MASTER OF SCIENCE (2021) McMaster University
(Statistics) Hamilton, ONT

TITLE: Designing Factorial Clinical Trials: An Examination of Power

AUTHOR: Ian J. Belo

SUPERVISOR: Dr. Stephen D. Walter

NUMBER OF PAGES: viii, 72.

Abstract

Investigators will sometimes consider clinical trials involving more than one treatment or intervention, as these trials allow for the simultaneous evaluation of the individual efficacy of multiple treatments. The most common design choice is a factorial trial, in which patients are randomized to all possible combinations of treatments, including control. Factorial trials are an attractive choice in examining individual treatment effects provided certain conditions are met, including the important assumption that no interaction exists between the treatments of interest.

However, even without interaction, the statistical power for a treatment can be substantially influenced by the effectiveness of the other treatment in the trial, an issue that has not been widely recognized. This issue is compounded by the fact that the impact on power depends on the scale on which interactions are defined.

In the current work, we evaluate how the power for a treatment in a binary outcome 2x2 factorial trial changes as a function of the effectiveness of a second treatment in the same trial, under a range of possible parameter conditions. We provide analytical results to describe the behavior of these functions on the additive, risk ratio, and odds ratio scales and attempt to determine where the maximum power occurs for each scale.

Sets of numerical evaluations were also implemented to support these analytic results, as well to evaluate how the minimum required sample size for the trial changes as a function of the first and second treatment effects. Controllable parameters within the evaluations include the event rate in the control group, sample size, treatment effect sizes, and Type-I error thresholds. Separate evaluations were created for scenarios where the treatments are assumed to not have an interaction on either the additive, risk ratio, or odds ratio scales. We also provide two examples of factorial trials using real data to illustrate our findings.

In general, we find that power for an individual treatment decreases as a function of the effectiveness of the other treatment if they do not interact on the risk ratio scale. A similar pattern is observed in the odds ratio case at low base rates, but at high base rates, power increases may occur if the first treatment is moderately more effective than its planned value. When treatments do not interact on the additive scale, power may either increase or decrease depending on the response rate in the control group.

Results from these analyses may benefit investigators in planning clinical trials. Assumptions about the anticipated effects of each treatment under study, even if there is no interaction between them, are critical in calculating a valid sample size that will yield sufficient power for individual treatments in the context of factorial studies.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Study Outline

The objective of the current work is to examine how the statistical power for an individual treatment effect in a binary-outcome 2x2 factorial randomized controlled trial (FRCT) may be influenced by the effectiveness of the other treatment in the trial.

FRCTs are a widely used class of clinical trial designs that allow researchers to investigate the effects of two or more treatments (or factors) simultaneously. These trials can provide a more comprehensive approach to studying treatment outcomes and are advantageous in certain scenarios. Mainly, if investigators wish to examine the effects of two treatments, an FRCT may be a more efficient design as it requires fewer total participants to test each treatment than would be needed to conduct two separate single factor trials. This “two-for-one” benefit in terms of sample size efficiency is valid in studies where there is assumed to be no interaction between the two treatments.

In planning for an FRCT, investigators determine the minimum number of patients needed for recruitment by using postulated main effects of the treatments along with the desired level of power for the trial. If the main effect for one of the treatments is

analysed and found to differ from its originally proposed value, investigators may use this new estimate yielded from the analysis in favor of the value proposed prior to the study. In doing this, however, the anticipated marginal effect for the other treatment in the FRCT may change as a result. This change in the expected marginal effect for the second treatment will then result in a modification to its associated power if the originally planned sample size is maintained for the analysis. Depending on the effect of the first treatment, this could result in either an increase or decrease in power for the second treatment. While power increases for the second treatment would be of obvious benefit to investigators, scenarios where the power is found to decrease would necessitate a larger sample to maintain the planned power for the second treatment effect.

When conducting a binary outcome FRCT, this issue is further complicated given that the anticipated event rate in the combined treatment group—and by extension the anticipated main effects—depends on the scale of measurement chosen to define treatment interactions. Separate examinations of the scenarios in which the treatments in a binary outcome FRCT are assumed to have no interaction on each scale (additive, risk ratio, and odds ratio) are therefore needed to have a comprehensive understanding of how the power for one treatment may change based on the observed effect of the other treatment.

The remainder of this chapter outlines the background and motivation for the current work. A brief introduction on FRCTs is provided with specific consideration for the scenario in which they are conducted for reasons of efficiency. The chapter finishes with

general outline of the problem under study and preludes the main methods that will be used to address it throughout.

In chapter 2, relevant formulas pertaining to power and sample size calculations as well as the main scales of measurements are introduced. The relationship between power and measurement scale are then examined to derive the expected change in power for a treatment in a FRCT given some updated treatment estimate for the other factor along with the assumption of no interaction on a specific scale. We also introduce the evaluation work conducted to show (both numerically and graphically) how the power for a treatment in a FRCT changes as a function of the effectiveness of the other treatment in the trial under a variety of scenarios and parameter settings. A corresponding set of evaluations are presented to show the sample size adjustment that may be needed for the analysis of the second factor to maintain a desired level of power.

Chapter 3 introduces two real-world examples of FRCTs to illustrate how the power and required sample size for an FRCT may change depending on the observed effect of the first treatment in the trial.

Finally, chapter 4 describes the main findings and trends of the evaluation results and provides general recommendations as to how the current work may be applied to the future conduction of FRCTs. Connections of the results to broader issues in clinical trial work are then discussed, along with areas of future work.

1.2 Overview of Factorial Randomized Clinical Trials (FRCTs)

FRCTs may be used to test the effectiveness of multiple treatments simultaneously by randomly assigning participants to groups in which they receive the treatments either alone or in combination. The simplest case of a FRCT involves two treatments (A and B), in which participants are randomly assigned to receive either treatment A, treatment B, both treatment A and B, or neither (control group). Typically, control groups receive either a placebo or an alternative treatment regimen to be compared to treatments A and B. Given that the two treatments are administered at two different levels (either present or absent), the scenario described constitutes a 2x2 (two treatments each at two levels) FRCT.

Following the randomization and treatment administration phases, the outcome of interest is measured in each group at some future point in the study as determined by the investigators. In clinical trials, binary outcomes—clinical endpoints that can take on one of two possible values—are commonly used. Examples of these outcomes include mortality, disease recurrence, or a continuous outcome dichotomized to a “yes/no” value. In measuring a binary outcome, the proportion of patients in each treatment group who exhibit the outcome of interest is determined. We refer to this proportion as the event rate. When conducting a binary-outcome 2x2 FRCT, these groups and their associated event rates can be described using the following table:

Group Event Rate	Description
p_{11}	Receives both treatments
p_{10}	Receives first treatment and control for second treatment
p_{01}	Receives second treatment and control for first treatment
p_{00}	Receives control for both treatments

Binary outcomes are useful to both investigators and patients, as they can provide a simplified understanding of whether a treatment has a beneficial effect on a clinically relevant outcome. Throughout the current work, we exclusively examine 2x2 FRCTs with binary outcomes, as these outcomes are both frequent and practical.

FRCTs are advantageous for two main reasons. By using this design, researchers may examine the individual effects of each treatment while gaining the additional advantage of studying potential treatment interactions—situations in which the effectiveness of a treatment differs depending on the levels of one or more other treatments in the trial. However, if the treatment effects act independently on the outcome (i.e., there is no interaction present between them), FRCTs may be used for the advantage of efficiency, which is a more common objective. A clinical trial review by McAlister et al. (2003) found that of 44 binary outcome 2x2 FRCTs, 36 of them (82%) were conducted for the main purpose of efficiency, with the remaining 18% conducted mainly to assess interaction effects between treatments.

In 2x2 FRCTs designed for efficiency, the marginal effects for each treatment are obtained by averaging the effectiveness of each treatment across the levels of the other. The magnitude of this effect is then obtained by comparing the average effectiveness among the groups where each treatment is present to the average when it is not. When using a binary outcome, this involves measuring the event rate in each of the four treatment groups, and comparing the average event rates in the treatment-present and treatment-absent groups.

By using this technique, investigators may assess the effects of multiple treatments within the same trial, thus providing a timely and economical alternative to conducting multiple single-factor experiments. In the latter case, a single treatment effect is usually measured by comparing the average proportion in the treatment group compared to a control group. In a FRCT, the observed marginal proportions are used to assess the main effects for both treatments. By computing the treatment effects in this way, all patients in the sample are included in the analysis of each main effect, allowing investigators to analyse the effects of multiple treatments with the same sample size that would have been needed to analyse them separately.

However, the interpretation of main effects in a FRCT can be misleading if an interaction exists between the two treatments (i.e., the effect of one treatment is dependent on the level(s) of the other treatment). Scenarios where even small interactions exist can lead to biased estimates of the main treatment effects (McAlister et al., 2003). Since the presence and potential magnitude of interactions are rarely

known in advance, preliminary interaction testing within a trial is sometimes used to determine how the main effects should be evaluated. There are points of contention surrounding this topic, as FRCTs are frequently underpowered to detect interactions and investigators may use inappropriate methods to assess them (Kahan et al., (2019), Montgomery (2003)). Moreover, the definition of an interaction depends on the scale of measurement used, and the absence of interaction on one scale directly implies interaction on another scale (Vanderweele, 2014). In the current work, we exclusively consider FRCTs where no interaction is assumed on a specified scale, and the analysis of the main effects is therefore unaffected. We separately consider the cases in which no interaction is assumed on the additive, risk ratio, and odds ratio scales.

1.3 Study Motivation

In a standard FRCT where no interaction is present, the minimum number of participants needed to obtain a desired level of power for a treatment main effect may be determined using sample-size calculations (the details of which are explained in the next chapter). In a 2x2 FRCT, sample size determinations are typically calculated for the anticipated main effects of both treatments, with the maximum of the two taken to be the total sample size needed for the FRCT. This practice is meant to ensure sufficient power for both treatment effects. In the case where the calculated sample size for one treatment is larger than the other, this will result in a total sample size that provides the desired target power for the treatment requiring the larger sample size while providing excess power (“overpowering”) for the treatment requiring the smaller sample size.

However, it is critical to note that the marginal event rates anticipated for a treatment in an FRCT will depend on the other treatment effect. Consider again two treatments, A and B, administered in a 2x2 FRCT with some binary outcome. If the treatment effect of A is analysed first and is found to differ from its originally planned value, investigators may use the estimate in favor of the planned value as their best information on the effect of A. Now using the estimate of A's effect, the expected marginal event rates for B will be modified, and by association, its power. This is true provided that the investigators maintain their originally planned estimate for the effect of B as well as the assumption of no interaction between the two treatments.

By realizing that the power for the second treatment may have changed due to the effectiveness of the first, investigators may be better positioned to adapt the study accordingly. For example, they may choose to recruit additional patients prior to the analysis of the second treatment if it is found that the power has decreased due to the effectiveness of the first. Understanding how the power of a main effect in a 2x2 FRCT is influenced by the effectiveness of the other factor can be useful in ensuring that investigators have a sufficient sample to maintain the planned power throughout the study.

This issue is complicated further by the impact of measurement scales. As mentioned previously, the analysis of treatment main effects in a 2x2 FRCT depends on the assumption of no interaction. However, when measuring a binary outcome, interactions may be defined differently depending on the scale of measurement. This

means that the event rate expected in the combined treatment group (A^+B^+) in an FRCT depends not only on the main effect estimates of both treatments but on the interaction scale as well. Descriptions of three common scales of measurement (additive, risk ratio, and odds ratio) and definitions of interactions on each respective scale are provided in chapter 2.

The main interest of the current work is to analyse how the power for a treatment effect in a binary 2x2 FRCT is impacted by the main effect estimate of the other treatment in the trial. Numeric evaluations are used to illustrate these effects through both graphs and tables. These evaluations consider a variety of parameter inputs relevant to the power of the factors in the study, including the base rate, sample size, α-level, and whether power is assessed using a one or two-tailed test. Separate evaluations are created for the cases where the treatments are assumed to have no interaction on the risk ratio, odds ratio, or additive scale, the results of which will be individually discussed.

A second set of evaluations is used to determine the change in the minimum sample size needed to achieve sufficient power, again using adjustable parameter settings. This set of evaluations may be of more practical use to investigators, as it depicts the adjustment in sample size that would be required to compensate for a potential loss in power.

Together, these evaluations may be used to both better understand how power changes under a variety of experimental scenarios and serve as a guide to anticipate how power may change in a planned 2x2 FRCT. In chapter 3, two examples are provided using real data to illustrate the potential for the power of a treatment to vary in a FRCT.

Chapter 2

Methodology

2.1 Power and Sample Size Calculations

The probability of finding a significant treatment effect in a clinical trial naturally depends on the sample size of the trial. The smaller the effect, the more difficult it will be to detect, and as such a larger sample will be required. In planning a study, investigators may use a postulated estimate of the treatment effectiveness to determine the power for the study, which is the probability of finding a statistically significant effect given that a treatment effect of the postulated size is present. In dealing with binary data, the estimated power to detect the difference two groups (treatment and control) relies upon the experimental parameters and typically involves a binomial approximation to the standard normal distribution. The power to detect the difference between two independent proportions, p_1 and p_2, using a two-sided test with respective sample sizes of n_1 and n_2 can be expressed by the formula:

$$\text{Power} = \Phi\left[\frac{|p_2 - p_1|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - Z_{(1-\alpha/2)}\frac{\sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}\right] \quad (\mathbf{2.1})$$

where $\bar{p} = (n_1p_1 + n_2p_2)/(n_1 + n_2)$, Φ is the cumulative distribution function of the standard normal distribution, and $Z_{(1-\alpha/2)}$ is the Z-score associated with the $(1-\alpha/2)^{th}$

percentile of the standard normal distribution. Assuming equal sample sizes in the two groups under comparison ($n = n_1 = n_2$), this equation simplifies to:

$$\text{Power} = \Phi\left[\frac{|p_2 - p_1| - Z_{(1-\alpha/2)}\sqrt{\bar{p}(1-\bar{p})(2/n)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)/n}}\right] \quad (\mathbf{2.2})$$

For a one-sided test, the $Z_{(1-\alpha/2)}$ terms in equations 1.1 and 1.2 are replaced by $Z_{(1-\alpha)}$. In a 2x2 FRCT, marginal estimates of the overall control and treatment group proportions for each treatment are obtained to individually assess the power for each treatment. A treatment's marginal estimates are calculated by taking the arithmetic average of the estimated group proportions for the treatment-present and treatment-absent groups across the levels of the other treatment in the trial. The calculation of marginal estimates is explained in Section 2.3, with consideration as to how these estimates differ depending on the scale of measurement used.

A modified version of the power equation may be used to determine the minimum sample size needed to obtain a specified power given an anticipated effect size. This equation is often of more practical use to investigators as trials are typically designed with a planned power in mind. A target power of 0.80 is commonly suggested by clinical guidelines (Sakpal, 2010). Assuming a two-tailed test, this formula is given by:

$$n^* = \left(\left[\sqrt{\bar{p}(1-\bar{p})} \times Z_{\left(1-\frac{\alpha}{2}\right)} - \sqrt{p_1(1-p_1) + p_2(1-p_2)} \times Z_{(1-\beta)}\right] / |p_2 - p_1|\right)^2 \quad (\mathbf{2.3})$$

where β is equal to one minus the planned power for the study (this is also known as the Type II error rate).

A continuity-corrected estimate of the sample size yielded from the above equation can be used to give a more accurate approximation of the sample size needed to reach the desired level of power. Once such correction, introduced by Casagrande, Pike, and Smith (1978), is given by:

$$n = \frac{n^*}{4}\left(1 + \sqrt{1 + \frac{4}{n^*|p_2 - p_1|}}\right)^2 \quad (\mathbf{2.4})$$

This version of continuity correction is regarded as a highly accurate approximation and is used throughout the evaluation work to obtain sample size estimates.

In a 2x2 FRCT, equations 2.3 and 2.4 can be used to calculate the required sample size for each treatment main effect, with p_1 and p_2 set as the expected marginal proportions for each treatment, resulting in two sample sizes. Afterwards, the maximum of the two sample sizes can be used as the total sample size for the trial. This results in having minimum sufficient power for the treatment with the larger calculated sample size and more-than-sufficient power for the treatment with the smaller sample size.

2.2 Scales of Measurement

In determining whether an interaction is present, it is critical to first establish the scale of measurement used to define them. Here, the additive, risk ratio, and odds ratio scales are introduced along with definitions of interactions on their respective scales.

On the additive scale, effect sizes are defined in terms of absolute risk reduction, which is the arithmetic difference in proportions between a group of participants administered a treatment and a group who does not receive the treatment (control). In a 2x2 FRCT, the interaction between two treatments on the additive scale is defined using the proportions of each treatment group:

$$p_{11} - p_{10} - p_{01} + p_{00}$$

If the above quantity does not differ from zero, then there is said to be no additive interaction between the two treatments. A positive or negative result would constitute a positive or negative interaction on the additive scale, respectively.

On the risk ratio scale, effect sizes are instead defined by the proportional difference between the treatment and control groups. Risk ratios are computed by taking the ratio of the average risk in a treatment group and the average risk in the control group. Thus, for two treatments in a 2x2 FRCT, the risk ratios for each treatment group are:

$$RR_{(11)} = \frac{p_{11}}{p_{00}}$$

$$RR_{(10)} = \frac{p_{10}}{p_{00}}$$

$$RR_{(01)} = \frac{p_{01}}{p_{00}}$$

Interactions on the risk ratio scale can be calculated by:

$$\frac{RR_{(11)}}{RR_{(10)} \times RR_{(01)}} = \frac{p_{11} \times p_{00}}{p_{10} \times p_{01}}$$

If the above ratio equates to one, then there is said to be no interaction between A and B on the risk ratio scale. A ratio greater than one indicates a positive interaction while a ratio less than one indicates a negative interaction.

Odds ratios are an alternative proportional measure that takes the proportion of the odds in a treatment group to the odds in the control group. The odds ratios for each of the three treatment groups in a 2x2 FRCT are:

$$OR_{(11)} = \frac{p_{11}/(1-p_{11})}{p_{00}/(1-p_{00})}$$

$$OR_{(10)} = \frac{p_{10}/(1-p_{10})}{p_{00}/(1-p_{00})}$$

$$OR_{(01)} = \frac{p_{01}/(1-p_{01})}{p_{00}/(1-p_{00})}$$

Interactions on the odds ratio scale can be assessed similarly by the ratio:

$$\frac{OR_{(11)}}{OR_{(10)} \times OR_{(01)}}$$

As before, no interaction is said to be present if the above ratio equals one. A ratio greater than one or less than one corresponds to a positive or negative interaction on the odds ratio scale, respectively.

2.3 Relationship Between Power and Scales of Measurement

To generate accurate sample size estimates in a 2x2 FRCT, investigators must estimate both the simple and combined effects of each treatment. Under the assumption of no interaction, the anticipated effect size in the combined treatment group depends on the scale of measurement used to define treatment effects. This in turn influences the main effect estimates of both treatments and the minimum sample size required to detect them at a specified level of power.

In this section, we show algebraically how the power for a treatment in a 2x2 FRCT may change as a direct consequence of the other treatment producing a main effect different from its planned value. We also show that the associated minimum required sample size required to achieve a desired level of power changes depending on the effectiveness of the other treatment in the trial. Separate cases are examined for when the treatments in the FRCT are presumed to have no interaction on the additive, risk ratio, and odds ratio scales.

For the first case, suppose a trial is planned with two treatments (A and B) which have some additive effect on a binary response outcome with no additive interaction between them. Let p_{00} denote the population base rate, with $\Delta_A = p_{10} - p_{00}$ and $\Delta_B = p_{01} - p_{00}$ denoting the postulated additive effects of treatments A and B, respectively. Under these assumptions, the expected marginal proportions in the control (p_1) and treatment (p_2) groups used to assess the main effect of treatment B are:

$$p_1 = \frac{p_{00} + (p_{00} + \Delta_A)}{2} = p_{00} + \frac{\Delta_A}{2} \quad \text{(Treatment B Absent)}$$

$$p_2 = \frac{(p_{00} + \Delta_B) + (p_{00} + \Delta_A + \Delta_B)}{2} = p_{00} + \Delta_B + \frac{\Delta_A}{2} \quad \text{(Treatment B Present)}$$

Suppose that treatment A is analysed first, and investigators observe that A appears to have a treatment effect different from its planned estimate. From this observation, the investigators update their estimate of Δ_A to $(\Delta_A + k)$, where k is bounded by $(-(p_{00} + \Delta_A) \leq k \leq 1 - p_{00} - \Delta_A)$. When determining the marginal effect of treatment B, the new expected observed proportions of the control (p_1') versus treatment (p_2') groups are:

$$p_1' = p_{00} + \frac{\Delta_A + k}{2} \quad \text{(Treatment B Absent)}$$

$$p_2' = p_{00} + \Delta_B + \frac{\Delta_A + k}{2} \quad \text{(Treatment B Present)}$$

Given these new proportions, if the power for treatment B were re-assessed using the same sample and effect size for treatment B as originally planned, the updated power for treatment B becomes:

$$\text{Power} = \Phi\left[|p_2' - p_1'| - Z_{(1-\alpha/2)}\sqrt{\bar{p}'(1-\bar{p}')(2/n)} \Big/ \sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}\right] \quad (\mathbf{2.5})$$

Thus, if the power for treatment B is determined according to some planned Δ_A and Δ_B, and the observed effect of A is found to be $(\Delta_A + k)$, the change in power for treatment B may be expressed as the difference in powers calculated using Δ_A and $(\Delta_A + k)$, holding all other parameters constant:

$$\Delta\text{Power} = \Phi\left[\frac{|p_2 - p_1| - Z_{(1-\alpha/2)}\sqrt{\bar{p}(1-\bar{p})(2/n)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)/n}}\right] - \Phi\left[\frac{|p_2' - p_1'| - Z_{(1-\alpha/2)}\sqrt{\bar{p}'(1-\bar{p}')(2/n)}}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right] (\mathbf{2.6})$$

In this same scenario, the difference in the minimum required sample size needed to achieve a desired power target can be calculated. Equations 2.3 may be used to determine the sample size estimates (${n^*}_A$ and ${n^*}_B$) needed to power for the

main effects of treatments A and B, respectively. The marginal proportions used in the calculation of n^*_A are:

$$p_1 = p_{00} + \frac{\Delta_B}{2} \quad \text{(Treatment A Absent)}$$

$$p_2 = p_{00} + \Delta_A + \frac{\Delta_B}{2} \quad \text{(Treatment A Present)}$$

Similarly, the proportions used in the calculation of n^*_B are:

$$p_1 = p_{00} + \frac{\Delta_A}{2} \quad \text{(Treatment B Absent)}$$

$$p_2 = p_{00} + \Delta_B + \frac{\Delta_A}{2} \quad \text{(Treatment B Present)}$$

Equation 1.4 can then be used to produce continuity corrected estimates of n^*_A and n^*_B (n_A and n_B), with the planned sample size set to $\max(n_A, n_B)$. If the observed effect of treatment A is found to be $(\Delta_A + k)$, the updated estimates for the marginal proportions for treatment B become:

$$p_1' = p_{00} + \frac{\Delta_A + k}{2} \quad \text{(Treatment B Absent)}$$

$$p_2' = p_{00} + \Delta_B + \frac{\Delta_A + k}{2} \quad \text{(Treatment B Present)}$$

The updated sample size calculation for treatment B, denoted as $n_{B'}$, can then be obtained using the same formulas as before by replacing $p_1$and p_2 with p_1' and p_2'. Therefore, the difference in the minimum sample size needed for treatment B when the effect size of treatment A is Δ_A versus $(\Delta_A + k)$ is:

$$\Delta n = n_{B'} - \max(n_A, n_B) \quad (\mathbf{2.7})$$

Next, we examine the risk ratio case. Let us first assume that A and B again are two treatments acting on a binary outcome this time with no interaction on the risk ratio scale. Here, let $\theta_A = p_{10}/p_{00}$ and $\theta_B = p_{01}/p_{00}$ denote the respective postulated

multiplicative effects of treatments A and B on the outcome variable. The expected marginal proportions in assessing the main effect of treatment B in this case are:

$$p_1 = \frac{p_{00} + (p_{00}\theta_A)}{2} = \frac{p_{00}(1+\theta_A)}{2} \quad \text{(Treatment B Absent)}$$

$$p_2 = \frac{p_{00}\theta_B + (p_{00}\theta_A\theta_B)}{2} = \frac{p_{00}\theta_B(1+\theta_A)}{2} \quad \text{(Treatment B Present)}$$

If treatment A is analysed first and its effect estimate is updated to $k\theta_A$ for some $k > 0$, then the new expected proportions for treatment B become:

$$p_1' = \frac{p_{00}(1+k\theta_A)}{2} \quad \text{(Treatment B Absent)}$$

$$p_2' = \frac{p_{00}\theta_B(1+k\theta_A)}{2} \quad \text{(Treatment B Present)}$$

As before, equations 2.6 and 2.7 may be used to determine the change in power and sample size difference for treatment B when the trial is planned with an effect size of $\theta_A$for treatment A but an effect size of $k\theta_A$is observed in the analysis.

Finally, we consider the odds ratio scale. Let θ_A and θ_B again represent the respective multiplicative effects of treatments A and B, respectively. If there is no interaction on the odds ratio scale, the expected marginal proportions for treatment B are:

$$p_1 = \frac{p_{00} + (p_{00}\theta_A)}{2} = \frac{p_{00}(1+\theta_A)}{2} \text{ (Treatment B Absent)}$$

$$p_2 = \frac{\theta_A\theta_B p_{00}(1-p_{00})}{2[(1-\theta_A p_{00})(1-\theta_B p_{00}) + \theta_A\theta_B p_{00}(1-p_{00})]} + \frac{p_{00}\theta_B}{2}$$

It can be shown that p_2 can also be expressed together in one term as:

$$p_2 = \frac{p_{00}\theta_B[\theta_A(1-p_{00}) + (1-\theta_A p_{00})(1-\theta_B p_{00}) + \theta_A\theta_B p_{00}(1-p_{00})]}{2[(1-\theta_A p_{00})(1-\theta_B p_{00}) + \theta_A\theta_B p_{00}(1-p_{00})]} \quad \text{(Treatment B Present)}$$

Again, if treatment A is analysed first resulting in an effect estimate of $k\theta_A$ for some $k > 0$, then the new expected marginal proportions for treatment B become:

$$p_1' = \frac{p_{00} + (p_{00} k\theta_A)}{2} = \frac{p_{00}(1 + k\theta_A)}{2} \text{ (Treatment B Absent)}$$

$$p_2' = \frac{k\theta_A \theta_B p_{00}(1 - p_{00})}{2[(1 - k\theta_A p_{00})(1 - \theta_B p_{00}) + k\theta_A \theta_B p_{00}(1 - p_{00})]} + \frac{p_{00}\theta_B}{2}$$

Equations 2.4 and 2.7 may then be used to determine the change in power for B given the updated effect of A by inserting the values of p_1, p_2, p_1', and p_2'. In all three cases, equations 2.3 and 2.4, and 2.7 may be used to determine the change in the minimum sample size needed to analyse treatment B with the same level of power planned prior to the analysis of treatment A.

2.4 Analysis of Power Across Scales of Measurement

Across the scales discussed, it is assumed that both treatments in the FRCT are intended to reduce the event rate of the outcome under study. As such, we examine cases for which the first treatment analyzed in the trial either reduces the event rate or has a zero (null) effect. For the second treatment, we similarly examine cases where the treatment reduces the event rate, omitting the trivial case where the second treatment has no effect (as this will render the power for the second treatment equal to α divided by the number of tails of the test, regardless of the first treatment's effect).

Given the possibility that the power for the second treatment in the trial may change depending on the effect of the first treatment, of particular interest are the cases

in which the second's power treatment attains a maximum or minimum based on the first treatment's observed effect. In the current section, we attempt to determine these points analytically for each of the three scales of measurement discussed (additive, risk ratio, and odds ratio). In sections 2.9 – 2.11, numerical evaluations are introduced to substantiate these results.

Beginning with the additive scale, given some postulated Δ_A, we wish to determine the value of k such that the observed effect of treatment A, $\Delta_A + k$, maximizes the power for treatment B. On this scale, the power for treatment B is calculated by:

$$\text{Power} = \Phi\left[\frac{|p_2' - p_1'| - Z_{(1-\alpha/2)}\sqrt{\bar{p'}(1-\bar{p'})(2/n)}}{\sqrt{[\,p_1'(1-p_1') + p_2'(1-p_2')\,]/n}}\right]$$

$$p_1' = p_{00} + \frac{\Delta_A + k}{2}, \qquad p_2' = p_{00} + \Delta_B + \frac{\Delta_A + k}{2}, \qquad \bar{p'} = \frac{p_1' + p_2'}{2}$$

where Φ is the standard normal cumulative density function. For simplicity in notation, we define $f = \Delta_A + k$, and $\boldsymbol{\phi}$ as the standard normal probability density function. In this section, as well as in the following sections examining the risk ratio and odds ratio scales, we will further assume that the sample size is sufficiently large such that the term $Z_{(1-\alpha/2)}\sqrt{\bar{p'}(1-\bar{p'})(2/n)}$ is approximately zero, thus simplifying the power expression to:

$$\text{Power} = \Phi\left[\frac{|p_2' - p_1'|}{\sqrt{[\,p_1'(1-p_1') + p_2'(1-p_2')\,]/n}}\right]$$

Note that we considered this modified power equation only to simplify the algebraic analysis in this section; the forthcoming numerical evaluations later in chapter 2 use the full power function. To further aid notation, we will define the argument inside

of the density function as $Q = \frac{|p_2' - p_1'|}{\sqrt{[\,p_1'(1-p_1') + p_2'(1-p_2')\,]/n}}$. Next, we calculate the change in power for treatment B with respect to f:

$$\frac{d(Power)}{df} = \big(\phi(Q)\big)\left(\frac{d(Q)}{df}\right)$$

where for the first term in the product:

$$\phi(Q) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-Q^2}{2}\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(p_2' - p_1')^2}{2([\,p_1'(1-p_1') + p_2'(1-p_2')\,]/n)}\right)$$

For the second term in the product, by separating and differentiating by parts, we have:

$$\frac{d(Q)}{df} = \frac{d}{df}\left[\frac{1}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right][|p_2' - p_1'|] + \frac{d}{df}[|p_2' - p_1'|]\left[\frac{1}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right]$$

It can be shown that for the partial derivative of the denominator portion is:

$$\frac{d}{df}\left(\frac{1}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right)$$

$$= \frac{d}{df}\left(\left(\sqrt{[(p_{00} + \frac{f}{2})\left(1 - (p_{00} + \frac{f}{2})\right) + (p_{00} + \Delta_B + \frac{f}{2})\left(1 - \left(p_{00} + \Delta_B + \frac{f}{2}\right)\right)]/n}\right)^{-1}\right)$$

$$= \frac{-\sqrt{n}\,(1 - 2p_{00} - f - \Delta_B)}{2\left(\left(p_{00} + \frac{f}{2}\right)\left(1 - (p_{00} + \frac{f}{2})\right) + (p_{00} + \Delta_B + \frac{f}{2})\left(1 - \left(p_{00} + \Delta_B + \frac{f}{2}\right)\right)\right)^{3/2}}$$

And for the partial derivative of the numerator portion:

$$\frac{d}{df}(|p_2' - p_1'|)$$

$$= \frac{d}{df}[\Delta_B\,] = 0$$

Together, we have:

$$\frac{d(Q)}{df} = \frac{-\sqrt{n}(1 - 2\,p_{00} - f - \Delta_B)(|p_2' - p_1'|)}{2\left[(p_{00} + \frac{f}{2})\left(1 - (p_{00} + \frac{f}{2})\right) + (p_{00} + \Delta_B + \frac{f}{2})\left(1 - \left(p_{00} + \Delta_B + \frac{f}{2}\right)\right)\right]^{\frac{3}{2}}}$$

From the above equations, we observe that $\phi(Q) \to 0$ as $Q \to \pm\infty$. This will occur when $\left(p_1'(1-p_1') + p_2'(1-p_2')\right)/n \to 0$, which is possible under several scenarios. First, if $p_1' = p_2' = 0$, a degenerate case in which the event rates in all four treatment group cells are zero. A second possibility is if $p_1' = p_2' = 1$, which is an alternative degenerate scenario where the event rates in all treatment cells are 1 with no effect of either treatment. Further analysis yielded $f = 1 - \Delta_B \pm \sqrt{1 - {\Delta_B}^2} - 2\,p_{00}$ as possible algebraic roots, which also has the feasible solution $f = 0$ when $\Delta_B = 0$ and $p_{00} = 1$ (again, the scenario with event rates of 1 in all treatment cells).

Next examining $\left(\frac{d(Q)}{df}\right)$, the partial derivative will equal 0 if either $|p_2' - p_1'| = 0$ or $(1 - 2\,p_{00} - f - \Delta_B) = 0$. The former clearly occurs when $p_2' = p_1'$, the trivial case where neither treatment has an effect (and therefore the power does not change). The latter occurs when $k = 1 - 2\,p_{00} - \Delta_A - \Delta_B$, which is a feasible solution. An example using this solution is provided in section 2.8. As we will also see, points of zero-change in the power for B tend to occur at base rates close to 0.5. At lower base rates, the power for B increases as a function of the effect of A, with the inverse relationship observed at larger base rates. This means that the power curves in the additive case may either have a minimum or maximum power depending on the base rate.

Next, we examine the risk ratio model. On the risk ratio scale, the postulated effect of the first treatment (A) is denoted by θ_{A}, with the observed effect of A denoted by $k\theta_{\mathrm{A}}$. As before, we wish to determine the value of k to maximize the power of the second treatment, B. The power for treatment B is calculated by (again ignoring the second numerator term):

$$\text{Power} = \Phi\left(\frac{|p_2' - p_1'|}{\sqrt{[\,p_1'(1-p_1') + p_2'(1-p_2')\,]/n}}\right)$$

$$p_1' = \frac{p_{00}(1+k\theta_{\mathrm{A}})}{2}, \quad p_2' = \frac{p_{00}\theta_{\mathrm{B}}(1+k\theta_{\mathrm{A}})}{2}, \quad \overline{p'} = \frac{p_1' + p_2'}{2}$$

Here, we will define the quantity $g = (1 + k\theta_{\mathrm{A}})$ with Q again denoting the argument inside of the distribution function. Analogous to the additive case, the partial derivative with respect to g can be calculated by:

$$\frac{d(Power)}{df} = (\phi(Q))\left(\frac{d(Q)}{dg}\right)$$

Where:

$$\phi(Q) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(p_2' - p_1')^2}{2([\,p_1'(1-p_1') + p_2'(1-p_2')\,]/n)}\right)$$

Following the same strategy as in the additive case and calculating $\left(\frac{d(Q)}{dg}\right)$ via differentiation by parts, we have:

$$\frac{d(Q)}{dg} = \frac{d}{dg}\left[\frac{1}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right][|p_2' - p_1'|\,] + \frac{d}{dg}[|p_2' - p_1'|\,]\left[\frac{1}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right]$$

We can then calculate the partial derivatives of the denominator and numerator portions separately, yielding:

$$\frac{d}{dg}\left[\frac{1}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right]$$

$$= \frac{d}{dg}\left[\frac{1}{\sqrt{\frac{p_{00}(1+g)}{2}\left(1-\frac{p_{00}(1+g)}{2}\right)+\left(\frac{p_{00}\theta_B(1+g)}{2}\right)\left(1-\frac{p_{00}\theta_B(1+g)}{2}\right)/n}}\right]$$

$$= \frac{\sqrt{n}\,(p_{00}-{p_{00}}^2 g+p_{00}\theta_B-{p_{00}}^2 g{\theta_B}^2)}{4\left(\left(\frac{p_{00}g}{2}\right)\left(1-\frac{p_{00}g}{2}\right)+\left(\frac{p_{00}g\theta_B}{2}\right)\left(1-\frac{p_{00}g\theta_B}{2}\right)\right)^{3/2}}$$

$$\frac{d}{dg}[|p_2'-p_1'|\,]$$

$$= \frac{d}{dg}\left[\frac{p_{00}g(1-\theta_B)}{2}\right] = \frac{p_{00}(1-\theta_B)}{2}$$

Together,

$$\frac{d(Q)}{dg} = \frac{\sqrt{n}\,(p_{00}-{p_{00}}^2 g+p_{00}\theta_B-{p_{00}}^2 g{\theta_B}^2)[|p_2'-p_1'|\,]}{4\left(\left(\frac{p_{00}g}{2}\right)\left(1-\frac{p_{00}g}{2}\right)+\left(\frac{p_{00}g\theta_B}{2}\right)\left(1-\frac{p_{00}g\theta_B}{2}\right)\right)^{3/2}} + \frac{\left(\frac{p_{00}(1-\theta_B)}{2}\right)}{\sqrt{p_1'(1-p_1')+p_2'(1-p_2')/n}}$$

As was observed in the additive case, $\phi(Q)=0$ when $p_1' = p_2' = 0$ or $p_1' = p_2' = 1$. In looking for solutions to $\left(p_1'(1-p_1')+p_2'(1-p_2')\right)=0$, which would also make $\phi(Q)=0$, the only root we find is when $k\theta_{\mathrm{A}} = \frac{4-\left(p_{00}(1+\theta_B)\right)}{p_{00}(1+\theta_B)}$, which only has a feasible solution of $k\theta_{\mathrm{A}}=1$ which occurs when $p_{00}=\theta_{\mathrm{B}}=1$ (the trivial case where the event rates are 1 in all of the treatment cells). For the second portion of the derivative, $\frac{d(Q)}{dg}$, the quantity is strictly negative at all constrained values of the parameters and has no real roots apart from the trivial case of $p_1' = p_2'$ (neither treatment has any effect).

Together, this means that except for the degenerate cases, $\frac{d(Power)}{df}$ is a strictly negative function, meaning that the power for treatment B always decreases as the multiplicative effect of treatment A increases, with a maximum power achieved when A

has no effect (i.e., a multiplicative effect of 1). The numerical evaluations for the relative risk model introduced in section 2.9 illustrate these trends.

Finally, in examining power on the odds ratio scale, we denote the postulated effect of treatment A in terms of its odds ratio as OR_{A}, with the observed effect of A denoted by kOR_{A}, which we will denote as h. The power for treatment B in this scenario is calculated by (ignoring the second term in the numerator):

$$\text{Power} = \Phi\left[\frac{|p_2' - p_1'|}{\sqrt{[\,p_1'(1-p_1') + p_2'(1-p_2')\,]/n}}\right]$$

$$p_1' = \frac{p_{00}(1-p_{00})}{2(1-p_{00}+h)} \qquad p_2' = \frac{p_{00}OR_{\mathrm{B}}(2hOR_{\mathrm{B}}p_{00} - hp_{00} - p_{00} + h + 1)}{2(OR_{\mathrm{B}}p_{00} - p_{00} + 1)(hOR_{\mathrm{B}}p_{00} - p_{00} + 1)}$$

As in the previous two cases, we have:

$$\frac{d(Power)}{dh} = (\phi(Q))\left(\frac{d(Q)}{dh}\right)$$

where:

$$\phi(Q) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(p_2' - p_1')^2}{2([\,p_1'(1-p_1') + p_2'(1-p_2')\,]/n)}\right)$$

Now, in evaluating the derivative of the inside power quantity with respect to h, we have:

$$\frac{d(Q)}{dh} = \frac{d}{dh}\left[\frac{1}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right][|p_2' - p_1'|\,] + \frac{d}{dh}[|p_2' - p_1'|\,]\left[\frac{1}{\sqrt{p_1'(1-p_1') + p_2'(1-p_2')/n}}\right]$$

where it can be shown that:

$$\frac{d}{dh}[|p_2' - p_1'|\,] = \frac{p_{00}\left(1 + OR_{\mathrm{B}}{}^2\frac{hp_{00}{}^2}{(1-p_{00})^2} - OR_{\mathrm{B}}(1+h^2+1)\right)}{2\left(1+\frac{hp}{1-p}\right)^2\left(1+\frac{OR_{\mathrm{B}}hp}{1-p}\right)^2}$$

Also, after denoting $m = \frac{p_{00}}{1-p_{00}}$:

$$\frac{d}{dh}\left[\frac{1}{\sqrt{p_1'(1-p_1')+p_2'(1-p_2')/n}}\right]$$

$$= \left[\left(\frac{-p_{00}}{2}-\frac{hm}{2(hm+1)}\right)\left(\frac{p_{00}}{2}+\frac{hm}{2(hm+1)}-1\right)+\left(\frac{-OR_{\mathrm{B}}}{2(OR_{\mathrm{B}}m+1)}-\frac{OR_{\mathrm{B}}hm}{2(OR_{\mathrm{B}}hm+1)}\right)\left(\frac{1}{n}\right)\right]^{-1/2}$$

$$\times\left[\left(\frac{1}{n}\right)\left(\frac{-m^2p_{00}h-mp_{00}+m}{2(hm+1)^3}-\frac{OR_{\mathrm{B}}m(OR_{\mathrm{B}}-OR_{\mathrm{B}}m+OR_{\mathrm{B}}{}^2hm-1)}{2(OR_{\mathrm{B}}m+1)(OR_{\mathrm{B}}hm+1)^3}\right)\right]$$

the above two portions can then be assembled into the full derivative $\frac{d(Q)}{dh}$.

As with the previous two scales, $\phi(Q)=0$ when $p_1'=\ p_2'=0$ or $p_1'=\ p_2'=1$. There are no other real solutions such that $\left(\ p_1'(1-p_1')+p_2'(1-p_2')\right)=0$. For the second part of the derivative, the roots such that $\frac{d(Q)}{dh}=0$ are difficult to solve in closed form, apart from the degenerate case of $p_1'=\ p_2'$. While we could not find a real solution to $\frac{d}{dh}\left(\left(\ p_1'(1-p_1')+p_2'(1-p_2')/n\right)^{-1/2}\right)=0$, we do find that $\frac{d}{dh}[|p_2'-p_1'|\ ]=0$ under two degenerate cases, when $p_{00}=0$ (and therefore, no event rates in any of the groups), and when $OR_{\mathrm{B}}=1$ (when treatment B has no effect). A third non-degenerate solution occurs when $h=kOR_{\mathrm{A}}=\frac{(1-p_{00})}{p_{00}\sqrt{OR_{\mathrm{B}}}}$. Under this third condition, $\frac{d(Q)}{dh}=\frac{d}{dh}\left(\left(\ p_1'(1-p_1')+p_2'(1-p_2')/n\right)^{-1/2}\right)[|p_2'-p_1'|\ ]$, which, if approximately equal to zero, implies $\frac{d(Power)}{dh}=\ 0$. Therefore, the effect of treatment A such that $kOR_{\mathrm{A}}=\frac{(1-p_{00})}{p_{00}\sqrt{OR_{\mathrm{B}}}}$ should correspond (approximately) to the maximum power for treatment B. As we will illustrate through an example in section 2.7, $kOR_{\mathrm{A}}=\frac{(1-p_{00})}{p_{00}\sqrt{OR_{\mathrm{B}}}}$ indeed yields a reasonable approximation for the effect of treatment A which maximizes power.

Across all three scales, we can use optimization methods to closely approximate the value of the effect of treatment A for which the maximum power of treatment B occurs. This involves testing for different values of the effect of treatment A within the constraints of the parameters (event rates in all treatment cells must be between 0 and 1, both treatments must not increase the event rate, the sample size must be a positive integer, etc.). This is particularly useful in the odds ratio case, as we could not find a closed form solution. As we will see in the odds ratio numerical evaluations in the subsequent section, the maximum power occurs at a moderate effect of the first treatment when the base rate is large. At smaller base rates, the odds ratio power trends closely approximate those observed in the risk ratio case, i.e., the power for treatment B strictly decreases as a function of the effectiveness of treatment A.

2.5 Introduction to Numerical Evaluation Work

In the previous section, analytic methods for calculating the power for a treatment effect based on the effect estimate of the other treatment in the trial were discussed. In the current section, we introduce numeric and graphical evaluations to show how power may change under a variety of different scenarios. Separate sets of evaluations were created for when the examined treatments are presumed to have no interaction on either the risk ratio, odds ratio, or additive scales. For each set of evaluations, example outputs are provided and discussed along with notes of the general trends observed. The evaluation examples presented throughout use two-sided testing with $\alpha = 0.05$. Manipulating the α-level or “number of tails” parameters changes the power at each

point in the evaluation table and graph but does not affect the overall power trends that are described in the following sections.

2.6 General Assumptions

- Evaluations examine the power and associated sample sizes of binary outcome 2x2 FRCTs with two treatments (A & B), each with some postulated effect size.
- Treatment A is analyzed first and its observed effect size is used in favor of the effect originally postulated.
- For all evaluations, the base rate, alpha level, and number of tails are taken as required inputs by the user.
- The evaluations calculating power require as input the total sample size, while the evaluations used to calculate the minimum required sample size take as required input the desired target power.
- In all cases, the total sample size is assumed to be randomized equally across the four treatment groups.
- In all cases, it is assumed that the outcome event is undesirable (e.g., mortality), and both treatments in the trial are intended to reduce the event rate.

2.7 Determination of Effect Size Range - Cohen's h

In designing the evaluations, the range of possible effect sizes for each treatment in the 2x2 FRCT needs to be established. The goal of setting a range is to enable the examination of a variety of possible combinations of effect sizes while restricting the focus to effect sizes that may be realistically predicted or estimated by investigators. In

other words, the range should ideally include effect sizes that are detectable while excluding those that are extremely large and unlikely in practical settings.

When considering binary outcomes, effect sizes are calculated by computing the absolute difference between two group proportions ($|p_1 - p_2|$). As such, a naïve assumption might be to use the additive difference between the proportions to directly measure the detectability of the effect, as larger effect sizes should intuitively be easier to detect. This is not sufficient however, as the power to detect an effect size does not depend solely on the difference $|p_1 - p_2|$. For example, if $p_1 = 0.2$ and $p_2 = 0.1$, a sample size of 200 patients per group would yield 80% power to detect this difference using a two-tailed test and $\alpha = 0.05$. Compare this to a second case where $p_1 = 0.6$ and $p_2 = 0.5$. Here, a sample size of 200 patients per group would now yield only 52% power using the same test, even though $|p_1 - p_2| = 0.1$ in both cases. A more refined measure is therefore required to indicate the detectability of effect sizes.

One popular solution, described by Cohen (2013), involves performing a nonlinear transformation ($\boldsymbol{\phi}$) on the proportions of interest, and calculates the difference in $\boldsymbol{\phi}$ rather than p. The transformation from to p to $\boldsymbol{\phi}$ results in approximately normalized proportions whose standardized effects can be determined by comparing the transformed values. The result of this calculation, denoted by h, serves as a more useful indicator of the detectability of effect sizes, as equal differences in $\boldsymbol{\phi}$ correspond to effect sizes that are equally detectable. Equations for $\boldsymbol{\phi}$ and h are given by:

$$\boldsymbol{\phi} = 2\sin^{-1}\sqrt{p}$$

$$h = |\boldsymbol{\phi_1} - \boldsymbol{\phi_2}|$$

Cohen further proposed guidelines to interpret (in general terms) the magnitude of the effect size given some calculated value of h. These are given below along with a range for the corresponding effect size in terms of the additive difference in proportions ($|p_1 - p_2|$). The difference ranges given below are approximations obtained from selecting a range of p_1 and p_2 values between 0.05 and 0.95. Note that not all possible additive differences are covered in these ranges, and "in-between" effects in terms of Cohen's h effect sizes are possible (e.g., $|p_1 - p_2| = 0.15$ could be considered between a small and medium effect size).

Small effect size: $h = 0.2$ $(0.05 \leq |p_1 - p_2| \leq 0.10)$

Medium effect size: $h = 0.5$ $(0.20 \leq |p_1 - p_2| \leq 0.25)$

Large effect size: $h = 0.8$ $(0.35 \leq |p_1 - p_2| \leq 0.39)$

While the ranges given on the right are a helpful in interpreting effect sizes in terms of additive differences, it is more conventional to use h values to label effect sizes. It should also be noted that h values and their associated descriptions are mainly designed as an aid to broadly interpret effect sizes and should not be recommended over more precise numerical methods pertinent to a specific study. For the current work, Cohen's h serves as reasonable guide in creating an effect size range for the evaluations, as we are considering a variety of different possible 2x2 FRCT cases. The range of effect sizes should therefore be flexible to apply to different scales of measurement, as well as to parameters such as the base rate which may be highly variable from study to study.

For the evaluations with no interaction assumed on the additive scale, additive proportional differences from 0 to 0.25 were considered for both treatments in increments of 0.05. This range roughly covers small and medium effect sizes in terms of Cohen's h. Large effect sizes—which correspond to an additive effect size of about 0.35 or more—were excluded as they could often result in predicting negative proportions in the combined treatment group for small and medium-sized base rates. For example, if both treatments had predicted additive effects of 0.35, the expected proportion in the combined treatment group would be negative for any base rate value lower than 0.7.

For the evaluations with no interaction assumed on either the risk ratio or odds ratio scales, multiplicative effect sizes from 1 (no effect) to 0.5 (50% relative risk reduction) were considered for both treatments in increments of 0.05. Unlike the additive case, there is no possibility of predicting negative proportions in the combined treatment group. For this reason, the chosen range was designed to consider effect sizes up to approximate "large" effects in terms of Cohen's h for base rates up to 0.8. A maximum of $h \approx 0.85$ is attained in the case of a treatment reducing the event rate from 0.8 to 0.4 (relative risk reduction of 0.5, or equivalent OR of 0.167).

2.8 Additive Scale Evaluation

The first set of evaluations were designed to examine power in the case where the treatments do not interact on the additive scale. Example outputs are shown below.

Figure 2.1: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.7$, $N = 400$, $\alpha = 0.05$, Two-Tailed Test)

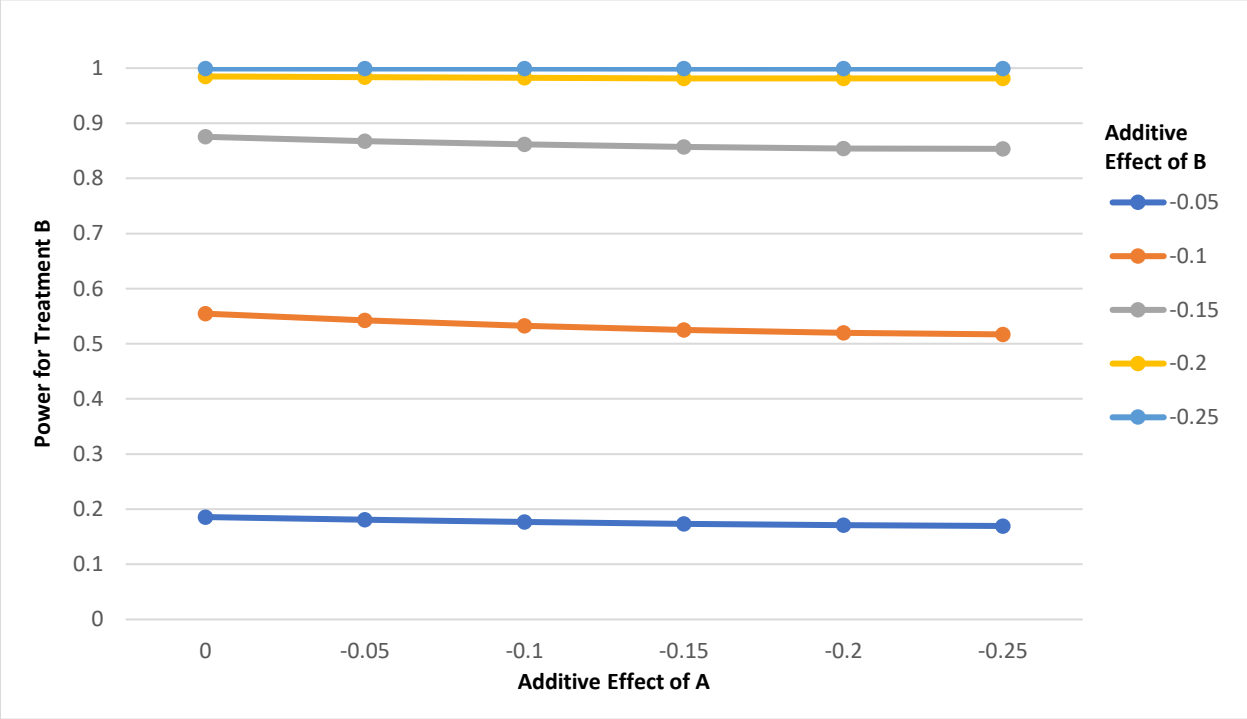


Table 2.1: Associated Power Table for Figure 2.1

Power for Treatment B		Additive Effect of A					
Additive Effect of B		**0**	**-0.05**	**-0.1**	**-0.15**	**-0.2**	**-0.25**
	-0.05	0.186	0.181	0.177	0.173	0.171	0.170
	-0.1	0.555	0.542	0.533	0.525	0.520	0.517
	-0.15	0.875	0.868	0.861	0.857	0.854	0.854
	-0.2	0.985	0.983	0.982	0.982	0.981	0.982
	-0.25	1	1	1	0.999	0.999	0.999

Figure 2.1 depicts the power changes for B as a as a function of the additive effective effect of treatment A under the specified parameter settings. The additive effect of A is incremented in quantities of 0.05 along the x-axis from 0 to -0.25, with the colored lines representing the additive effects of B along the same interval. Table 2.1 shows the numeric values of the power for B at each combination of the treatment effects, rounded to the nearest three decimal places.

From the graph and table, we can see that the power for treatment B decreases as a function of the additive effect of A, though the decreases at each increment tend to be small. As a result, the power for B does not change substantially even if the estimate of A is considerably more than its postulated effect. In summary, when the treatments in a 2x2 FRCT are assumed to have no additive interaction, the resulting estimate following the analysis of the first treatment does not have a strong negative influence on the power for the second treatment, and so it is unlikely that investigators would need to compensate with an increased sample size to maintain power. This is true in cases where the base rate is high as in the current example.

As was described in section 2.4, $k = 1 - 2\,p_{00} - \Delta_A - \Delta_B$ is a solution to the point of zero-change in the power for treatment B. Using Figure 2.1 and Table 2.1 as an example, we will look at the case where $\Delta_A = -0.2$ and $\Delta_B = -0.15$. Given a base rate of $p_{00} = 0.7$, we should expect no change in the power when $k = 1 - 2(0.7) + 0.2 + 0.15 = -0.05$. Therefore, we should expect no change in power for B if the postulated effect of A is $\Delta_A = -0.2$ and an effect of $\Delta_A + k = -0.25$ is observed. Looking at Table 2.1, we can see

that the power for treatment B when $\Delta_A = -0.2$ and $\Delta_B = -0.15$ is 0.854. When $\Delta_A = -0.25$, the power for treatment B remains at 0.854, in accordance with our algebraic result.

Interestingly, in cases of moderate or small base rates, an increased effect of A can result in a slightly increased power for B. As an example, consider a second set of outputs given below:

Figure 2.2: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.5$, $N = 400$, $\alpha = 0.05$, Two-Tailed Test)

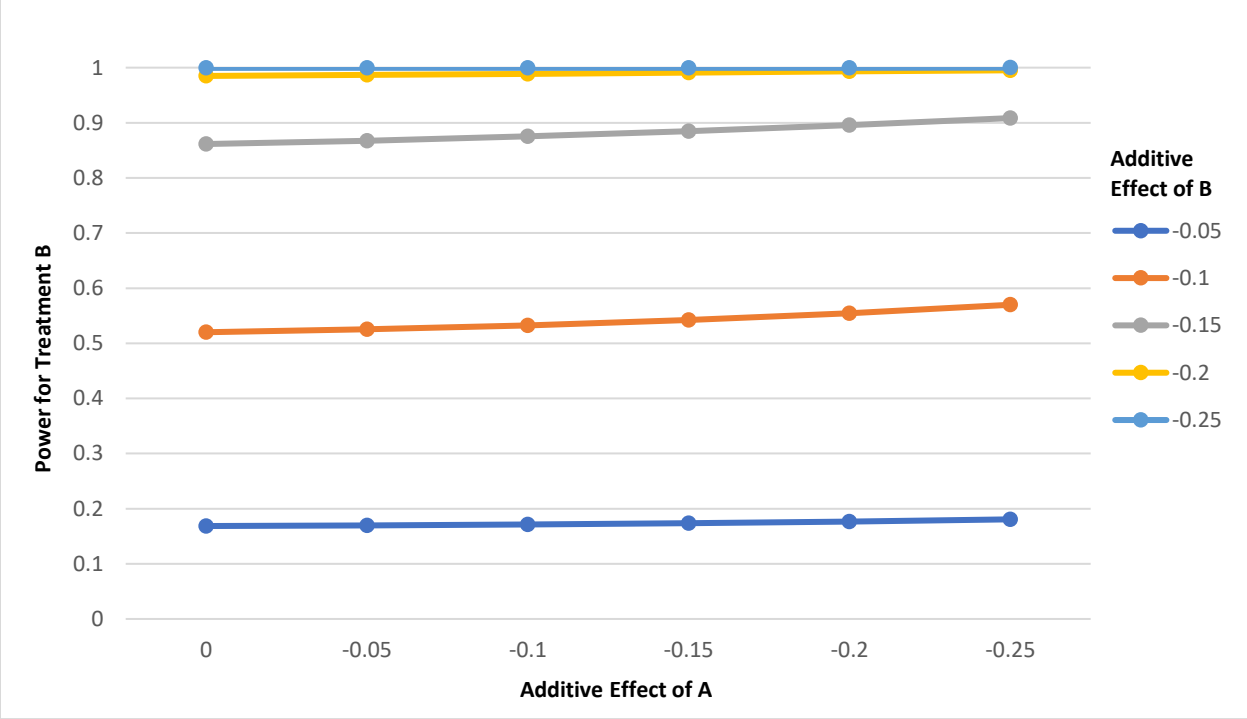


Table 2.2: Associated Power Table for Figure 2.2

Power for Treatment B	Additive Effect of A					
Additive Effect of B	**0**	**-0.05**	**-0.1**	**-0.15**	**-0.2**	**-0.25**
-0.05	0.169	0.170	0.171	0.173	0.177	0.181
-0.1	0.520	0.525	0.533	0.542	0.555	0.570
-0.15	0.861	0.868	0.875	0.885	0.896	0.908
-0.2	0.985	0.987	0.989	0.991	0.993	0.995
-0.25	1	1	1	1	1	1

Figure 2.2 and Table 2.2 are evaluation outputs similar to the previous example with the parameters held constant except for the base rate, which has been reduced from 0.7 to 0.5. From this we can see that an increased effect of A is associated with a slightly increased power for B. In general, when holding all other parameters constant, starting with a high base rate will result in small power decreases for B, which gradually reduce in magnitude and shift to small power increases as the base rate is decremented. At which specific base rate the power relationship shifts from decreases to increases depends on the settings of the other parameters, mainly the sample size, but is generally in the range of 0.5 to 0.6. Figures 2.8-2.14 in the Appendix provide an example that show the gradual shift in the additive power relationships by covering a wide range of base rates. Overall, both power increases and decreases may be observed with the additive scale depending on the parameters, though the power differences in either case tend to be small.

2.9 Risk Ratio Scale Evaluation

The graph and table below depict an example output from an evaluation calculating the power for treatment B, where both treatments are assumed to have no interaction on the risk ratio scale. Here, the parameters are set with a control rate of 0.3, total sample size of 2000 (500 per 2x2 group), $\alpha = 0.05$, and as a two-sided test.

Figure 2.3: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.3$, $N = 2000$, $\alpha = 0.05$, Two-Tailed Test)

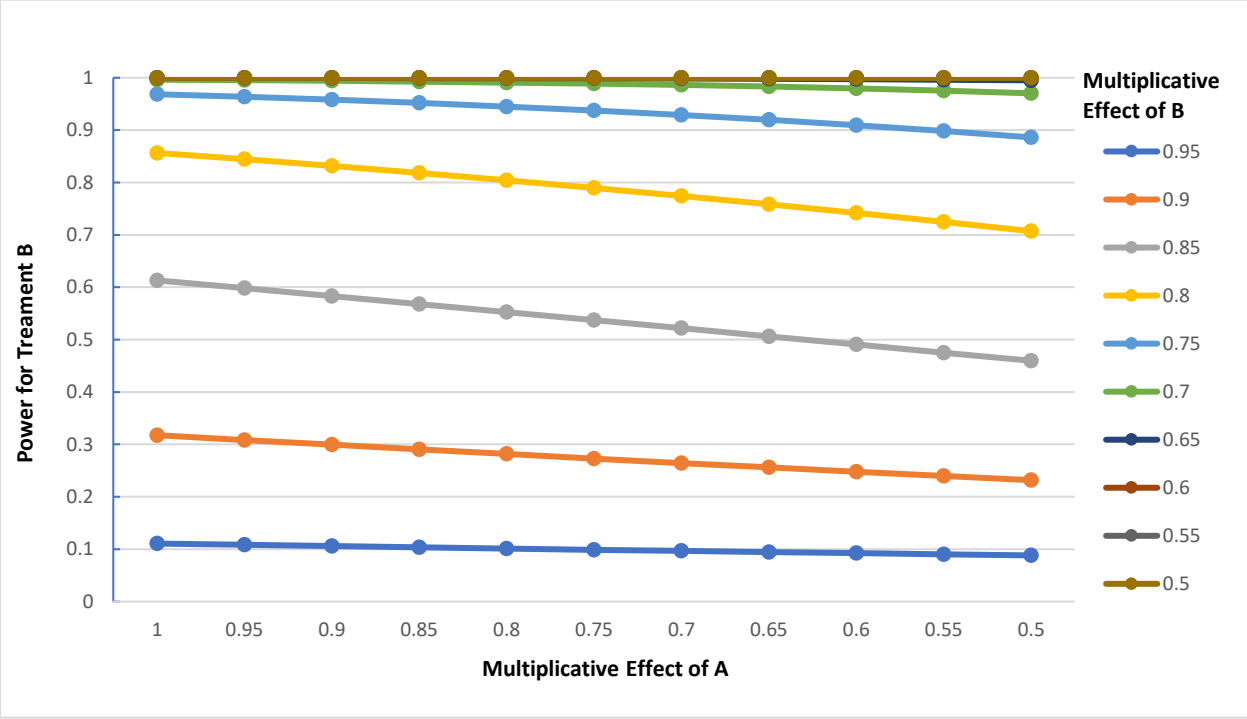


Table 2.3: Associated Power Table for Figure 2.3

Power for Treatment B Multiplicative Effect of B	Multiplicative Effect of A										
	1	**0.95**	**0.9**	**0.85**	**0.8**	**0.75**	**0.7**	**0.65**	**0.6**	**0.55**	**0.5**
0.95	0.111	0.108	0.106	0.104	0.101	0.099	0.097	0.095	0.092	0.090	0.089
0.9	0.318	0.308	0.299	0.291	0.282	0.273	0.265	0.256	0.248	0.240	0.232
0.85	0.613	0.598	0.583	0.568	0.553	0.537	0.522	0.506	0.491	0.475	0.460
0.8	0.856	0.844	0.832	0.818	0.804	0.790	0.774	0.758	0.742	0.725	0.707
0.75	0.968	0.964	0.958	0.952	0.945	0.937	0.929	0.920	0.909	0.898	0.886
0.7	0.996	0.995	0.994	0.993	0.991	0.989	0.986	0.983	0.979	0.975	0.970
0.65	1	1	1	1	1	0.999	0.998	0.998	0.997	0.996	0.995
0.6	1	1	1	1	1	1	1	1	1	1	1
0.55	1	1	1	1	1	1	1	1	1	1	1
0.5	1	1	1	1	1	1	1	1	1	1	1

Figure 2.3 illustrates how the power to detect the effect of treatment B changes as a function of the effectiveness of treatment A. On the x-axis is the multiplicative effect of A ranging between 0.5 and 1, with each of the colored lines representing a different multiplicative effect of treatment B. The y-axis shows the power for treatment B. Together, the points on each line correspond to the power for treatment B at a specific combination of the multiplicative effects of both treatments. Table 2.3 shows the numerical values of the power for B at each combination of the treatment effects, rounded to the nearest three decimal places.

These outputs can then be used to examine the differences in power for treatment B under a variety of effect size possibilities. To illustrate using the example outputs, suppose a 2x2 FRCT is planned with postulated multiplicative effects of 0.8 for both treatments, again with a base rate of 0.3. From the table, we can see that this would result in having a power of 0.804 for treatment B using the example parameters. Suppose treatment A is analysed first and is estimated to have no effect on the base

rate (i.e., a multiplicative effect of 1), and this estimate of treatment A's effect is used instead of its proposed value of 0.8. If treatment B is now analysed using the same parameters, its power would now be 0.856, roughly 6% more than it was prior to the analysis of treatment A. This increase in power corresponds to a 28% relative decrease (from 0.20 to 0.144) in the type II error test for the main effect of treatment B.

Alternatively, suppose that A was found to be greatly more effective than anticipated with an estimated multiplicative effect of 0.5. Now, if B is analysed using the same example parameters, its power drops to 0.707, an absolute decrease of over 9% from the original power. This represents a 46.5% relative increase (from 0.20 to 0.293) in the type II error rate of the test for the main effect of treatment B. Cases such as this where power decreases can occur may be of interest to investigators, as they may then choose to recruit additional participants prior to the analysis of the second treatment to increase the power back to a desirable level.

Continuing with this example, sample size evaluations were implemented to determine the number of additional patients needed to maintain power for the trial. Figure 2.4 shows the minimum required total sample size needed to achieve a power of 0.8 for both treatments given some postulated effects of both treatments. Table 2.4 on the following page displays the associated values of each point on Figure 2.4.

Figure 2.4: Minimum Required Total Sample Size as a Function of the Effectiveness of Treatment A and Treatment B

($P_{00} = 0.3$, Power $= 0.8$, $\alpha = 0.05$, Two-Tailed Test)

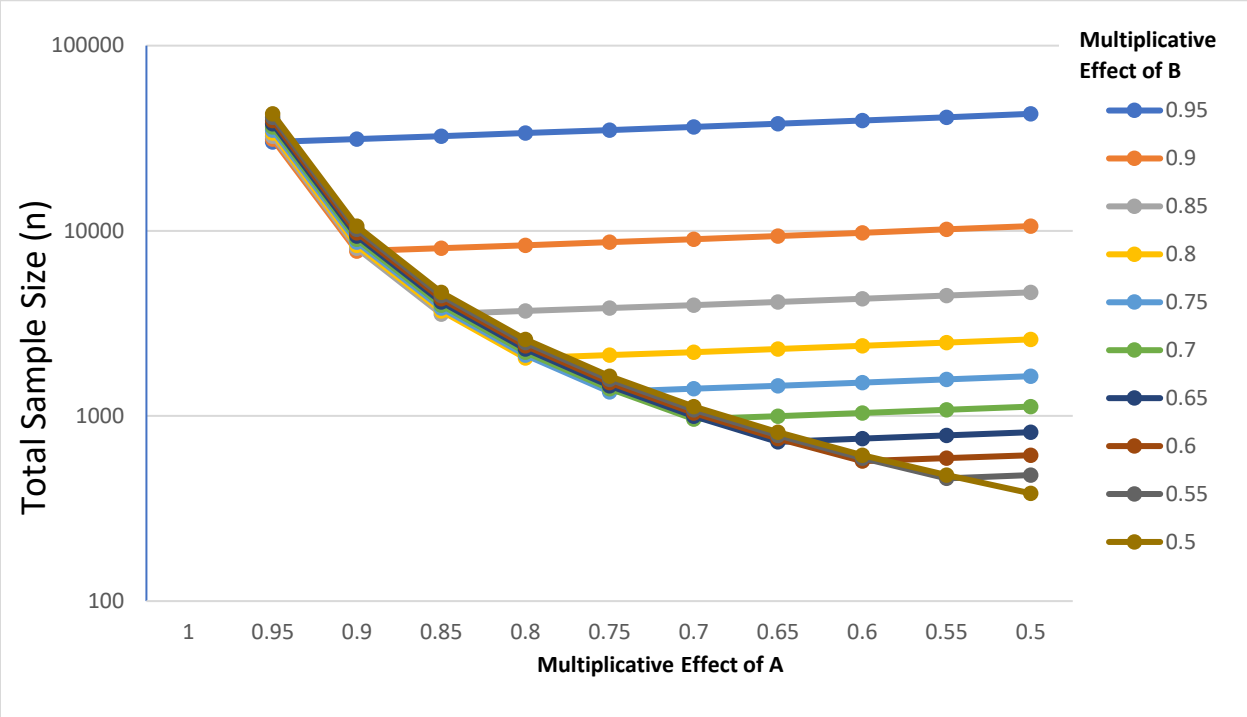


Table 2.4: Associated Sample Size Table for Figure 2.4

Total N		Multiplicative Effect of A										
Multiplicative Effect of B		1	0.95	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.55	0.5
	0.95	#N/A	30196	31304	32472	33706	35012	36392	37856	39412	41068	42836
	0.9	#N/A	31304	7770	8058	8360	8680	9020	9378	9758	10166	10600
	0.85	#N/A	32472	8058	3552	3686	3824	3970	4128	4294	4470	4660
	0.8	#N/A	33706	8360	3686	2054	2130	2212	2300	2390	2488	2592
	0.75	#N/A	35012	8680	3824	2130	1350	1402	1454	1512	1574	1640
	0.7	#N/A	36392	9020	3970	2212	1402	962	998	1038	1080	1124
	0.65	#N/A	37856	9378	4128	2300	1454	998	724	754	784	816
	0.6	#N/A	39412	9758	4294	2390	1512	1038	754	570	592	614
	0.55	#N/A	41068	10166	4470	2488	1574	1080	784	592	460	480
	0.5	#N/A	42836	10600	4660	2592	1640	1124	816	614	480	382

The minimum total sample size at each point is determined by separately calculating the minimum required sample size to detect each treatment main effect (with a power of 0.80) and then taking the maximum of the two. As before, the multiplicative effect of A is depicted on the x-axis with each colored line representing a different multiplicative effect of B. Whereas in Figure 2.3 the y-axis measured the power for B given a total sample size input of 2000, the y-axis in Figure 2.4 measures the total sample size needed given a desired power threshold of 0.80. Table 2.4 gives the values of the continuity-corrected sample sizes needed at each combination of the multiplicative effects of the treatments, rounded up to the nearest integer.

Using these outputs with the example given before (multiplicative effect of 0.80 for both treatments with a base rate of 0.30), a minimum sample size of 2054 patients would be needed to achieve a power of 0.80.

If the analysis for treatment A is conducted first and its multiplicative effect is estimated to be 0.50, we previously noted that this would drop the power for treatment B from 0.80 to 0.717. Using the new estimate of the effect of A, we can determine the sample size needed in the analysis of treatment B to maintain a power of 0.8. Figure 2.5 and Table 2.5 show the minimum required sample size needed to achieve a power of 0.80 for treatment B given some effect combination of the treatments:

Figure 2.5: Minimum Required Sample Size for Treatment B as a Function of the Effectiveness of Treatment A and Treatment B ($P_{00} = 0.3$, Power $= 0.8$, $\alpha = 0.05$, Two-Tailed Test)

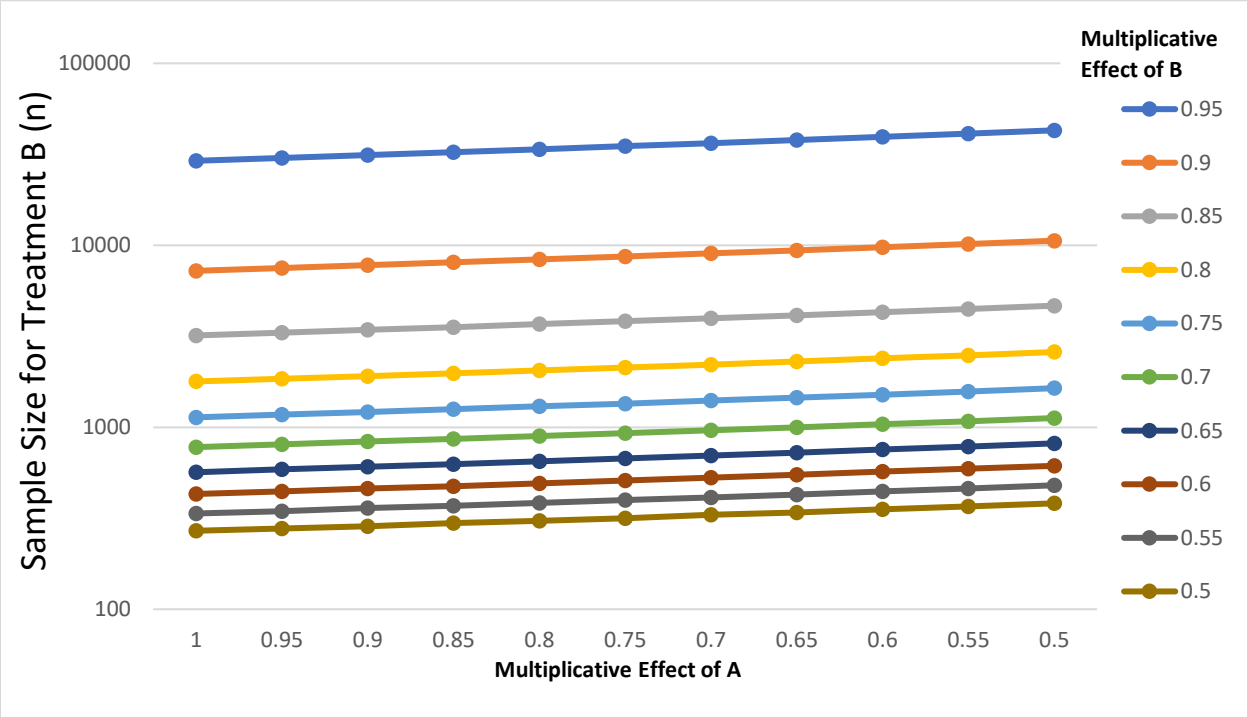


Table 2.5: Associated Sample Size Table for Figure 2.5

Sample Size for Treatment B	Multiplicative Effect of A										
Multiplicative Effect of B	**1**	**0.95**	**0.9**	**0.85**	**0.8**	**0.75**	**0.7**	**0.65**	**0.6**	**0.55**	**0.5**
0.95	29142	30196	31304	32472	33706	35012	36392	37856	39412	41068	42836
0.9	7242	7500	7770	8058	8360	8680	9020	9378	9758	10166	10600
0.85	3198	3310	3428	3552	3686	3824	3970	4128	4294	4470	4660
0.8	1786	1846	1912	1980	2054	2130	2212	2300	2390	2488	2592
0.75	1134	1172	1212	1256	1302	1350	1402	1454	1512	1574	1640
0.7	778	806	834	864	894	928	962	998	1038	1080	1124
0.65	566	588	606	628	650	674	700	724	754	784	816
0.6	430	444	460	474	492	510	528	548	570	592	614
0.55	336	346	360	370	384	398	412	426	444	460	480
0.5	270	278	286	298	306	316	330	340	354	368	382

From the output, we can see that if treatments A and B have multiplicative effects of 0.50 and 0.80, respectively, a total sample size of 2592 would be needed to achieve a power of 0.80 for treatment B. This is 528 patients more than the 2054 needed originally when using the postulated effect value of 0.8 for treatment A, which can be expressed as a roughly 26% increase in the total sample size.

In terms of general observations, a larger effectiveness of A always results in a decreased power for B when the treatments do not interact on the risk ratio scale, holding all other parameters constant. Broadly speaking, if the estimate of A's treatment effect is within 10% of its postulated value, the difference it makes on the power for B tends to be small (< 5%), regardless of the base rate. Large drops in power are generally observed only when the effect of A is estimated to be far larger than anticipated, in which case it may be beneficial to consider recruiting additional patients prior to the analysis of B. An interesting note is that greatest overall power drop is observed for some intermediate value of the effect of treatment B, rather than either the largest or smallest values. From the example given in Figure 2.3, the largest power drop was observed when the multiplicative effect of B was 0.85 (power = 0.631 when the effect of A is 1 vs. power = 0.460 when the effect of A is 0.5, for an overall power drop of 0.153). For which value of B's effect the greatest power drop is observed varies depending on the parameter settings.

2.10 Odds Ratio Scale Evaluation

The final set of evaluations were designed to examine power in the case where the treatments do not interact on the odds ratio scale. Example outputs are provided below using the same parameter settings as in the risk ratio case.

Figure 2.6: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.3$, $N = 2000$, $\alpha = 0.05$, Two-Tailed Test)

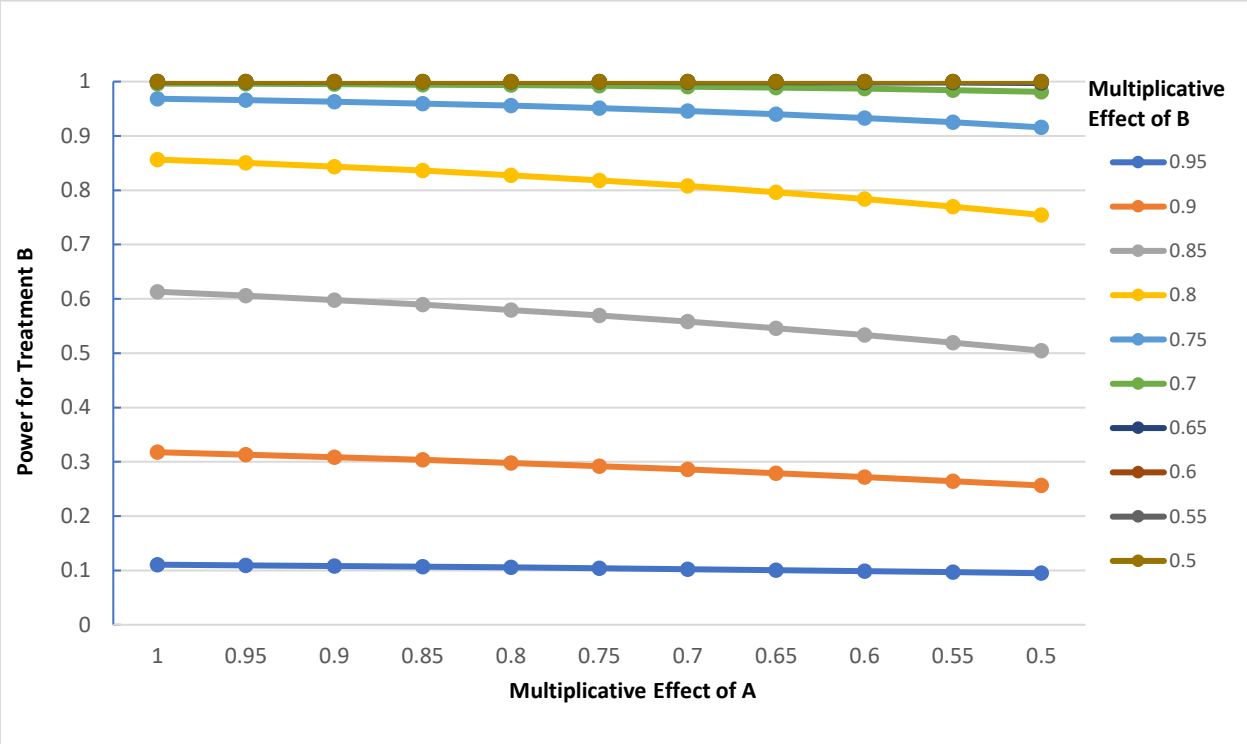


Table 2.6: Associated Power Table for Figure 2.6

Power for Treatment B Multiplicative Effect of B	Multiplicative Effect of A										
	1	**0.95**	**0.9**	**0.85**	**0.8**	**0.75**	**0.7**	**0.65**	**0.6**	**0.55**	**0.5**
0.95	0.111	0.110	0.108	0.107	0.106	0.104	0.103	0.101	0.099	0.097	0.095
0.9	0.318	0.313	0.309	0.304	0.298	0.292	0.286	0.279	0.272	0.264	0.257
0.85	0.613	0.606	0.598	0.589	0.580	0.569	0.558	0.546	0.533	0.519	0.505
0.8	0.856	0.850	0.844	0.836	0.828	0.818	0.808	0.796	0.784	0.770	0.754
0.75	0.968	0.966	0.963	0.960	0.956	0.951	0.946	0.940	0.933	0.925	0.916
0.7	0.996	0.996	0.995	0.994	0.993	0.992	0.990	0.989	0.987	0.984	0.981
0.65	1	1	1	1	1	1	1	1	0.999	0.998	0.997
0.6	1	1	1	1	1	1	1	1	1	1	1
0.55	1	1	1	1	1	1	1	1	1	1	1
0.5	1	1	1	1	1	1	1	1	1	1	1

Analogous to before, the graph (Figure 2.6) and associated table (Table 2.6) show the power for the effect of treatment B over a large range of treatment effect combinations. Comparing these outputs to Figure 2.3 and Table 2.3, we see that the calculated powers in each table are close numerically, and the power curves in both figures exhibit the same general shape. In both cases we observe an overall decrease in the power for B as the effectiveness of A increases. Holding the parameters constant, the power relationships observed in the relative risk and odds ratio evaluations tend to be more similar the closer the base rate is to zero.

However, there is a caveat as the base rate becomes larger ($P_{00} > 0.6$) in the odds ratio case. As the base rate increases beyond this point, the power functions which were elsewhere strictly decreasing for lower base rates begin to show slight concavity for some effect sizes of B. The degree of the concavity varies depending on

the other parameter settings. Figure 2.7 and Table 2.7 provide an evaluation output demonstrating this relationship.

Figure 2.7: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.7$, $N = 600$, $\alpha = 0.05$, Two-Tailed Test)

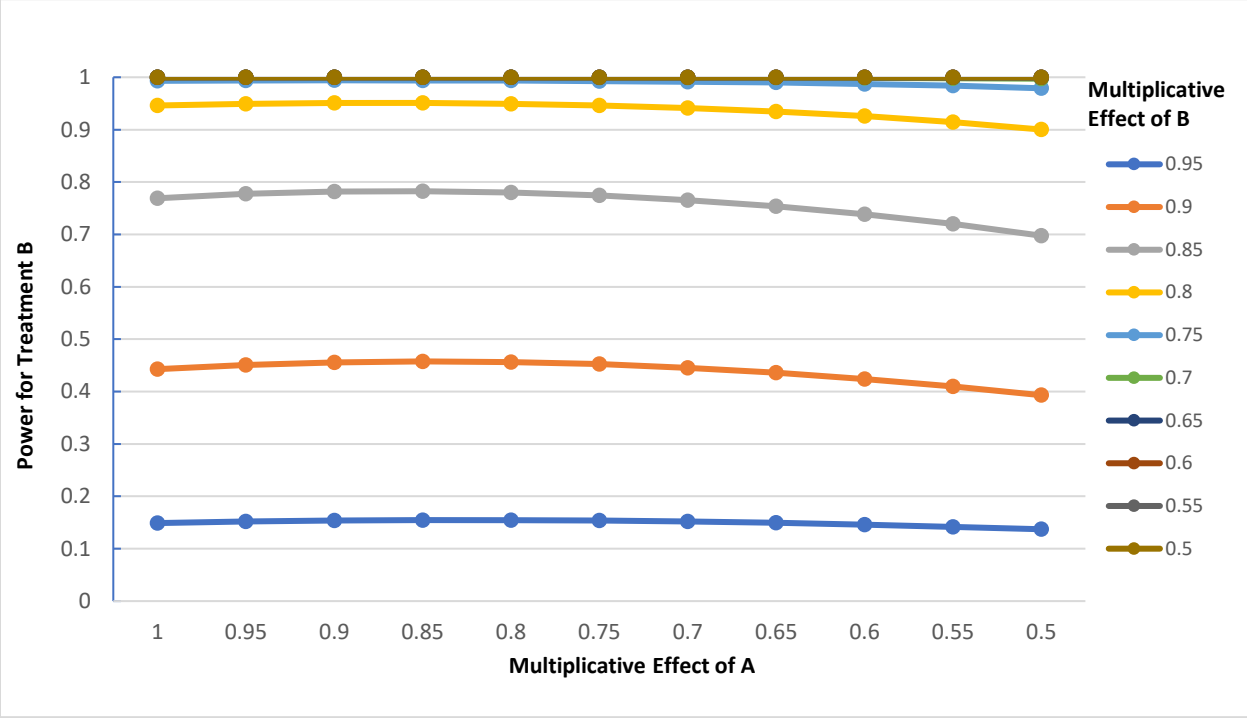


Table 2.7: Associated Power Table for Figure 2.7

Power for Treatment B Multiplicative Effect of B	Multiplicative Effect of A										
	1	**0.95**	**0.9**	**0.85**	**0.8**	**0.75**	**0.7**	**0.65**	**0.6**	**0.55**	**0.5**
0.95	0.149	0.152	0.154	0.155	0.154	0.154	0.152	0.149	0.146	0.142	0.137
0.9	0.443	0.451	0.456	0.458	0.456	0.452	0.445	0.436	0.424	0.410	0.393
0.85	0.769	0.777	0.782	0.782	0.780	0.774	0.766	0.754	0.738	0.720	0.698
0.8	0.946	0.949	0.951	0.951	0.949	0.946	0.941	0.935	0.926	0.915	0.900
0.75	0.993	0.994	0.994	0.994	0.993	0.993	0.992	0.990	0.987	0.984	0.979
0.7	1	1	1	1	1	0.999	0.999	0.999	0.999	0.998	0.997
0.65	1	1	1	1	1	1	1	1	1	1	1
0.6	1	1	1	1	1	1	1	1	1	1	1
0.55	1	1	1	1	1	1	1	1	1	1	1
0.5	1	1	1	1	1	1	1	1	1	1	1

In Section 2.4, we posited that $kOR_{\mathrm{A}} = \frac{(1-p_{00})}{p_{00}\sqrt{OR_{\mathrm{B}}}}$ should yield an approximate solution in calculating the effect of A which maximizes the power for treatment B. Using Figure 2.7 and Table 2.7 as an example, suppose we are interested in estimating the effect of treatment A such that the power for treatment B is maximized when the multiplicative effect of treatment B is 0.85 (corresponding to the grey curve in Figure 2.7). Given that $P_{00} = 0.7$ and treatment B reduces the event rate to $(0.7)(0.85) = 0.595$, we can easily calculate that $OR_{\mathrm{B}} = \frac{(0.595)/(1-0.595)}{(0.7)/(1-0.7)} = 0.63$. From here, we calculate that the effect of treatment A to maximize the power for treatment B is $kOR_{\mathrm{A}} = \frac{(1-0.7)}{0.7\sqrt{(0.63)}} = 0.54$.

$kOR_{\mathrm{A}} = 0.54$ corresponds to treatment A reducing the event rate from 0.7 to 0.56, or a multiplicative effect of approximately 0.8. Looking at Figure 2.7 and Table 2.7, we see that the power for B is approximately 0.780 when the multiplicative effect of A is 0.8. This is reasonably close to the true maximum power for B, which was verified by software to be 0.783 and occurs when the multiplicative effect of treatment A is 0.863 (OR = 0.66). In summary, $kOR_{\mathrm{A}} = \frac{(1-p_{00})}{p_{00}\sqrt{OR_{\mathrm{B}}}}$ can be used to approximate the effect of A such that the power for treatment B is maximized, though these approximations do have a noticeable degree of error.

Examining the general power trends, we observe that slight power increases for B are observed when the effect of A is smaller (from approximately 1 to 0.8). As the effectiveness of A increases past this point, these power increases level off and begin to transition into decreases. Compared to the risk ratio case, the overall drop in power across the range of the effectiveness of A tends to be substantially smaller when the

parameters are held constant at a high base rate. For example, from Table 2.5, the largest overall drop in power is seen when the effect of B is 0.85 (a raw difference of about 0.07 when the effect of A is 1 versus 0.5). Using the same parameters in the risk ratio evaluation, the largest drop in power is also when the effect of B is 0.85, but with a much larger absolute decrease (a difference of about 0.26 when the effect of A is 1 versus 0.5). Overall, the risk ratio and odds ratio evaluations produce similar results when the base rate is small, which might be expected given the close numeric similarity between risk ratios and odds ratios calculated at small base rates. However, we observe a clear divergence in the evaluations between the two scales when the base rate is high.

Chapter 3

Applications Using Example Data

In this section, two examples are provided using real data FRCT data to show how the power (and associated minimum sample size), needed to detect a treatment effect may change depending on the effectiveness of the other treatment in the trial.

3.1 Illustration Using Example Data (Heyland et al., 2013)

In the following example based on a study by Heyland et al. (2013), we examine how the observed effect of the first treatment in a 2x2 FRCT may influence the analysis of the second treatment and its associated power. In this trial, critically ill patients experiencing multiorgan failure were randomly assigned to receive supplements of antioxidants, glutamine, both, or neither (placebo), with 28-day mortality measured as the primary outcome. The study was planned to detect a 25% relative risk reduction from 0.30 to 0.225 for both treatments (RR = 0.75, OR = 0.677), with a calculated total sample size of 1200 to obtain 80% power for each treatment in the trial. The study employed two-tailed testing with $\alpha = 0.05$. The authors were interested in assessing the main effects of each treatment and did not anticipate an interaction on the odds ratio scale. Assuming these conditions are met, the anticipated 2x2 table for the trial is:

Table 3.1: Anticipated 2x2 Table (Heyland et al.)

		Antioxidants		
		Present	Absent	Marginal Event Rate (Glutamine)
Glutamine	Present	0.164	0.225	0.195
	Absent	0.225	0.300	0.263
	Marginal Event Rate (Antioxidants)	0.195	0.263	

* Cells indicate the expected event rate of patient death in each of the four treatment groups.

A total of 1223 patients were enrolled and randomized to each of the four treatment groups after being assessed for eligibility, with 1218 patients included in the primary analysis. The table below shows the results with the observed numbers of patient deaths out of the total patient count per group with the associated group proportions.

Table 3.2: Resulting 2x2 Table (Heyland et al.)

		Antioxidants		
		Present	Absent	Marginal Event Rate (Glutamine)
Glutamine	Present	101/310 (0.326)	97/301 (0.322)	194/611 (0.318)
	Absent	89/307 (0.290)	76/300 (0.253)	165/607 (0.272)
	Marginal Event Rate (Antioxidants)	190/617 (0.308)	173/601(0.288)	

* Cells indicate the event rate of patient death in each of the four treatment groups.

After determining that there was no significant interaction between the two treatments on the odds ratio scale ($p = 0.49$), the investigators assessed the main effects of each treatment. The authors reported no significant difference in mortality among patients who received antioxidants versus no antioxidants (0.308 versus 0.288, OR = 1.09, $p = 0.48$) and a tendency of increased mortality for patients who received glutamine versus no glutamine (0.324 versus 0.272, OR = 1.28, $p = 0.05$).

Using the data from this study, we explore the hypothetical scenario in which the main effect of one treatment is analysed first and assess the resulting impact on the power for the other treatment in the trial. In accordance with the original trial, we assume no interaction on the odds ratio scale. Suppose the analysis for the main effect of antioxidants had been conducted first, with the investigators still blinded to the effect of glutamine. Here, the authors failed to find evidence of a non-zero effect of antioxidants ($p = 0.48$). If the investigators update their estimate of the effect of antioxidants to zero based on this test, the anticipated marginal effect of glutamine and its associated power will change as a result. Originally, a 25% risk reduction from 0.30 to 0.225 was anticipated for both antioxidants and glutamine, with no interaction between the two treatments. Table 3.1 depicts the proportions anticipated in each treatment group under these assumptions. In this case, marginal proportions of 0.195 and 0.263 would be expected for the glutamine-present versus glutamine-absent groups, respectively. As determined by the investigators, a total sample size of approximately 1200 patients would be needed to provide 80% power to detect this effect.

In the new hypothetical case, the main effect of antioxidants is analyzed and found to have no effect, with a subsequent analysis planned to assess the main effect of glutamine. If the investigators maintain the originally postulated value for glutamine's effect and assumption of no interaction between the two treatments, the anticipated 2x2 table becomes:

Table 3.3: Updated 2x2 Table (Heyland et al.)

		Antioxidants		
		Present	Absent	Marginal Event Rate (Glutamine)
Glutamine	Present	0.225	0.225	0.225
	Absent	0.300	0.300	0.300
	Marginal Event Rate (Antioxidants)	0.263	0.263	

* Cells indicate the event rate of patient death in each of the four treatment groups.

Here, marginal event rates of 0.225 and 0.30 would be expected for the glutamine-present and glutamine-absent groups, respectively. If the marginal analysis for glutamine is conducted maintaining the originally planned sample size of 1200, the power to detect the main effect for glutamine increases from 0.80 to approximately 0.84. This change represents a 20% relative decrease in the type II error rate of the test (from 0.20 to 0.16), reducing the chance of investigators failing to find a significant main effect for glutamine. As we have shown through this example, the anticipated marginal estimates of the glutamine-present and glutamine-absent groups are changed by the

updated information on the effect of antioxidants, consequently improving the power to detect the effect of the glutamine treatment.

3.2 A Second Example (Poldermans et al., 1999)

As previously illustrated, an increase in power for a treatment (glutamine) occurred when the observed effect size of the other treatment (antioxidants) was found to be smaller than the planned estimate. In the second example, based on a study by Poldermans et al. (1999), we examine a counterexample in which a treatment's power may decrease if the observed effect size for the other treatment is found to be larger than the planned estimate.

In this study, patients at high risk of cardiac complications and who were undergoing major vascular surgery were randomized to receive a standard care regimen with or without bisoprolol, a beta blocker medication used to reduce the risk of adverse cardiac events. The primary outcome measure was the overall proportion of patients in each group who either died of a cardiac event or had a non-fatal myocardial infarction. The study was planned to detect a 50% relative risk reduction (or equivalently, OR = 0.41) in the primary outcome from 0.30 to 0.15 at 80% power using a two-tailed test ($\alpha = 0.05$). An interim analysis was conducted following the enrolment of the first 100 patients and a stop for safety was planned if a significant difference ($p = 0.001$) in the primary endpoint was found between the bisoprolol and control groups.

The results of the interim analysis are shown in the table below, with each cell indicating the number and proportion of patients in each group who exhibited the primary outcome (non-fatal myocardial infarction or death):

Table 3.4: Results Table (Poldermans et al.)

Bisoprolol Group (N = 59)	Standard Care Group (N = 53)
2/59 (0.034)	18/53 (0.34)

* Cells indicate the event rate of non-fatal myocardial infarction or death.

From the interim results, a significant difference was found in the primary outcome between the bisoprolol and standard care groups (0.034 versus 0.340, $p < 0.001$), leading the investigators to stop the study following the interim analysis.

Using this data, consider a scenario in which a hypothetical second treatment, which we will denote as treatment B, had been administered along with bisoprolol in a 2x2 FRCT, such that participants received either bisoprolol, treatment B, both, or neither. Further suppose the investigators had planned for a 50% relative risk reduction in the primary outcome for both bisoprolol and treatment B, with no interaction between the treatments on the odds ratio scale. Under these assumptions, the anticipated 2x2 table for the trial is:

Table 3.5: Anticipated 2x2 Table (Poldermans et al.)

		Treatment B		
		Present	Absent	Marginal Event Rate (Bisoprolol)
Bisoprolol	Present	0.068	0.150	0.109
	Absent	0.150	0.300	0.225
	Marginal Event Rate (Treatment B)	0.109	0.225	

* Cells indicate the event rate of non-fatal myocardial infarction or death.

From this table, marginal event rates of 0.109 and 0.225 would be expected for the treatment-present and treatment-absent groups for B, respectively. In order to obtain a power of 0.80 for the main effect treatment B (using a two-tailed test and $\alpha = 0.05$), a minimum sample size of approximately 324 participants would be required.

Next, suppose that the main effect of bisoprolol is analysed first, and is observed to reduce the rate of the primary outcome from 0.30 to 0.10, an effect that would be considered significant ($p \approx 0.02$) but not enough to stop the study for benefit. Analogous to the previous example, this new information on the effect of the first analysed factor (bisoprolol) may be used to re-compute the power for the second factor (treatment B). Updating the 2x2 table using the observed estimate of bisoprolol's effect, while maintaining the assumption of no interaction, we get:

Table 3.6: Updated 2x2 Table (Poldermans et al.)

		Treatment B		
		Present	Absent	Marginal Event Rate (Bisoprolol)
Bisoprolol	Present	0.044	0.100	0.072
	Absent	0.150	0.300	0.225
	Marginal Event Rate (Treatment B)	0.097	0.200	

* Cells indicate the event rate of non-fatal myocardial infarction or death.

From the updated table, the marginal estimates for treatment B become 0.097 in the B-present group and 0.2 in the B-absent group. If the analysis of B was then carried out using the originally planned sample size of 324 participants, the power for the main effect of B drops to approximately 0.74, a relative 7.8% drop from the original power of 0.8. This also corresponds to a 30% relative increase in the type II error rate (from 0.2 to 0.26). An additional 48 patients would be required in the analysis of treatment B to maintain a power of 0.8, an approximate 15% increase from the originally planned sample size of 324.

Chapter 4

Discussion

4.1 Summary of Findings

The current study examined how the power for a treatment in a binary outcome 2x2 FRCT may change depending on the effect estimate of the other treatment in the trial. Numerical evaluations were implemented to examine how power relationships change under a variety of parameter settings, mainly by modifying the base rate and sample size. Separate evaluations examined these relationships when interactions were either defined on either the risk ratio, odds ratio, or additive scales. The main findings for each scale are summarized in Table 4.1 on the following page. Note that what is considered “larger” and “smaller” base rates depends on the parameter settings, but in general tends to be approximately 0.6.

Table 4.1: Summary of Main Findings

Scale	Base Rate	Estimate for first treatment greater than postulated value?	Effect on Power for second treatment
Risk Ratio	Smaller	Yes	Decrease
	Smaller	No	Increase
	Larger	Yes	Decrease
	Larger	No	Increase
Odds Ratio	Smaller	Yes	Decrease
	Smaller	No	Increase
	Larger	Yes	May Increase or Decrease
	Larger	No	May Increase or Decrease
Additive	Smaller	Yes	Increase
	Smaller	No	Decrease
	Larger	Yes	Decrease
	Larger	No	Increase

When treatments do not interact on the risk ratio scale, a treatment estimate that is greater than its postulated value always leads to a power decrease for the other treatment, regardless of the other parameters. Generally, this drop in power is between 1-3% for every additional increment of 0.05 greater the first treatment's estimated multiplicative effect is over its postulated value. Conversely, when the first treatment estimate is smaller than its postulated value, a 1-3% increase in power is observed for

each 0.05 increment greater the postulated multiplicative effect is over the estimate. At high base rates, large power drops (e.g., 40%) are theoretically possible, but only in extreme cases where the effect estimate of the first treatment is far greater than its assumed value (e.g., an estimated effect of 0.5 versus a postulated effect of 0.05). We also observe that the magnitude of the power decrease is greatest for some intermediate value of the second treatment's postulated effect. For which value the largest power drop is observed varies depending on the parameter settings.

When the treatments do not interact on the odds ratio scale, the power relationship closely approximates that observed in the risk ratio case when the base rate is small. That is, estimates for the first treatment effect greater than its assumed value results in a power decrease for the second treatment effect, while estimates smaller than the assumed value result in a power increase. This result is expected, as odds ratios and risk ratios closely approximate one another when the base rate is small. However, at larger base rates (> 0.5), the odds ratio power curves begin to show increased concavity. When this occurs, greater estimates for the first treatment effect result in slight power increases for the second treatment, which eventually shift to power decreases as the first treatment estimate approaches extreme values. Across the cases considered, the change in power when using the odds ratio scale tends to be small to moderate, with power differences of at most 15% even for extreme first treatment effect estimates, regardless of sample size or base rate. As before, the greatest change in power tends to be for some intermediate value of the postulated effect of the second treatment which varies based on the parameter settings.

Finally, in the case where the treatments do not interact on the additive scale, the power relationship may either show increases or decreases depending on the base rate. As the base rate approaches one, additive effect estimates for the first treatment that are greater than its assumed value result in power decreases for the second treatment, with power increases observed when the estimate is less than the assumed value. However, as the base rate approaches smaller values an opposite pattern is observed. Additive effect estimates for the first treatment that are greater than its assumed value now results in power increases for the second treatment, while estimates less than the assumed value result in power decreases. At which base rate this change occurs depends on the parameter settings but is generally between base rates of 0.5 and 0.6. Across a large range of parameter combinations, the change in power (either increase or decrease) tends to be less than 5% if the additive effect estimate of the first treatment is within a 0.05 increment of its assumed value. In extreme cases where the estimate of the first treatment is much greater than the assumed value (e.g., a difference of 0.2), power changes of 10-15% are possible but less likely to come up in practical studies.

4.2 Implications and Broader Connections to Clinical Trials

In examining the evaluation results across different scales of measurement and parameter settings, we find that the estimate of the first treatment effect can have a noticeable impact on the power for the other treatment in the trial. Although large changes in power for the second treatment are possible, only small to moderate changes are observed when the estimate of the first treatment effect is close to its postulated value. In practice, effect estimates larger than the planned value are not common (Zakeri et al, 2018), and when they occur investigators may invoke stopping rules if the treatment is substantially more beneficial than anticipated. There is contention over when it is a correct decision for investigators to stop randomization to a treatment due to benefit (Montori et al., 2005, Pocock, 2005, Walter et al., 2019), as trials that are stopped early for benefit tend to overestimate the treatment benefit observed in the interim analysis, particularly in trials with small sample sizes. Regardless of the justification, stopping an FRCT for benefit will nonetheless affect the power relationships discussed here.

However, as has been noted in the examples throughout, even a moderate change in power for a treatment may still represent a considerable change in the relative type II error rate when analyzing the effect (e.g., a 10% drop in power from 0.8 to 0.7 represents a 50% increase in the type II error rate, from 0.2 to 0.3). Investigators conducting binary outcome FRCTs should be aware of the possibility of the first treatment analysis affecting the power for the other treatment in the trial, with particular

vigilance paid to cases in which the first treatment effect estimate differs greatly from its proposed value.

These results also suggest that it may not be ideal to analyse both treatments simultaneously, even in situations where that is possible. Analyzing the effect of one treatment first, while still blinded to the effect of the second, affords investigators the opportunity to adjust the sample size prior to the second treatment analysis if a decrease in power has occurred. In other words, investigators gain the ability to adapt the study accordingly to prevent a loss in power.

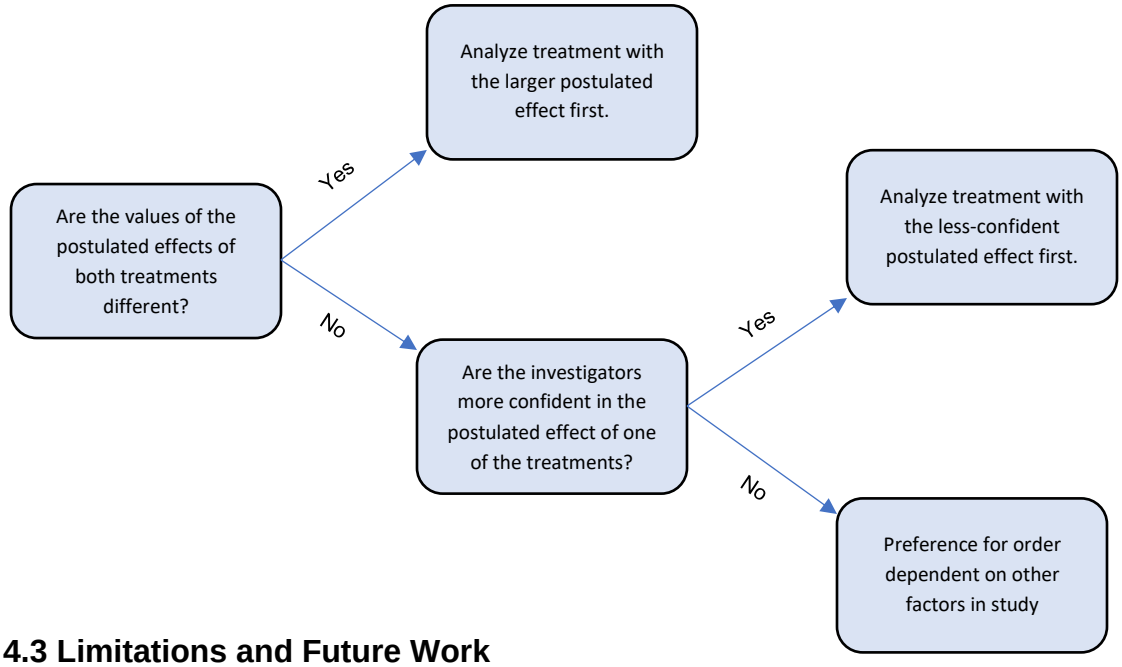
Additional considerations relate to which treatment effect should be analysed first in cases where the option is available. If the postulated effect for one of the treatments is smaller than the other, it is arguably preferable to conduct the first analysis for the treatment with the larger postulated effect. The justification for this stems from the fact that the sufficient total sample size for the trial is normally calculated based on the maximum of the minimum sufficient sample sizes needed to detect each of the two postulated effects. Using this method will result in having power for the smaller effect at the desired threshold level and more-than-sufficient power for the larger effect. For this reason, the power for the treatment with the smaller effect is at greater risk of falling below the desired target power depending on the effect estimate of the other treatment. It may therefore be beneficial to analyse the treatment with the larger postulated effect first, knowing that a sample size adjustment may have to be made to maintain power for the second (lower power) treatment.

In the case where the postulated effects of both treatments are the same, it may instead be preferable to first analyse the treatment for which the investigators are less confident in the assumed effect. This might be the case, for example, in a trial where a new treatment is being administered alongside an established treatment for which estimates of its effect have been obtained from previous data. Since it can be reasonable to assume that the treatment with the less established effect is likelier to produce an effect estimate different from its postulated value, investigators may be concerned that the less predictable effect of the new treatment may influence the power for the second (more predictable) treatment. By analyzing the new treatment first, investigators are better positioned to adapt the sample size for the second treatment analysis should the first analysis yield an unanticipated estimate that results in decreased power for the second treatment. Alternatively, if the first treatment estimate results in increased power for the second treatment, investigators then benefit by having a reduced type II error rate for the second treatment analysis. In either case, analyzing the treatment with the less predictable effect is beneficial for the second treatment analysis. the treatment whose effect is less predictable, assuming that the postulated effects of both treatments are the same.

Figure 4.1 outlines the previously described recommendations for which treatment to analyze first in a 2x2 binary FRCT, assuming investigators do not have a prior preference. Under these assumptions, these recommendations may be used to aid in designing the study protocol. As a final note, the preferred choice of which treatment

effect to analyse first may depend on other factors specific to the design of a trial and is left to the judgement of the investigators in these cases.

Figure 4.1: Recommendations for Order of Treatment Analyses



4.3 Limitations and Future Work

We conclude by considering study limitations and directions for future study. One key limitation of the current work is that we exclusively examine 2x2 FRCTs with binary outcomes. The same issues discussed here have yet to be examined in FRCTs using continuous outcomes and remain left open to future investigation. Examining power changes with continuous measures adds additional dimensions to the issue, as considerations of different treatment variances, the possibility of non-normal effects, and potential heterogeneity in the variances of the different treatment effects in the trial are all factors that may have a significant contributing effect to the power relationship between treatments.

A second limitation is that we exclusively consider scenarios in which the entire available sample is used to determine the effect estimate of the first treatment. We do not consider cases where this estimate comes from an interim analysis where some subset of the sample is used, which could alter the impact on the marginal estimates for the second treatment. For example, If the first treatment effect estimate is obtained using a subset of the sample and subsequent decision is made to stop randomization due to benefit, the change in the marginal effect estimate of the second treatment will now be weighted, based on how many patients were included in the first analysis. It is also possible that an interim analysis on the first treatment will yield more extreme estimates of the first treatment effect, as the variance of the estimator will be larger as a result of using a smaller sample. This in turn can increase the impact on the marginal estimates for the second treatment. Overall, the potential for the first treatment estimate to be obtained from an interim analysis provides an additional layer of complexity that is not discussed here, and remains an avenue for future research in this area.

In conclusion, the numeric evaluations implemented and discussed throughout are intended to be a framework for investigators to understand how the power for a treatment in a 2x2 binary outcome FRCT is influenced by the effect estimate of the other treatment in the trial. These evaluations are constructed under ideal trial assumptions, including that there is no interaction between the two treatments of interest on some scale of measurement. In the future, the examination other types of FRCTs may provide a more complete picture of how a treatment's power may change

under other experimental scenarios, and can help develop a more thorough understanding of how investigators may adapt a study accordingly.

Chapter 5

Appendix

Figure 2.8: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.8$, $N = 400$, $\alpha = 0.05$, Two-Sided Test)

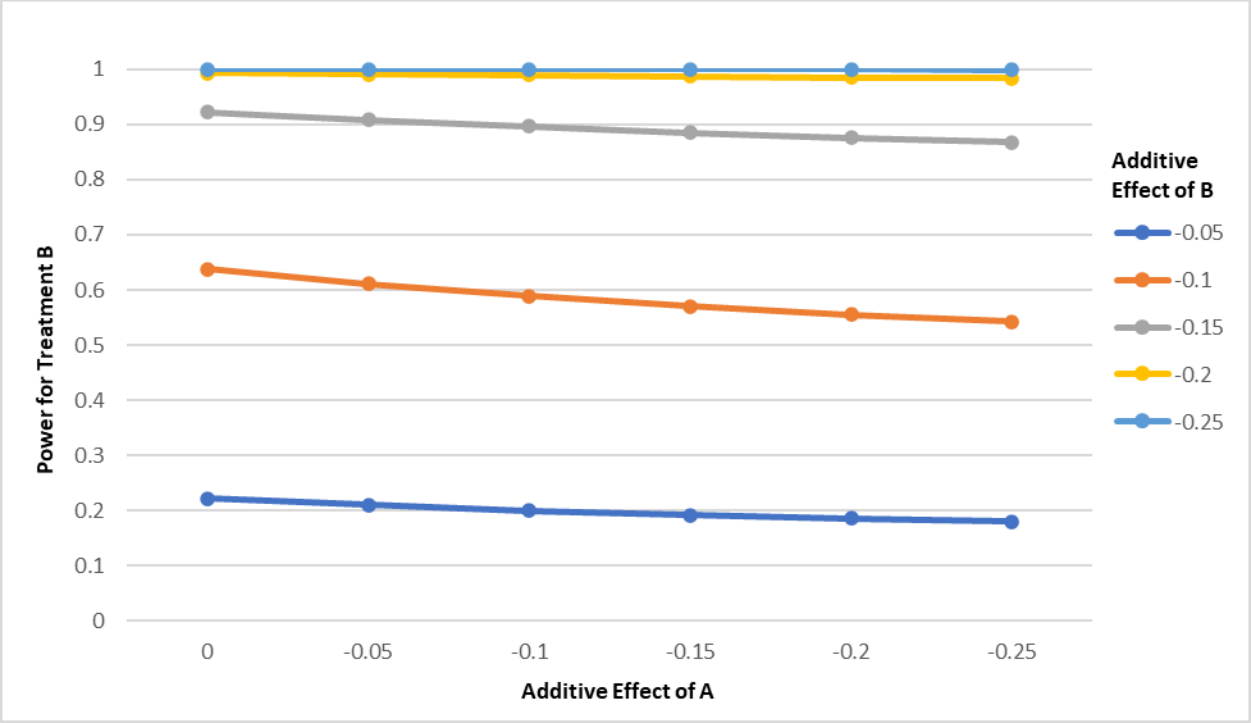


Figure 2.9: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.75$, $N = 400$, $\alpha = 0.05$, Two-Sided Test)

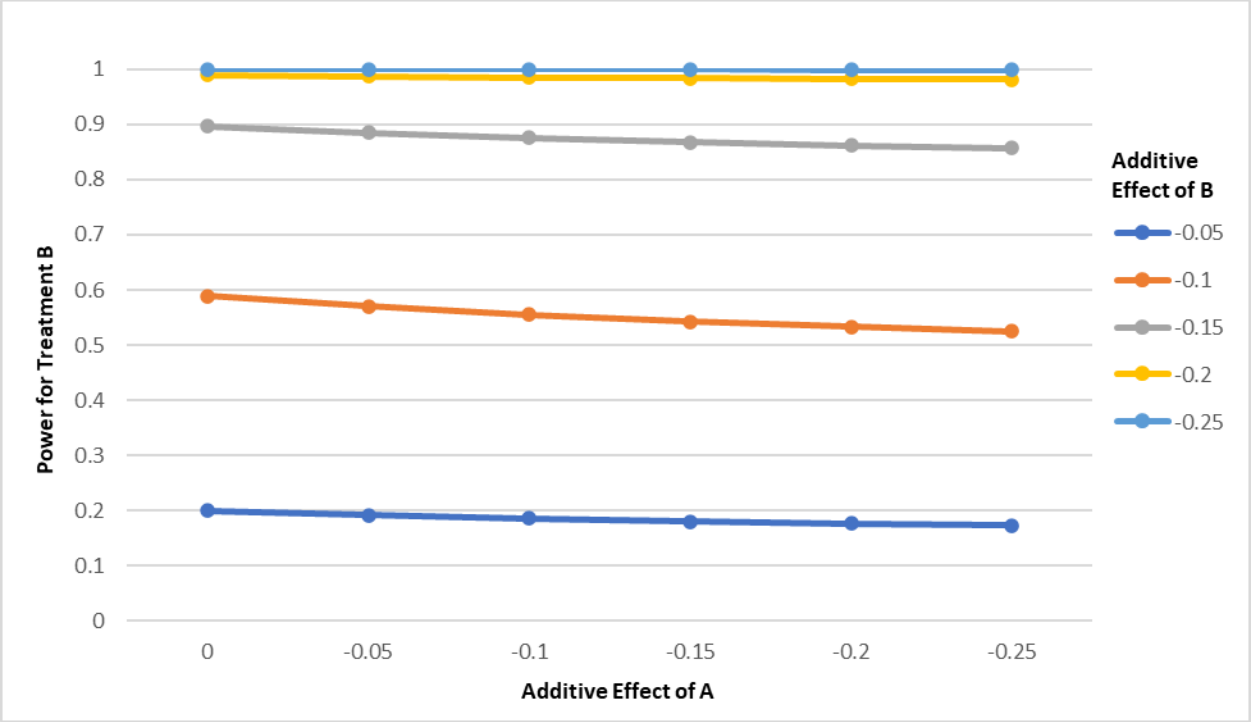


Figure 2.10: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.7$, $N = 400$, $\alpha = 0.05$, Two-Sided Test)

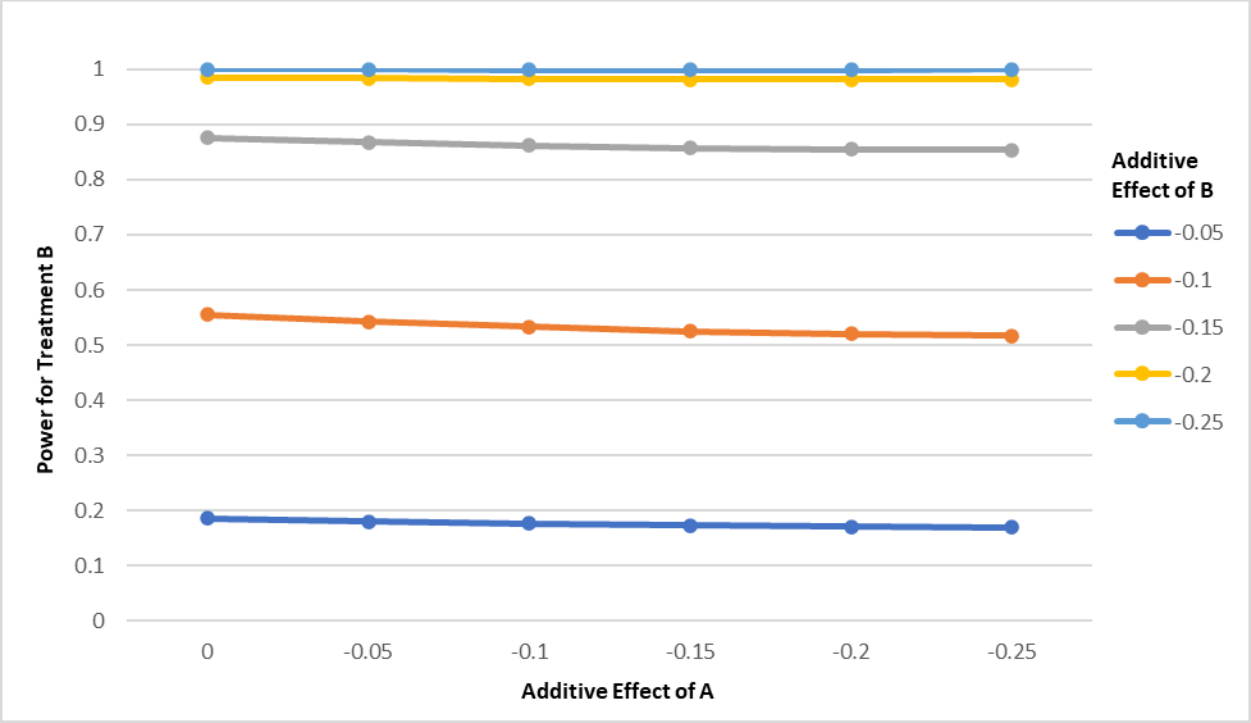


Figure 2.11: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.65$, $N = 400$, $\alpha = 0.05$, Two-Sided Test)

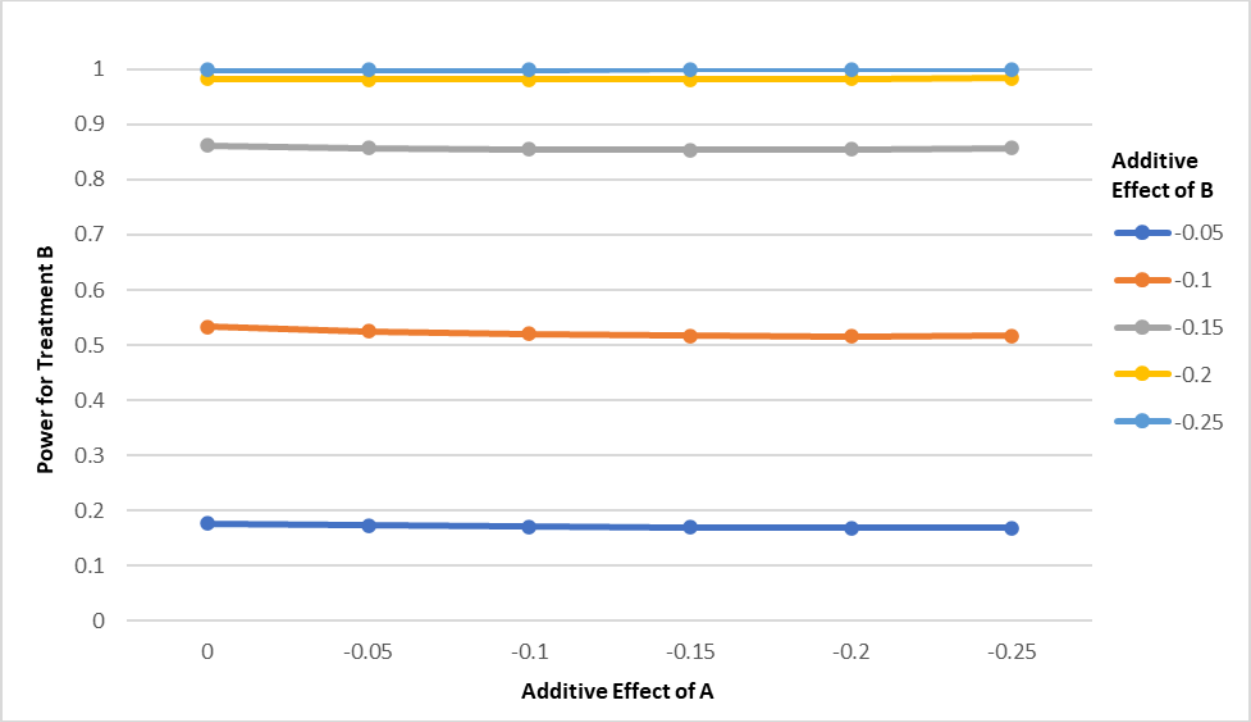


Figure 2.12: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.60$, $N = 400$, $\alpha = 0.05$, Two-Sided Test)

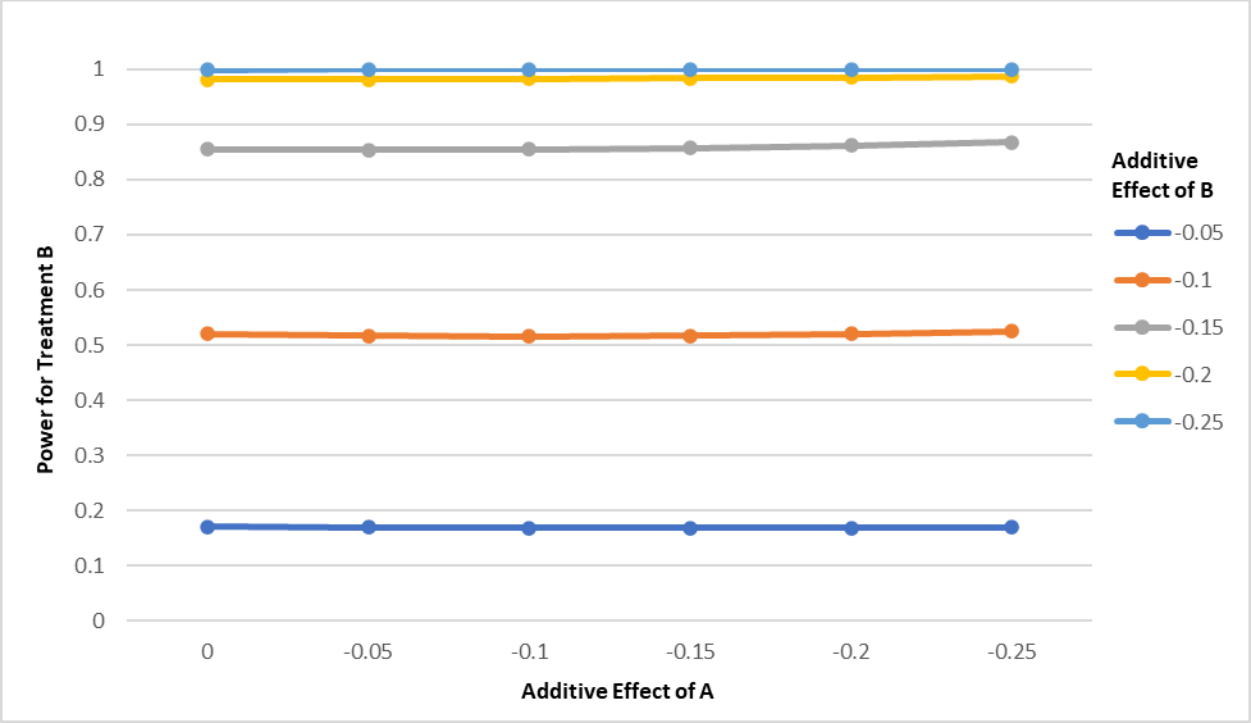


Figure 2.13: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.55$, $N = 400$, $\alpha = 0.05$, Two-Sided Test)

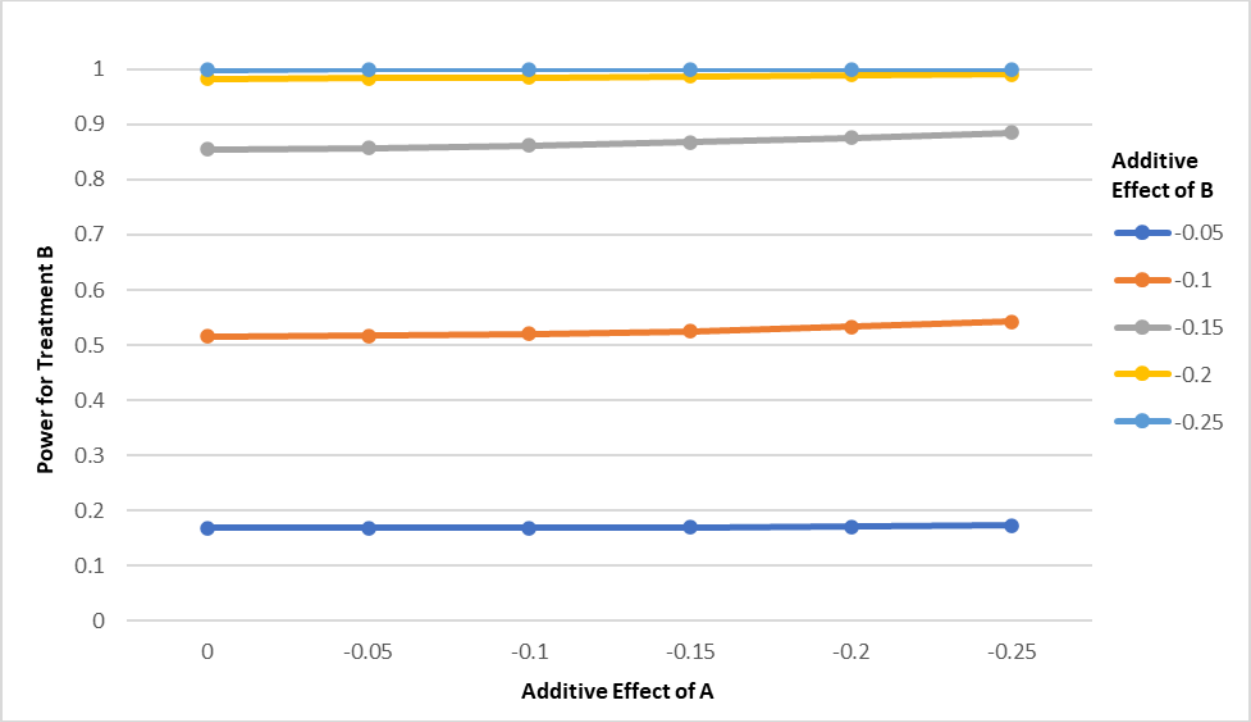
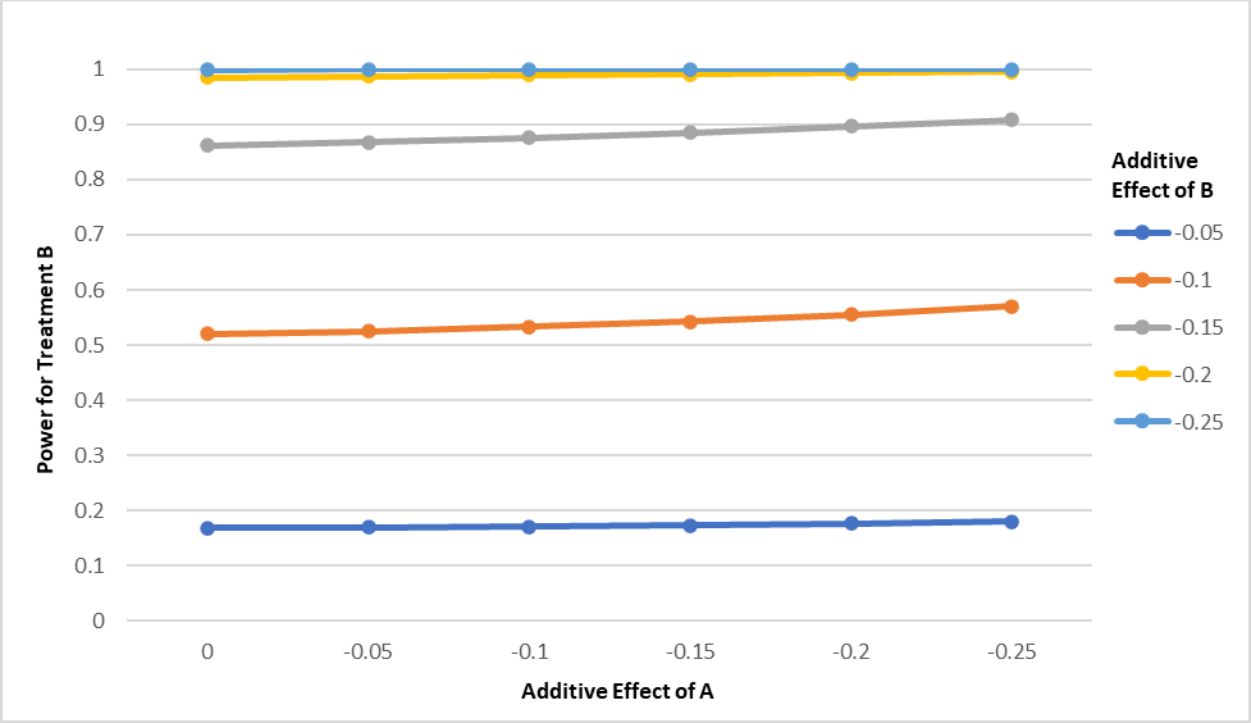


Figure 2.14: Power for Treatment B as a Function of the Effectiveness of Treatment A ($P_{00} = 0.50$, $N = 400$, $\alpha = 0.05$, Two-Sided Test)



Bibliography

1. Casagrande JT, Pike MC, Smith PG. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*. 1978:483-6.
2. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Academic press; 2013.
3. Heyland D, Muscedere J, Wischmeyer PE, Cook D, Jones G, Albert M, Elke G, Berger MM, Day AG. A randomized trial of glutamine and antioxidants in critically ill patients. *New England Journal of Medicine*. 2013;368(16):1489-97.
4. Kahan BC, Tsui M, Jairath V, Scott AM, Altman DG, Beller E, Elbourne D. Reporting of randomized factorial trials was frequently inadequate. *Journal of Clinical Epidemiology*. 2020; 117:52-9.
5. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA*. 2003;289(19):2545-53.
6. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. *BMC Medical Research Methodology*. 2003;3(1):1-5.
7. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, Lacchetti C, Leung TW, Darling E, Bryant DM, Bucher HC. Randomized trials stopped early for benefit: a systematic review. *JAMA*. 2005;294(17):2203-9.
8. Pocock SJ. When (not) to stop a clinical trial for benefit. *JAMA*. 2005;294(17):2228-30.

9. Poldermans D, Boersma E, Bax JJ, Thomson IR, Van De Ven LL, Blankensteijn JD, Baars HF, Yo TI, Trocino G, Vigna C, Roelandt JR. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. *New England Journal of Medicine*. 1999;341(24):1789-94.
10. Poole C, Shrier I, VanderWeele TJ. Is the risk difference really a more heterogeneous measure?. *Epidemiology*. 2015;26(5):714-8.
11. Sakpal T. Sample size estimation in clinical trial. *Perspectives in Clinical Research*. 2010 Apr 1;1(2):67-.
12. VanderWeele TJ, Knol MJ. A tutorial on interaction. *Epidemiologic Methods*. 2014;3(1):33-72.
13. Walter SD. Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology*. 2000;53(9):931-9.
14. Walter SD, Guyatt GH, Bassler D, Briel M, Ramsay T, Han HD. Randomised trials with provision for early stopping for benefit (or harm): the impact on the estimated treatment effect. *Statistics in Medicine*. 2019;38(14):2524-43.
15. Zakeri K, Noticewala S, Vitzthum L, Sojourner E, Shen H, Mell L. 'Optimism bias' in contemporary national clinical trial network phase III trials: are we improving?. *Annals of Oncology*. 2018;29(10):2135-9.