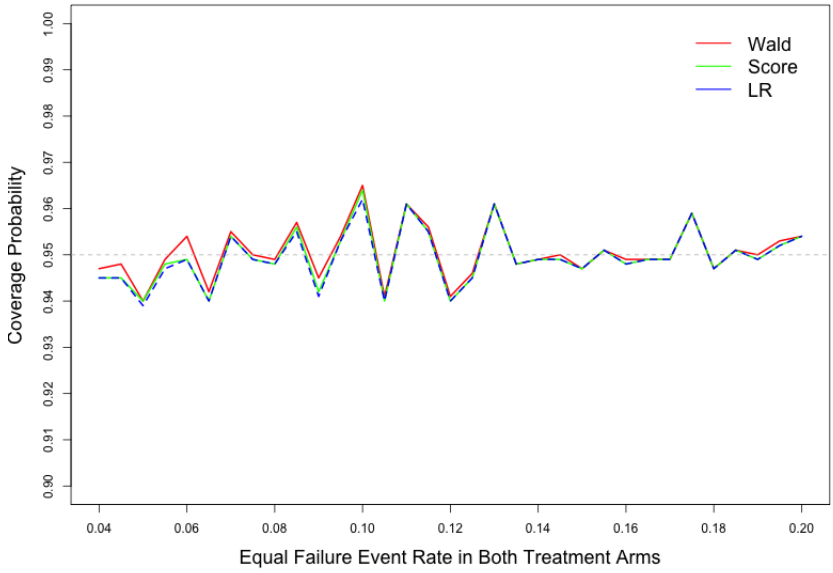


Figure 4.3: Fixed Trial RR Coverage Probability for Equal Failure Rates



4.5.2 Sequential Analyses

Under a sequential trial design with the same equal failure rates and threshold, we again see the overall trend of greater variance in confidence interval coverage in comparison to a fixed trial design, with sequential coverage ranging anywhere between 90.2% and 96.0%. More precisely, relative risk confidence interval coverage ranges between 90.2% and 95.0%, risk difference interval coverage between 90.5% and 95.1%, and threshold design coverage between 90.3% and 96.0%. As shown in Figures 4.5 and 4.6, confidence interval coverage is closest to the nominal 95% when relative risk intervals are applied to rare failure rates (less than 8%), but risk difference intervals perform better as failure rates increase. As shown in Figure 4.7, the threshold design's confidence intervals have a tendency to perform somewhere in between these differences, and the LR-based intervals outperform the Wald and score-based intervals for nearly all failure rates.

Figure 4.5: Sequential Trial RD Coverage Probability for Equal Failure Rates

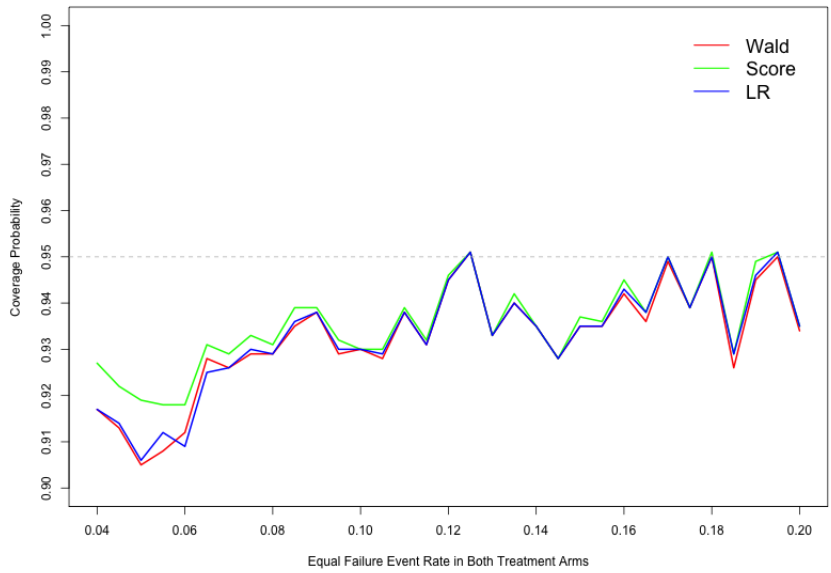


Figure 4.6: Sequential Trial RR Coverage Probability for Equal Failure Rates

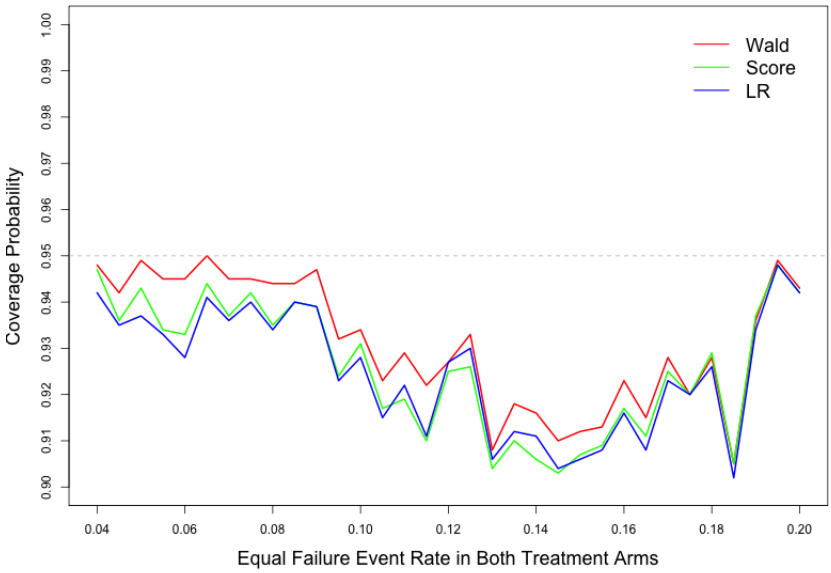
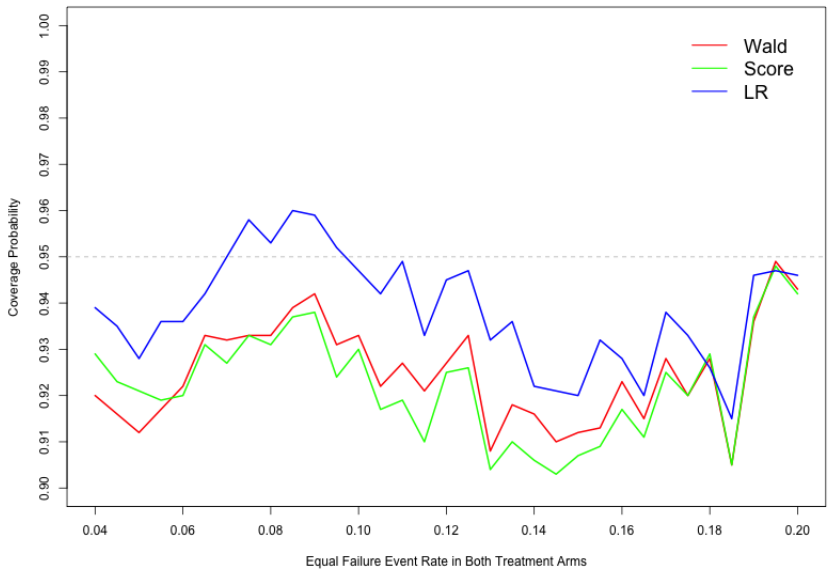


Figure 4.7: Sequential Trial Proposal's Coverage Probability for Equal Failure Rates



Chapter 5

CONCLUSIONS

In this research we describe Wald, score and LR-based NI inference using relative and absolute differences to compare rare event rates. We propose a method for making inference in which the choice of a relative or absolute comparison of event rates is based on the observed SOC event rate. In our proposal, inferiority assumptions are relaxed for extremely rare SOC outcomes by switching from a relative to a more conservative absolute comparison when SOC rates are below a pre-defined threshold T. When switching scales, the absolute NI margin is parametrized according to the threshold and relative NI margin.

We first emphasize that the proposed framework only be applied under an appropriate trial setting in which the sample population adequately reflects the underlying target population of interest to ensure the risk difference (and similarly the number needed to treat) is a meaningful measure of treatment effect. We further emphasize that due diligence be given in defining the relative NI margin and the threshold by which failure event rates are considered rare enough in the SOC group to allow slightly higher rates in the EXP group. That is, the threshold design must maintain the original goal in which the relative NI margin is defined such that the effectiveness of the EXP will be proven, and after all clinical information is considered, the use of the EXP over the SOC will not result in a meaningful loss of efficacy. Once this has been defined, the threshold must be chosen such that the relaxed restrictions on inferiority lead to a more meaningful comparison of treatment groups and to a potentially greater benefit beyond the scope of the trial, such as resource allocation or lower resistance to antibiotics.

Our results indicate little difference between the type 1 error, power, and coverage prob-

ability of the Wald-, score- and LR-based inference methods under a fixed or sequential trial design. Our work emphasizes that relaxed inferiority assumptions for rare outcome rates can lead to increased power to detect a more meaningful NI without significant increase in type 1 error or extreme loss in coverage probability (although coverage probabilities based on a sequential design did have more variability than those based on a fixed design and tended to be slightly lower than nominally expected for some settings). In the setting of the ITIP-1 trial, we emphasize the higher power to detect non-inferiority if there truly is no difference in failure rates between subjects on placebo versus those on the SOC (up to 40% for a 7% threshold in Scenario B1 and up to 30% for a 5% threshold in Scenario E1). We further emphasize that the most meaningful gains in power will be under alternatives which are NI by both the relative and absolute margins, but which are analyzed according to the absolute difference (Scenarios A1 and D1).

We look exclusively at a fixed trial design with a 3% threshold in Appendix D(avoiding sequential designs due to limitations within the RCTdesign software). To highlight Scenario H1, the probability of achieving non-inferiority increases by up to 75% when failure rates are 1% in both treatment groups. In fact, as the equal failure rates decrease from the threshold of 3% down to 1%, the probability of achieving non-inferiority also increases. Thus, we see further evidence suggesting that our hybrid threshold design produces gains in power to detect a meaningful non-inferiority without substantial loss of control of type 1 error or coverage probability.

Limitations of this research include the scope of scenarios simulated and presented, as many more exist that might be of interest to study investigators. Within the presented scenarios, one particular limitation to this research is that determination of confidence interval coverage does not adjust for biased estimates due to multiple testing procedures within a sequential trial design (although the determination of whether or not to stop the trial early via the standardized z-statistic does). This may or may not explain the increased variability

for confidence interval coverage under a sequential sampling scheme. Additionally, while our focus on extremely rare outcome rates is most relevant for the ITIP-1 trial, it is also believed that similar statistical behavior and gains in meaningful comparisons might be made in the cases of either relaxing inferiority assumptions for extremely frequent *positive* outcome rates by switching from an absolute to a less conservative relative NI margin, or relaxing inferiority assumptions for extremely frequent *negative* outcome rates by switching from a relative to a more conservative absolute NI margin. Future work relating to this research might expand on these limitations or consider alternative proposals for relaxing the definition of non-inferiority under extreme outcome conditions in order to produce the most appropriate assessment of an intervention's impact on public health.

BIBLIOGRAPHY

[1] Food, D. Administration, *et al.*, "Guidance for industry non-inferiority clinical trials," *Rockville (Maryland): FDA*, 2010.

[2] T. R. Fleming, "Current issues in non-inferiority trials," *Statistics in Medicine*, vol. 27, no. 3, pp. 317–332, 2008.

[3] T. R. Fleming, "Design and interpretation of equivalence trials," *American Heart Journal*, vol. 139, no. 4, pp. S171–S176, 2000.

[4] G. Piaggio, D. R. Elbourne, S. J. Pocock, S. J. Evans, D. G. Altman, C. Group, *et al.*, "Reporting of noninferiority and equivalence randomized trials: extension of the consort 2010 statement," *JAMA*, vol. 308, no. 24, pp. 2594–2604, 2012.

[5] T. R. Fleming, K. Odem-Davis, M. D. Rothmann, and Y. L. Shen, "Some essential considerations in the design and conduct of non-inferiority trials," *Clinical Trials*, vol. 8, no. 4, pp. 432–439, 2011.

[6] M. D. Rothmann, B. L. Wiens, and I. S. Chan, *Design and analysis of non-inferiority trials*. CRC Press, 2011.

[7] R. Temple and S. S. Ellenberg, "Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: ethical and scientific issues," *Annals of Internal Medicine*, vol. 133, no. 6, pp. 455–463, 2000.

[8] T. M. de Boo and G. A. Zielhuis, "Minimization of sample size when comparing two small probabilities in a non-inferiority safety trial," *Statistics in Medicine*, vol. 23, no. 11, pp. 1683–1699, 2004.

[9] A. P. S. M. N. L. Amy Ginsburg, Salim Sadruddin, "Innovative treatments in pneumonia," 2014.

[10] D. Shah, "3-day or 5-day oral antibiotics for non-severe pneumonia in children.," *Indian Pediatrics*, vol. 45, no. 7, pp. 577–578, 2008.

[11] P. C. O'Brien and T. R. Fleming, "A multiple testing procedure for clinical trials," *Biometrics*, pp. 549–556, 1979.

[12] S. J. Pocock, "Interim analyses for randomized clinical trials: the group sequential approach," *Biometrics*, pp. 153–162, 1982.

[13] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, vol. 54, no. 3, pp. 426–482, 1943.

[14] D. Katz, J. Baptista, S. Azen, and M. Pike, "Obtaining confidence intervals for the risk ratio in cohort studies," *Biometrics*, pp. 469–474, 1978.

[15] R. W. Mee and D. Anbar, "Confidence bounds for the difference between two probabilities," 1984.

[16] O. Miettinen and M. Nurminen, "Comparative analysis of two rates," *Stat Med*, vol. 4, no. 2, pp. 213–226, 1985.

[17] M. Nurminen, "Confidence intervals for the ratio and difference of two binomial proportions," *Biometrics*, vol. 42, no. 3, pp. pp. 675–676, 1986.

[18] A. Agresti, "Score and pseudo-score confidence intervals for categorical data analysis," *Statistics in Biopharmaceutical Research*, vol. 3, no. 2, pp. 163–172, 2011.

[19] P. A. R. Koopman, "Confidence intervals for the ratio of two binomial proportions," *Biometrics*, vol. 40, no. 2, pp. pp. 513–517, 1984.

[20] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[21] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. CRC Press, 1979.

[22] L. M. Friedman, C. Furberg, D. L. DeMets, *et al.*, *Fundamentals of Clinical Trials*. Springer, 4 ed., 2010.

[23] S. S. Emerson, J. M. Kittelson, and D. L. Gillen, "Frequentist evaluation of group sequential clinical trial designs," *Statistics in Medicine*, vol. 26, no. 28, pp. 5047–5080, 2007.

[24] C. Jennison and B. W. Turnbull, *Group sequential methods with applications to clinical trials*. CRC Press, 1999.

Appendix A

DELTA METHOD

A.1 Univariate Definition

Suppose an estimator $\hat{\theta}$ has an asymptotic distribution

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \sigma_\theta^2).$$

Then providing function f is differentiable at θ, $f(\hat{\theta})$ will have an asymptotic distribution

$$\sqrt{n}(f(\hat{\theta}) - f(\theta)) \xrightarrow{D} N\left(0, \left(\frac{df(\theta)}{d\theta}\right)^2 \cdot \sigma_\theta^2\right).$$

Thus, we approximate the distribution of $\hat{\theta}$ as

$$\hat{\theta} \dot{\sim} N(\theta, \left(\frac{df(\theta)}{d\theta}\right)^2 \cdot \sigma_\theta^2)$$

A.2 Application to Relative Risk

We define the binomially distributed random variable $X_j \sim B(n_j, p_j)$ as the number of subjects in treatment arm j to have an event, n_j as the total number of subjects, and p_j as the true unknown probability of an event. We apply asymptotic theory to $\hat{p_j} = \frac{X_j}{n_j}$ to obtain

$$\sqrt{n}[\hat{p_j} - p_j] \xrightarrow{D} N(0, p_j(1 - p_j)).$$

Owing to the general tendency for ratios to be less well behaved in small samples, when a relative risk is the target of inference it is common to apply a logarithmic transformation.

Using the delta method with asymptotic theory, an approximate distribution for $\log(\hat{p_j})$ is:

$$\log(\hat{p_j}) \dot{\sim} N(\log(p_j), \left(\frac{1}{p_j}\right)^2 \cdot \frac{p_j}{(1-pj)}) \\ \dot{\sim} N(\log(p_j), \frac{1-p_j}{p_j n_j}).$$

Assuming the two treatment arms are independent, p_0 and p_1 are statistically independent random variables and the variance of their difference is equal to the sum of the individual variances. An approximate distribution of the estimated log-relative risk comparing p_0 to p_1 based on asymptotic theory is thus

$$\log(RR) = \log(\hat{p}_1) - \log(\hat{p}_0) \dot{\sim} N(\log(p_1) - \log(p_0), \frac{1-p_1}{p_1 n_1} + \frac{1-p_0}{p_0 n_0})$$

Appendix B

INVERSION OF TEST STATISTIC TO FORM 95% CONFIDENCE INTERVALS

B.1 Guidelines

The following flow charts describe the iterative process used to determine the upper limit of a 95% confidence interval using the inversion of a test statistic approach. As described, the final value of Pre-Δ in B.1 and Pre-θ in B.2 are defined as the upper bounds for Δ and θ, respectively. The lower limits are determined using a similar process, but by subtracting rather than adding increments to the inital or previous Pre-Δ and Pre-θ values. This iterative process was used to determine the 95% confidence bounds based on the score and likelihood ratio statistics, wherein the only difference lies in the calculation of the test statistic.

Figure B.1: Guide for Upper Confidence Limit for Δ

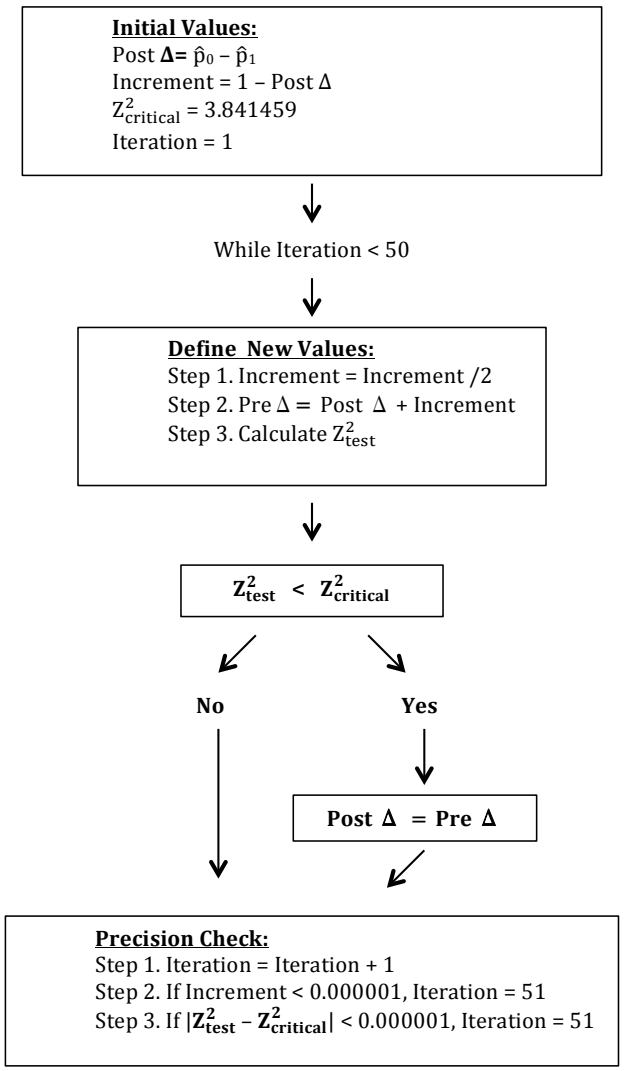
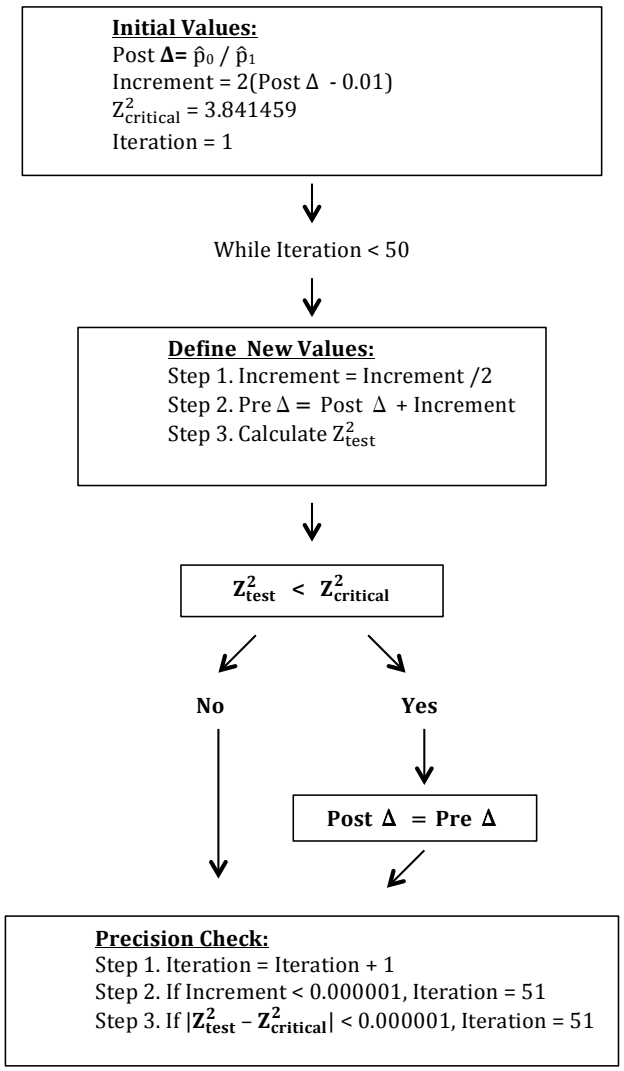


Figure B.2: Guide for Upper Confidence Limit for θ



B.2 Example R Code

```
scoreRD <- function(x0, x1, n0, n1, delta0, conf.level=0.95)
{

    if(x0 == 0 && x1 == 0){
      estimate <- "NA"
      CIlo <- -Inf
      CIhi <- Inf
      Z <- "NA"
    }

    else{
      p0hat <- x0/n0
      p1hat <- x1/n1
      z2_crit <- qchisq(conf.level,1)

      deltahat <-  p0hat - p1hat
      inc <- 1 - deltahat
      iter  <- 1

      while(iter <= 50){
        inc <- inc/2
        shiftup <- deltahat + inc
        z2_test <- scoreTestRD(p0hat, p1hat, n0, n1, shiftup)

        if(z2_test < z2_crit){
          deltahat <- shiftup
        }

        iter <- iter + 1

        if( (inc < 0.000001) || ( abs(z2_test - z2_crit) < 0.000001) ){
          iter <- 51
          CIhi <- shiftup
        }
      }

     deltahat <-  p0hat - p1hat
     inc <- 1 + deltahat
```

```
     iter  <- 1

     while(iter <= 50){
        inc <- inc/2
        shiftdown <- deltahat - inc
        z2_test <- scoreTestRD(p0hat, p1hat, n0, n1, shiftdown)

        if(z2_test < z2_crit){
          deltahat <-  shiftdown
        }

        iter <- iter + 1

        if( (inc < 0.000001) || ( abs(z2_test - z2_crit) < 0.000001) ){
          iter <- 51
          CIlo <- shiftdown
        }
      }
    }

    result <- data.frame(S_RDlo = CIlo, S_RDhi = CIhi)
    return(result)
}

scoreTestRD <- function(p0hat, p1hat, n0, n1, shift)
{
    num <- p0hat - p1hat - shift

    if( abs(num) == 0 ){
      z2_test <- 0
    }
    else{
      t = n1/n0
      a = 1 + t
      b = -( 1 + t + p0hat + t * p1hat + shift * (t + 2) )
      c = shift * shift + shift * (2 * p0hat + t + 1) + p0hat + t * p1hat
      d = -p0hat * shift * (1 + shift)
      v = (b / a / 3)^3 - b * c / (6 * a * a) + d / a / 2
      s = sqrt( (b / a / 3)^2 - c / a / 3 )
```

```
      if(v>0){u = s} else{u = -s}

      w = ( pi + acos(v / u^3) ) / 3
      p0mle = 2 * u * cos(w) - b / a / 3
      p1mle = p0mle - shift

      var <- p0mle * (1 - p0mle) / n0 + p1mle * (1 - p1mle) / n1
      z2_test <- num^2 / var
    }

    return(z2_test)
}

lrRR <- function(x0, x1, n0, n1, theta0, conf.level=0.95)
{

    if(x0 == 0 && x1 == 0){
      estimate <- "NA"
      CIlo <- -Inf
      CIhi <- Inf
      Z <- "NA"
      p.value <- "NA"
      reject <- "false"
    }

    else{
      p0hat <- x0/n0
      p1hat <- x1/n1
      z2_crit <- qchisq(conf.level,1)

      thetahat <-  p0hat/p1hat
      inc <- (thetahat-0.01)*2
      iter  <- 1

      while(iter <= 50){
        inc <- inc/2
        shiftup <- thetahat + inc
        z2_test <- lrTestRR(p0hat, p1hat, n0, n1, shiftup)

        if(iter == 1 & z2_test < z2_crit){z2_test = z2_crit + 0.01}
```

```
      if(z2_test < z2_crit){
        thetahat <- shiftup
      }

      iter <- iter + 1

      if( (inc < 0.000001) || ( abs(z2_test - z2_crit) < 0.000001) ){
        iter <- 51
        CIhi <- shiftup
      }
    }

    thetahat <-  p0hat/p1hat
    inc <- (thetahat-0.01)*2
    iter  <- 1

    while(iter <= 50){
      inc <- inc/2
      shiftdown <- thetahat - inc
      z2_test <- lrTestRR(p0hat, p1hat, n0, n1, shiftdown)

      if(iter == 1 & z2_test < z2_crit){z2_test = z2_crit + 0.01}

      if(z2_test < z2_crit){
        thetahat <-  shiftdown
      }

      iter <- iter + 1

      if( (inc < 0.000001) || ( abs(z2_test - z2_crit) < 0.000001) ){
        iter <- 51
        CIlo <- shiftdown
      }
    }
  }

  result <- data.frame(LR_RRlo = CIlo, LR_RDhi = CIhi)
  return(result)
}
```

```
lrTestRR <- function(p0hat, p1hat, n0, n1, shift)
{

    x0 = p0hat * n0
    x1 = p1hat * n1

    A = (n0+n1)*shift
    B = -(n0*shift + x0 + n1 + x1*shift)
    C = x0+x1
    p1mle = ( -B - sqrt(B^2-4*A*C))/(2*A)
    p0mle = p1mle*shift

    if(p1mle == 0){ p1mle = 0.0000000001}
    if(p0mle == 0){ p0mle = 0.0000000001}

    logL_R <- x0 * log(p0mle) + (n0 - x0) * log(1 - p0mle) + x1 * log(p1mle) + (n1 - x1) * log(1 - p1mle)
    logL_Un <- x0 * log(p0hat) + (n0 - x0) * log(1 - p0hat) + x1 * log(p1hat) + (n1 - x1) * log(1 - p1hat)
    z2_test <-  2 * (logL_Un - logL_R) # Wilk's Theorem

    return(z2_test)
}
```

B.3 Example Calculations

Using the above code to form a 95% confidence interval (CI) for $\Delta = p_0 - p_1$ based on the score statistic and sample estimates $\hat{p_0} = 0.094$ and $\hat{p_1} = 0.065$, the score-based CI equals (0.00534 - 0.05307), which compares to a (0.00532 - 0.05268) Wald-based CI and a (0.00537 - 0.05291) LR-based CI.

Using the above code to form a 95% CI for $\theta = p_0/p_1$ based on the LR statistic and sample estimates $\hat{p_0} = 0.077$ and $\hat{p_1} = 0.073$, the LR-based CI equals (1.43168, 3.08300), which compares to a (1.42013, 3.04965) Wald-based CI and a (1.42440, 3.04467) score-based CI.

Appendix C

RCTDESIGN R CODE

C.1 Sample Power Calculation

```
> genPower <- function(p1, n0=1000, n1=1000, theta0=1.5, analyses){
+   p0 <- theta0*p1
+   delta0 <- p0 - p1
+   StudyDsn <- seqDesign(prob.model = "proportions",
+     arms = 2, ratio = c(1,1),
+     null.hypothesis = c(p1+delta0, p1), # c(placebo, SOC)
+     alt.hypothesis = c(p1, p1),
+     test.type = "less",
+     sample.size = n0+n1,
+     size = 0.025,
+     nbr.analyses = analyses,
+     P = c(1,Inf,Inf,0.5), # c(OBF bound, none, none, Pocock bound)
+     power="calculate"
+   )
+     return(seqExtract(StudyDsn, "power"))
+ }

> genPower(p1=0.03, analyses=1)
[1] 0.5024915

> genPower(p1=0.03, analyses=2)
[1] 0.483055

> genPower(p1=0.03, analyses=4)
[1] 0.461032
```

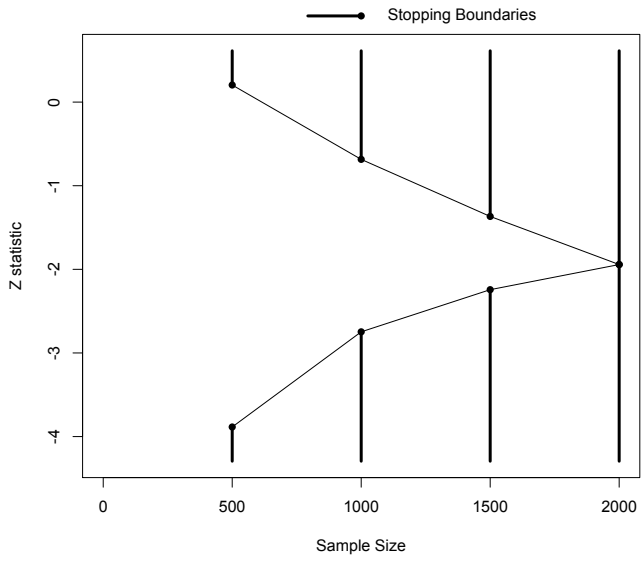
C.2 Sample Design used in Simulation

```
> genDsn <- function(p0, p1, n0, n1, theta0){
+
+   # Study Design (RR and RD will be equivalent since working on z scale)
+
+   Studydsn <- seqDesign(prob.model = "rate",
+     arms = 2, ratio = c(1,1),
+     null.hypothesis = c(p1*theta0, p1), # c(placebo, SOC)
+     test.type = "less",
+     sample.size = n0+n1,
+     size = 0.025,
+     power = .975,
+     nbr.analyses = 4,
+     P = c(1,Inf,Inf,0.5), # c(OBF bound, none, none, Pocock bound)
+     display.scale = "Z"
+   )
+     return(Studydsn)
+ }

> genDsn(p0=0.07, p1=0.05, n0=1000, n1=1000, theta0=1.5)
Call:
seqDesign(prob.model = "rate", arms = 2, null.hypothesis = c(p1 *
    theta0, p1), ratio = c(1, 1), nbr.analyses = 4, sample.size = n0 +
    n1, test.type = "less", size = 0.025, power = 0.975, P = c(1,
    Inf, Inf, 0.5), display.scale = "Z")

PROBABILITY MODEL and HYPOTHESES:
   Theta is rate ratio (Treatment : Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >= 1.5000    (size  = 0.025)
    Alternative hypothesis : Theta <= 0.5369    (power = 0.975)

STOPPING BOUNDARIES: Normalized Z-value scale
                     Efficacy Futility
    Time 1 (N=  500)  -3.8854   0.2043
    Time 2 (N= 1000)  -2.7474  -0.6850
    Time 3 (N= 1500)  -2.2432  -1.3674
    Time 4 (N= 2000)  -1.9427  -1.9427
```

C.3 Sample Plot

```
> seqPlotBoundary(SampleDsn, fixed=FALSE, display.scale="Z", dsnLbls="Stopping Boundaries")
```

Figure C.1: Sample Plot of `SampleDsn` Boundaries



Appendix D

ADDITIONAL RESULTS

D.1 Scenarios G - H

Table D.1: Additional Simulated Parameter Definitions

θ_0 = 1.5, T=3% and Δ_0=1.5%						
Scenario	p_1	p_0	θ	Δ	n_1	n_0
G1	1.0%	1.5%	1.50	0.5%	1000	1000
G2	1.0%	2.5%	2.50	1.5%	1000	1000
G3	2.0%	3.0%	1.50	1.0%	1000	1000
G4	2.0%	3.5%	1.75	1.5%	1000	1000
G5	3.0%	4.5%	1.50	1.5%	1000	1000
G6	4.0%	6.0%	1.50	2.0%	1000	1000
G7	4.0%	5.5%	1.38	1.5%	1000	1000
G8	5.0%	7.5%	1.50	2.5%	1000	1000
G9	5.0%	6.5%	1.30	1.5%	1000	1000
H1	1.0%	1.0%	1.0	0.0%	1000	1000
H2	2.0%	2.0%	1.0	0.0%	1000	1000
H3	3.0%	3.0%	1.0	0.0%	1000	1000
H4	4.0%	4.0%	1.0	0.0%	1000	1000
H5	5.0%	5.0%	1.0	0.0%	1000	1000
H6	6.0%	6.0%	1.0	0.0%	1000	1000

N simulations = 5000

Table D.2: Scenario G Fixed Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
G1	2.72	2.72	2.72	54.1	46.4	50.8	54.1	46.4	50.8
G2	0.02	0.02	0.02	3.18	2.36	2.66	3.18	2.36	2.66
G3	2.50	2.70	2.50	11.9	10.5	11.2	11.3	10.1	10.7
G4	0.48	0.50	0.48	3.02	2.70	2.86	2.70	2.44	2.58
G5	2.76	2.88	2.76	2.80	2.60	2.68	2.78	2.90	2.78
G6	2.84	2.84	2.74	0.76	0.74	0.74	2.84	2.80	2.74
G7	6.86	6.88	6.72	2.94	2.82	2.82	6.86	6.88	6.72
G8	2.62	2.72	2.52	0.26	0.20	0.26	2.62	2.72	2.52
G9	12.5	12.7	12.3	2.54	2.42	2.54	12.5	12.7	12.3
Coverage Probability (%)									
G1	95.5	94.9	91.9	95.1	95.2	94.8	95.1	95.2	94.8
G2	95.8	95.2	94.3	94.7	94.9	94.6	94.7	94.9	94.6
G3	95.1	94.8	94.5	94.3	94.6	94.4	94.0	94.2	94.1
G4	94.2	94.0	93.8	94.5	94.3	94.3	94.0	93.8	93.8
G5	95.0	94.7	94.6	94.7	94.8	94.7	94.8	94.6	94.7
G6	94.9	94.8	94.7	94.9	95.0	94.9	95.3	95.2	95.2
G7	94.9	94.6	94.6	94.8	94.8	94.8	95.3	95.0	95.1
G8	95.5	95.4	95.4	95.4	95.3	95.3	95.5	95.4	95.4
G9	95.6	95.4	95.4	95.4	95.4	95.3	95.6	95.4	95.4

Table D.3: Scenario H Fixed Trial Results

	Relative Risk			Risk Difference			Threshold Design		
Scenario	Wald	Score	LR	Wald	Score	LR	Wald	Score	LR
Rejection Percentages									
H1	15.2	15.2	15.2	91.4	87.5	90.0	91.4	87.5	90.0
H2	24.5	25.2	24.5	67.3	64.1	65.7	67.3	64.1	65.7
H3	35.4	36.2	35.4	50.9	48.3	49.5	44.9	44.3	44.4
H4	46.3	47.0	46.3	41.8	40.4	40.8	46.4	47.1	46.4
H5	54.3	54.5	53.9	34.9	34.1	34.3	54.3	54.5	53.9
H6	62.5	63.0	62.1	30.3	29.5	30.0	62.5	63.0	62.1
Coverage Probability (%)									
H1	96.2	95.2	89.2	95.2	95.2	95.0	95.2	95.2	95.0
H2	95.0	94.5	94.2	94.5	94.5	94.5	94.5	94.5	94.5
H3	95.1	94.8	94.6	94.8	94.8	94.6	94.9	94.8	94.6
H4	95.0	94.8	94.7	94.7	94.8	94.7	94.9	94.8	94.7
H5	94.7	94.6	94.5	94.5	94.6	94.5	94.7	94.6	94.5
H6	95.0	94.9	94.9	94.9	94.9	94.9	95.0	94.9	94.9