University of Sheffield

School of Health and Related Research

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Quantifying Effect Sizes in Clinical Trials

Author:
Joanne C. Rothwell

Supervisors:
Prof. Cindy L. Cooper
Prof. Steven A. Julious

July 2, 2018

Acknowledgements

The work presented in this thesis would not have been completed without the support of a number of individuals and organisations. Firstly, I would like to thank my supervisors Professor Cindy Cooper and Professor Steven Julious for their unwaivering support, guidance and advice throughout the process of this research. They have pushed me to keep going when I was struggling and always had faith that I would get there in the end!

There are many staff and students in ScHARR whom I would like to thank, in particular Dawn Teare and Oscar Bortolami for their constructive comments and ideas following the confirmation review and Stephen Walters for playing devil's advocate in meetings and discussions.

I would like to acknowledge those individuals who provided external advice which enhanced this research. Professor Barney Reeves and Professor David Richards discussed the processes they went through to design their respective trials published in the Health Technology Assessment, as well as allowing me to use their trials as examples in a publication. Other trialists who provided examples and anecdotes include Sallie Lamb, Helen Rodgers, Smitaa Patel and Martin Tickle.

I am grateful for the opportunity to collaborate with Jonathan Cook as part of the DELTA2 project funded by the Medical Research Council. This has allowed me a variety of platforms to present my work, as well as the possibility to work alongside many leading academics throughout the course of the project.

This PhD would not have been possible in any way without the studentship and matched funding I received from ScHARR, the Faculty of Medicine and Dentistry and the Sheffield CTRU. The funding enabled me to focus solidly on my research without external concerns, develop my teaching skills and enable me to go to conferences to present my work. All of this has built me up to this point, so thank you.

Thanks must go to my parents for their support throughout this process, my Dad for continually challenging my thinking and endless proof-reading when I could do no more, and my Mum for the emotional support she constantly provides! My sisters for providing a different view of life and distracting me when I needed it.

Finally, to Michael for sitting with me through long sleepless nights and letting me vent, being chief tea-maker and your constant support throughout this journey, and for getting Argyle - he makes us smile.

Research Achievements

Peer Reviewed Publications

Published

Walters SJ, Bonacho dos Anjos Henriques-Cadby I, Bortolami O, Flight L, Hind D, Jacques RM, Knox C, Nadin B, **Rothwell J**, Surtees M, Julious SA. (2017) Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom Health Technology Assessment Programme. BMJ Open 2017;7:e015276. doi: 10.1136/bmjopen-2016-015276

Cook JA, Julious SA, Sones W, **Rothwell JC**, Ramsay CR, Hampson LV, Emsley R, Walters SJ, Hewitt C, Bland M, Fergusson DA, Berlin J, Altman D, Vale LD. (2017) Choosing the target difference ("effect size") for a randomised controlled trial - DELTA2 guidance protocol. Trials 18:271. 2017. http://rdcu.be/tpqA

In Publication

Rothwell JC, Julious SA, Campbell MJ, Cooper C. Handbook of Statistical Methods for Randomised Controlled Trials: Chapter 12 - Sample Size Calculations. Chapman and Hall (Under Review)

Rothwell JC, Julious SA, Cooper C. A Study of Target Effect Sizes in Randomised Controlled Trials published in the Health Technology Assessment Journal. Trials (Under review).

Cook JA, Julious SA, Hampson L, Hewitt C, Berlin J, Ashby D, Sones W, Emsley R, Fergusson D, Walters S, Wilson E, MacLennan G, Stallard N, **Rothwell J**, Bland M, Smith R, Brown L, Ramsay C, Cook A, Armstrong D, Altman D, and Vale L. Choosing the target difference (effect size) for a randomised controlled trial - MRC guidance. (Under Review).

Conferences and Meeting Presentations

Invited Sessions

Rothwell JC and Julious SA (presented by Julious SA). Quantifying Effect Sizes in Clinical Trials - Joint Statistical Meetings, Baltimore 2017. *Voted as one of three honorable mentions for the JSM Biopharmaceutical Section Contributed Paper Awards.*

Rothwell JC. A Review of Methods to Quantify Effect Sizes in Clinical Trials-Statisticians in the Pharmaceutical Industry, London 2017

Rothwell JC. Quantifying Effect Sizes in Clinical Trials - Health Technology Assessment Review Update. - DELTA2 Workshop, Oxford 2016

Rothwell JC. Quantifying Effect Sizes in Clinical Trials - Health Technology Assessment Review. Society for Clinical Trials, Montreal 2016

Contributed Sessions

Rothwell JC. "Quantifying Effect Sizes in Clinical Trials - Simulation of Trial Scenarios." - Poster. Society for Clinical Trials, Liverpool 2017. *Finalist for the SCT Best Poster Prize.*

Walters SJ, Henriques-Cadby I, Bortolami O, Flight L, Hind D, Jacques R, Knox C, Nadin B, **Rothwell JC**, Surtees M, Julious SA "Recruitment and retention of participants in randomised controlled trials: a review of trials funded by the United Kingdom Health Technology Assessment Programme" .- Poster. Society for Clinical Trials, Liverpool 2017

Marshall E, Wilson DA, **Rothwell JC**, Karadimitriou SM, Smith S. "Tackling barriers to statistics learning" - Poster. University of Sheffield Learning and Teaching Conference, Sheffield 2017 and Learning support in STEM: Learning differences and teaching innovations conference, Sheffield 2017

Abstract

Introduction: This thesis discusses the importance of the target effect size (ES) used in clinical trials of health interventions. It investigates the common methods of eliciting the target difference and whether target values are optimistic or unrealistic. Regression to the mean (RTM) is shown for trials in sequence, which is assessed through simulations and an adjustment developed to adapt for this bias.

Research Question: Investigating currently used methods for eliciting the target difference and optimal methods for adjusting for RTM.

Methods: Firstly, a review of the Health Technology Assessment (HTA) journal trial reports of parallel-group randomised controlled trials (RCTs) was performed. The standardised observed and target ES were compared for various clinical areas and elicitation methods. Second, performing simulations of trials in sequence to investigate the effect of RTM. A mathematical solution was evaluated to confirm these simulated results.

Results: A review of 107 HTA reports showed the median standardised target ES is 0.30 (mean= 0.30), and the median standardised observed ES is 0.11 (mean= 0.19). Use of previous research was the most common method of elicitation. Simulations showed RTM occurs for trials in sequence, an adjustment method has been developed and proven mathematically, which depends only on the power of the first trial.

Conclusions: This thesis demonstrates the most common method of target difference elicitation is the use of previous research. This method leads to RTM of the observed ES. An adjustment based on the power used in the initial trial power and the progression criteria in pilot studies has been developed and tested. If trialists adopt this adjustment then trial sample sizes, though slightly inflated, would potentially provide more realistic estimates of the target ES.

Contents

List of Abbreviations

Abbreviation	**Meaning**
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
BMJ	British Medical Journal
DELTA	Difference ELicitation in TriAls
EMEA	European Medicines Agency
ES	Effect Size
GP	General Practitioner
HTA	Health Technology Assessment
ICC	Intracluster Correlation Coefficient
ICH	International Conference of Harmonisation
JAMA	Journal of the American Medical Association
JSM	Joint Statistical Meetings
MCID	Minimum Clinically Important Difference
MDD	Minimum Detectable Difference
MRC	Medical Research Council
NEJM	New England Journal of Medicine
NIHR	National Institute of Health Research
PhRMA	Pharmaceutical Research and Manufacturers of America
PSI	Promoting Statistical Insight
QoL	Quality of Life
RCT	Randomised Controlled Trial
RSS	Royal Statistical Society
RTM	Regression to the Mean
ScHARR	School of Health and Related Research
SCT	Society of Clinical Trials
SIC	Sufficiently Important Difference
UKCRN	UK Clinical Research Network

1. Introduction

1.1 Introduction

Controlled experiments have been used informally since the 1930s in research areas such as medicine, education and social welfare (Shapiro and Louis, 1983), with the first randomised clinical trial in the 'modern' sense being reported in 1948 (Forbes and Holt, 1948; Julious and Zariffa, 2002). In a medical context, these controlled experiments are more commonly known as controlled clinical trials and are usually performed with human participants (Bland, 2000). Trials are used to assess how well a particular health technology or treatment works, as well as the safety of the treatment (Durham, 2008). However, controlled experimental trials can be expensive and time consuming, and in some cases are not feasible to conduct (Shapiro and Louis, 1983).

The main purpose of a clinical trial it to get an unbiased, reliable assessment of a treatment or therapy (Julious, 2010b). This is achieved with a number of design features. These include randomisation of participants (Guideline, 1995; Lewis, 1999), blinding of participants and/or trialists, and the inclusion of a control group to quantify the response of the experimental treatment (Julious and Zariffa, 2002). The first documented case of randomisation was in 1948 by Austin Bradford Hill (Bradford-Hill, 1990; Forbes and Holt, 1948; Yoshioka, 1998), and it is now part of the International Conference on Harmonisation (ICH) guidelines (ICH-E9) recommendations for clinical trials (Lewis, 1999). With the number of trials in the United Kingdom in the set-up or recruiting stages exceeding 5500 (2015), according to the UK Clinical Research Network (UKCRN) there is a constant need to use suitable methodology in the design and analysis of these trials (Network, 2015).

1.1.1 Chapter Aims

This chapter aims to introduce the various types of clinical trial available, as well as introducing a standard sample size calculation which is used in the design of the trial. Each variable in the calculation will be defined and it will be briefly discussed which variable is most sensitive to change. This will inform the basis of the research aims for this PhD project.

1.2 Background Information

1.2.1 Types of Clinical Trial

There are a number of considerations when designing a clinical trial. First, the research question needs to be established and the outcomes of interest specified (Julious, 2010b). There are broadly three main types of outcome; these are continuous outcomes, binary (or survival) outcomes and ordinal outcomes (Julious, 2010b). A more specialised type of outcome is used in survival data, where the outcome is time-to-event. A continuous outcome is assessed on a scale where the measurement takes a number which can vary by the precision of the instrument taking the reading. For example, systolic blood pressure is a continuous outcome because it can be any number, though there are certain ranges which are implausible or the instrument cannot measure more accurately. A binary outcome is categorical and can take one of two possible values, for example a participant could be alive or dead, they could have responded or not responded to treatment. Ordinal data is, like binary data, categorical data; however it has a set order to the categories. There are a number of different aims for clinical trials (Flight and Julious, 2016) which can be grouped into three main objectives:

- Superiority Trials to show that one therapy or treatment is better than another,
- Equivalence Trials to show the treatments or therapies are equal,
- Non-Inferiority Trials to show that one therapy or treatment is no worse than another.

Finally, there are two main study designs, namely parallel group designs and crossover designs. These designs will be discussed further in chapter 2. In pharmaceutical clinical research there are four main stages, known as phases (Durham, 2008). These are called Phases I-IV which are further defined in chapter 3, but Phase III trials are the major large scale comparative investigations used to assess a treatments efficacy compared to other treatments (Altman, 1999; Durham, 2008). Phase III trials are investigating a formal hypothesis, whereas the focus of the preceding phases is assessing dosage, safety and determining if the trials are worth continuing

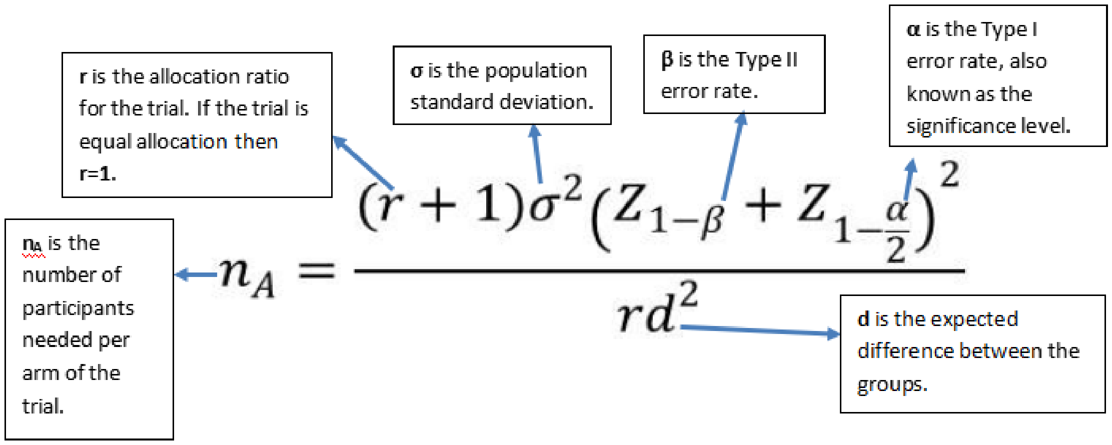

r is the allocation ratio for the trial. If the trial is equal allocation then r=1.
σ is the population standard deviation.
β is the Type II error rate.
α is the Type I error rate, also known as the significance level.
nA is the number of participants needed per arm of the trial.
$n_A = \frac{(r+1)\sigma^2\left(Z_{1-\beta}+Z_{1-\frac{\alpha}{2}}\right)^2}{rd^2}$
d is the expected difference between the groups.

Figure 1.1: A standard sample size formula

1.2.2 Sample Size Calculations

Once the type of trial to be performed has been determined, the next consideration is the calculation of the sample size. This is an important step in the trial design as a study which recruits too many or too few participants may be deemed unethical (Altman, 1980). A study which recruits too many participants runs the risk of reaching the study goal before the trial has completed, therefore exposing more participants to a treatment which is less effective (Altman, 1980). A study which has recruited too few participants has a high chance of failing to meet the study aims, putting participants through the trial when there would be no benefit of doing so as no conclusive results could emerge (Altman, 1980).

A sample size calculation for a continuous outcome measure can be estimated from an equation of the type seen in Figure 1.1 (Brush, 1988; Durham, 2008). It consists of 5 different variables;

- the population standard deviation for the outcome of interest (σ),
- the target population difference between the treatments (d),
- the Normal value for the power ($Z_{1-\beta}$) of the trial,
- the Normal value for the significance of the trial ($Z_{1-\alpha/2}$)
- and the allocation ratio between the groups (r) (Brush, 1988; Fleiss, 1986).

These variables will be further defined in chapter 2. All the variables in the sample size calculation are fixed through design in the case of r, β and α, and estimated in the case of σ. The only exception to this is d (Friedman, 2010), which is dependent on the intervention and the results from elicitation, to be discussed in section 1.2.3.

The most sensitive part of the sample size formula is the target difference, d; if the target difference is halved whilst the remaining parameters are unchanged, the

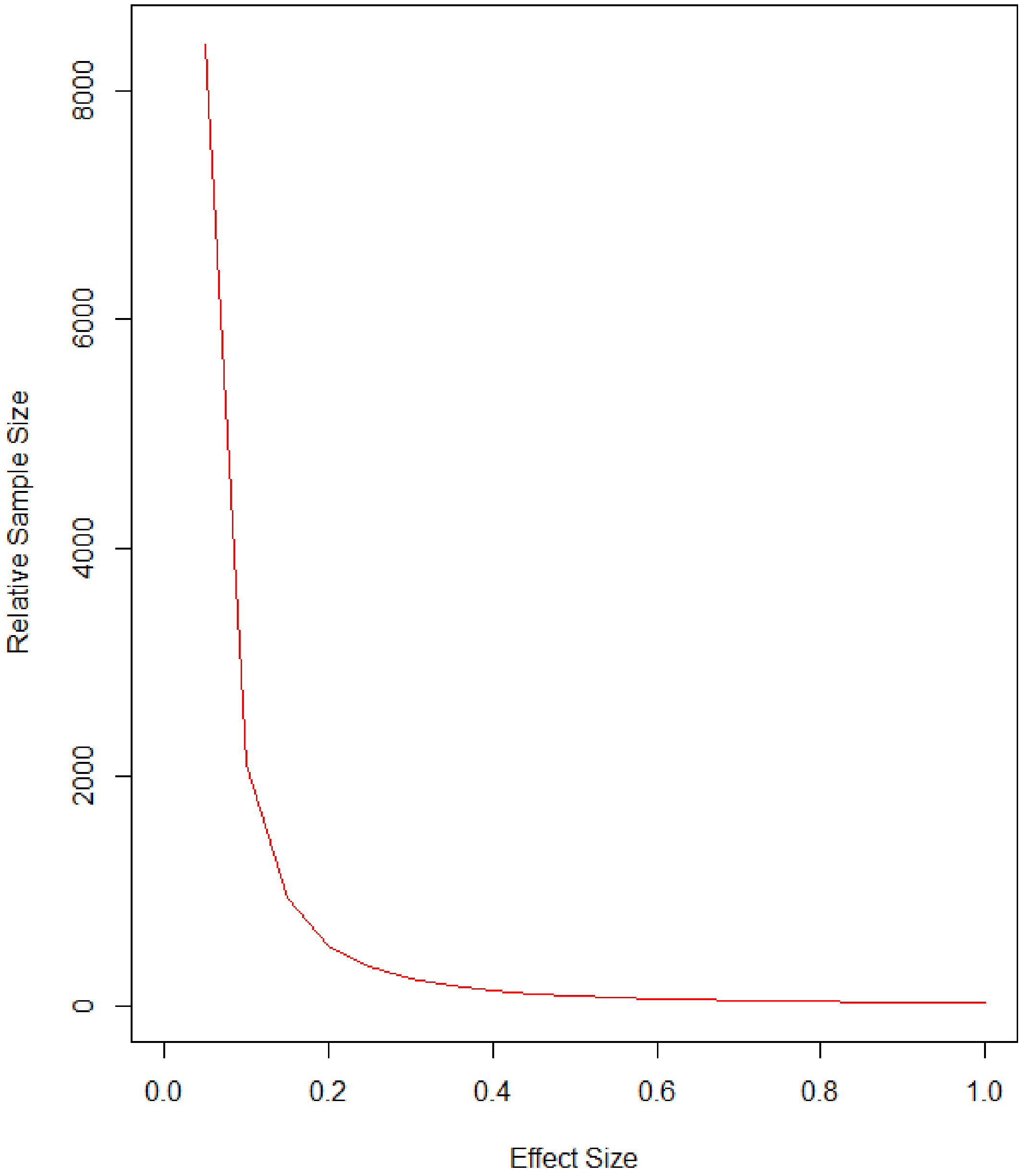


Figure 1.2: The impact that the effect size has on the sample size for 90% power and 5% significance level.

required sample size is multiplied by four (Fayers and Machin, 1995).This shows that this variable is extremely sensitive in the calculation and it should be chosen with the utmost care. This can also be seen in Figure 1.2, which clearly shows as the effect size gets smaller the sample size required increases dramatically. This sensitivity to d will form the basis of the PhD. If the target difference is over-estimated, the sample size will be smaller and this could potentially result in an under-powered trial (Fayers and Machin, 1995; Friedman, 2010). This means that even if there is a true effect to be observed, the trial is not powered highly enough to observe this effect. Studies with low power run the risk of rejecting treatments which are beneficial (Friedman, 2010).

There are two main approaches which can be used to lessen the likelihood of having a low-powered trial, these are to increase the number of participants or to increase the significance level (Cohen, 1973). These approaches have their own associated problems, such as a trial becoming too expensive if more participants are needed, or have a Type I error rate undesirably high (Cohen, 1973).

1.2.3 Determining the Target Effect Size

The task of determining d for trial design is difficult, with a number of different methods used. These will be discussed in further detail in chapter 3. There is the consideration that the method used to estimate the target effect size could cause a bias depending on the method chosen. The definition of the type of bias which will be discussed in this document is "Any experiment, study or measuring process is said to be biased if it produces an outcome that differs from the truth in a systematic way" (Everitt and Palmer, 2010). Bias may occur when moving from one trial to the next, conditional on observing encouraging results in the first trial. In order for there to be no bias, the second trial would be performed irrespective of the results of the first trial. However, in practice, this would not occur as if the results of the first trial were not encouraging then the second trial would usually not be performed. The effect of this bias will be investigated further in chapter 4.

1.3 Research Aims and Questions

Having illustrated the potential implications of over-estimating the estimated effect size, it prompts the question of how many trials are actually achieving their target or estimated effect size used in the original sample size calculation.

Effect sizes are known to vary across research areas, with areas such as nutrition and genetics consistently reporting extremely small effect sizes (Siontis and Ioannidis, 2011). If it is in fact true that trialists are over-estimating the target effect size in their sample size calculations, it would be useful to know, given a particular disease area or research area, a range of plausible effect sizes. This could be used to assist trialists when designing trials for which there is no previous research or data. The discussion in this chapter leads to the following research questions

1. How are effect sizes quantified in the design of clinical trials?
2. Are observed effect sizes similar to *a priori* (target) effect sizes? Is this effect size clinically important?
3. What range of observed effect sizes are being seen in different clinical areas or populations?
4. Are there more optimal methods for quantifying the effect size?
5. Are there more optimal methods to adjust for the bias of moving from one trial to the next?

The term "similar" that is used in research question two will be formally defined later in the thesis. The research questions numbered one to three aim to investigate what is the current situation with regards to observed effect sizes in clinical trials, particularly focusing on whether trialists are being overly optimistic or pessimistic with the target effect size when designing the trial, and in which clinical areas is this more likely to occur.

Research questions four and five will aim to investigate methods to adjust for the potential bias which can occur when moving from the early trial to a later stage trial based on the early results. This question focuses primarily on the scenario of moving from one trial to the next, moving on to investigate the scenario of a pilot to a main trial. There are other designs which are outside the scope of this thesis, such as surrogate end-points and other more complex trial designs. These are briefly discussed throughout the thesis and discussed in more detail in chapter 8.

Once the current situation is known, it will allow for practical advice to be formed for those designing trials based on prior results. The fourth and fifth research questions will be investigated using simulations to test different scenarios, in order to propose an adjustment method for the target difference when moving from one trial to the next. There is also the opportunity to test this adjustment mathematically, along with determining the adjustment through the use of algebra. The proposed adjustment has also been tested on real data to assess its validity in comparison to simulated data.

1.3.1 Further details of Research Questions

It is vital for any research to consider the importance of performing the research itself, and the importance of the research questions. This section will explain why each research question is important and also the methods which are to be used to derive a solution.

- Are the observed effect sizes similar to *a priori* (target) effect sizes?

The target difference can be defined as the difference which the trial is aiming to observe. This can be further sub-categorised as shown below, though there are many other distinctions available (Cook et al., 2014; Harris et al., 2017).

- **Target Difference** - the value which is used in the sample size calculation. It could be larger than the clinically important difference, it is the value which would prompt a change in clinical practice. Alternatively, it could be the difference by which a drug shows superiority to other similar drugs on the market. The target difference can be further categorised as follows:
  - **Minimum Clinically Important Difference (MCID)** - the smallest value which you could accept that the treatment actually works or has clinical benefit (Andrews et al., 2011; Lenth, 2001).
  - **Sufficiently Important Difference (SID)**- the smallest benefit that an intervention would require to justify costs and risks (Barrett et al., 2007)
  - **Minimum (statistically) Detectable Difference (MDD)** - the minimum value which is statistically detectable at the pre-specified type I error rate (Hanson et al., 2003a; Piva et al., 2004).

This question considers two main contexts. The first context is that the target difference is set to be the minimum clinically important difference (MCID). The second context is that the target or anticipated difference is chosen based on previous research, or is unjustified. These two contexts need to be considered separately, since the premise of each is very different.

The reason this research question is of interest is that there are a number of articles (and it is the common belief) that researchers are overly optimistic when choosing the anticipated effect size for a sample size calculation (Campbell, 2013; Fayers and Machin, 1995; Friedman, 1985). If these anticipated effect sizes are consistently overly optimistic, then there is the possibility that they need to be re-assessed if they are not being achieved. This leads to the need to have to consider the method by which the anticipated difference is chosen.

If the anticipated difference is chosen as a MCID, then it cannot be deemed to be overly optimistic if that is truly the minimum important difference. If it is a true MCID then the interventions which are shown not to be achieving the MCID may be truly clinically ineffective. If the MCID is optimistic, however, then this leads us to ask the question of where the MCID came from.

Consideration will be given to whether the target difference is based on well-founded and advised methods of elicitation, such as opinion-seeking or review of the current evidence (Cook et al., 2018) or whether it has been set arbitrarily or is unjustified. All these factors will influence whether the target difference is deemed realistic or not. There are many reasons why the target difference could be unrealistic, there could be a need to re-evaluate it if it is unachievable.

Another reason could be that currently it is unachievable but it is truly the minimum clinically important difference so should still be used as different interventions are developed. This could lead to the need to determine whether it can be re-evaluated or whether it is set as the MCID because it truly is the MCID.

For the trials which are not based on a MCID, a question to consider is where are these anticipated differences coming from. To be able to identify which methods are commonly used to elicit the target difference would be useful, as well as investigating whether particular methods producing higher estimates of effect size. As highlighted earlier it should also be considered that there is the possibility that a trial is not statistically significant ($P \geq 0.05$) and the MCID is not met, this could be due to the MCID being accurate but the treatment simply not being effective enough to reach the pre-specified minimum clinically important difference. The assessment of the methods used to determine the MICD could be used to establish whether certain methods are providing unrealistic or unachievable MCIDs or target effect sizes.

Trials of a common clinical condition, diabetes, consistently fail to demonstrate the identified MCID. In informal correspondence with Professor Simon Heller, a chief investigator at the University of Sheffield who receives substantial NIHR funding for research in this area, he raised concerns that funders will start refusing funding because the treatments are deemed to be not effective enough in diabetes, yet funding is needed to continue to work towards treatments which are effective. There is also empirical evidence of the MCID falling over time in trials for the treatment of depression (Voehringer and Ghaemi, 2011).

- What range of observed effect sizes are being seen in different clinical areas, populations or interventions?

This research question stems from the previous question and they are linked together. It would be useful to be able to demonstrate plausible ranges of target effect sizes for various interventions or clinical areas.

Summary of Questions 1 and 2

The key things which will be taken away from these two questions are the different methods used to elicit the target difference when designing the trial, the anticipated and observed effect sizes, the sample size, target power and significance. It is also interesting to stratify the results to explore differences in these factors by: clinical area; outcome measure and intervention type. It is known that half of all publicly funded trials have a non-significant P-value (Sully et al., 2013), therefore it is target that at least half of the trials included in this review will have a non-significant P-value. This does not mean that those interventions are ineffective, simply that there was not enough evidence to show there is an effect there. The methods which are planned to research these questions are discussed in more detail in chapter 4.

This part of the research will focus on informing what the current situation is. This is an important area of research, demonstrated by the fact that the MRC has recently called for research into quantifying effect sizes and the Medical Statistics group at ScHARR have been successful in its co-application for this funding. The research done in these two research questions will contribute to this research.

- Are there more optimal methods for specifying the anticipated effect size?

This question is important because it should lead to providing advice for trialists with regards to eliciting an appropriate target effect size when designing future trials. A common design which is used in ScHARR and other research groups is basing a large trial on a proof-of-concept trial, or similar published studies. These specifically focus on target effect sizes, and are not necessarily based on the MCID.

This question will be dependent on the results discovered in questions 1 and 2. Based on which methods are most commonly used and which ones provide a 'close' estimate of the effect size compared to the observed effect size. It would be interesting to know if there are differences in the sample sizes used based on the methods applied for elicitation of the anticipated difference. This research question will aim to reflect and evaluate the results from questions 1 and 2, as well as make recommendations for future trials. It would also be interesting to assess the relationship between the method used and whether the anticipated effect size is an over- or under-estimation of the observed effect size.

- Are there more optimal methods to adjust for the bias of moving from one trial to the next?

This question is of interest, as it will allow us to provide some advice for future trialists. This research question is important because it will bring all the recommendations and advice from the previous questions and aim to produce a method which can be used in practice. The methods used for this question are discussed in chapter 7; however the main consideration is the evolving nature of this question. Chapter 6 discusses the methods which are currently published, as well as the limitations of these methods. They will then be compared to methods developed in chapter 7 for trials in sequence. The context of the trials in sequence is determined by the most commonly observed methods of target difference elicitation found in chapter 4. For example, if it was observed in chapter 4 that the most common method of target difference elicitation reported were pilot study to main trial, then the development of another method to adjust for the bias will focus on the context of pilot study to main trial. It is key to tailor this question to contexts which will prove most useful for trialists in the future.

1.4 Development since Start of the PhD

Following the commencement of this PhD, the Medical Statistics group in the School of Health and Related Research (ScHARR) acquired a grant from the MRC as co-applicants to research the area of choosing the target effect size (Cook et al., 2017). There are 5 research questions stipulated in the grant proposal which focus on review of the current guidance provided by funders to identify any key methodological developments or changes in practice. The objectives are, amongst other things, to determine the scope of the guidance required that would aid researchers and address funders needs in terms of determining a target effect size for sample size calculations. The research in this PhD contributed to the study by performing the review of the HTA monographs to establish how effect sizes are chosen.

As part of the process of the research for this MRC grant funded study, the work presented in chapter 4 was presented at the Society for Clinical Trials (SCT) conference in Montreal as part of a structured session. This work, along with work from chapter 6, has also been presented at the SCT-MRC conference (Liverpool, 2017), the PSI conference (London, 2017), JSM conference (Baltimore, 2017) and the Royal Statistical Society seminar (Sheffield, 2017).

1.5 Chapter Summary

This introduction has provided a brief overview of clinical trials and sample size calculations, as well as details about the proposed research questions outlined in section 1.3 and their relevance to the fields of medical statistics and clinical trials. The work presented in this thesis will follow a frequentist framework only, since Bayesian methodology does not include the use of a target difference in the same context as the frequentist methodology.

The format for each chapter will be a brief recap of previous chapters in each introduction, followed by the chapter aims and the chapter content. They will then finish with discussion and conclusions in the context of work already covered and signposting work which is to be covered in later chapters.

The different sample size calculations used for different trial designs will be discussed in chapter 2, with the methods commonly used to determine the target difference, d, will be discussed further in chapter 3. Chapter 4 will present a review of the Health Technology Assessment journals, showing the most common methods of elicitation, as well as considering possible problems with these methods in terms of potential bias. This bias is further explored in a literature review in chapter 5, setting the context for simulations in chapter 6. From this, a discussion and possible avenues for further work are described in chapter 8.

2. Background to Sample Size Calculations

2.1 Introduction

Chapter 1 highlighted the importance of design in clinical trials, as well as the fundamental components of a typical sample size calculation. This chapter will discuss the various sample size formulae used for different trial designs and outcome measures, as well as discussing the sensitivity of the calculation to the anticipated effect.

The sample size calculation is an important part of a trial protocol. The calculation estimates the minimum number of patients needed in the trial for a given power and significance level for a pre-specified clinically meaningful difference between the two treatments (d).

The clinically meaningful difference is important for a number of reasons- both clinical and statistical. The difference in treatments could be, for example, the difference in the proportion of patients surviving on an experimental treatment compared to the standard treatment, or the expected change in systolic blood pressure on an experimental treatment compared to an active control treatment. It could be argued that the sample size calculation is most sensitive to the target difference. If the difference is halved then the sample size quadruples (Fayers and Machin, 1995). This is a serious consideration since very large sample sizes may not be achievable due to cost, time or other resources.

2.1.1 Chapter Aims

This chapter will describe the sample size calculations for parallel group trials, focusing on superiority studies. Brief discussion of equivalence and non-inferiority studies is included, along with crossover trials, binary outcome measures and cluster trials. This thesis chapter is based on a book chapter written by myself in my first year of this PhD which has been accepted for publication (Rothwell et al., 2018a), which draws heavily on work by Julious (Julious, 2004).

This chapter also aims to show the importance of d in each calculation as well as how the effect size impacts on the sample size for the various trial types. In this context (and in context for the PhD) before we can truly understand the sensitivity of the sample size calculation to d, we need to understand its operational impact on the calculations.

2.2 Types of Trials

For the purpose of this chapter we shall be focusing on trials comparing two treatments in the form of a parallel group trial.

2.2.1 Parallel Group Trials

In a parallel group trial, patients are randomly assigned to one of two treatment groups. The ideal scenario is that all other baseline characteristics of the patients are roughly similar in each group (i.e. equal number of males and females, location of hospital, patient ages). If we consider a diagram of the trial, it would look like Figure 2.1.

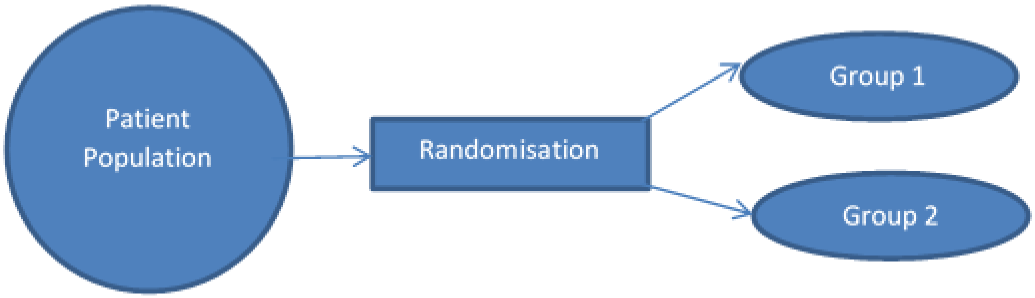


Figure 2.1: Illustration of a typical randomisation for a parallel group trial.

Each group is given a different intervention (e.g. treatment versus placebo, experimental treatment versus current standard treatment) then the two groups are compared directly with each other. There are a number of considerations when a randomisation occurs, one of which is what type of randomisation method will be employed. There is simple randomisation, which is similar to tossing a coin and assigning each patient based on the outcome of the coin (Tails = Group A, Heads = Group B). However, due to the random nature of this method, there could be an imbalance in the number of patients in each group. The size of this imbalance is larger in small trials compared to large trials, thus having a bigger impact on the power. One method to ensure there are equal numbers of patients in each group is to use blocked randomisation. This is best illustrated using an example. Consider the basic case of two treatment groups. The aim is to randomly allocate the patients to either Group A or Group B. If we consider a block size of 4, we would need to know that at the end of each block there had been allocations to each of the groups.

Consider creating a randomisation list by tossing a coin. We would allocate Group A to heads (H) and Group B to tails (T). Now in our block size of 4, we need to ensure there are 2 heads (Group A) and 2 tails (Group B) after 4 tosses of the coin. The blocks would look like those shown below.

Block 1	A	B	A	B
Block 2	A	A	B	B
Block 3	A	B	B	A
Block 4	B	A	A	B

The first patient would then be randomly allocated within a block, with the following three patients also being allocated within that block. This ensures that after each block has been completed there are equal patients in Group A and B.

Another randomisation method is to use stratification. This method is similar to blocking but whilst the blocking method ensures balance with defined block sizes, stratification also ensures balance by strata. These strata are subgroups of clinical importance and are usually patient characteristics, for example, age or gender. Once again we shall illustrate this using an example. Suppose we are tossing a coin to create the randomisation list. For this randomisation, not only are we trying to ensure equal numbers of patients in the two groups, but we are stratifying by gender so there will be two strata. We then generate the randomisation list separately for each stratum. An example of this is

Stratum 1				
Block 1	AB	AB	BA	BA
Block 2	AB	BA	AB	BA
Block 3	BA	BA	AB	AB
Block 4	BA	AB	BA	AB
Stratum 2				
Block 1	BA	AB	AB	BA
Block 2	AB	BA	BA	AB
Block 3	AB	BA	AB	BA
Block 4	BA	AB	BA	AB

Now, in this artificial example, after 16 recruited or randomised participants we can see that there is balance in terms of overall group allocation as well as a balance in group allocation by strata. This method is commonly used for crossover trials which are discussed below.

2.2.2 Crossover Trials

The aim for this type of trial is that all patients experience all treatments. This is done by randomly allocating the patients to different treatment sequences. For two period crossover studies these would be AB or BA. Thus, each group has a different order of treatments assigned to it. This is illustrated in Figure 2.2.

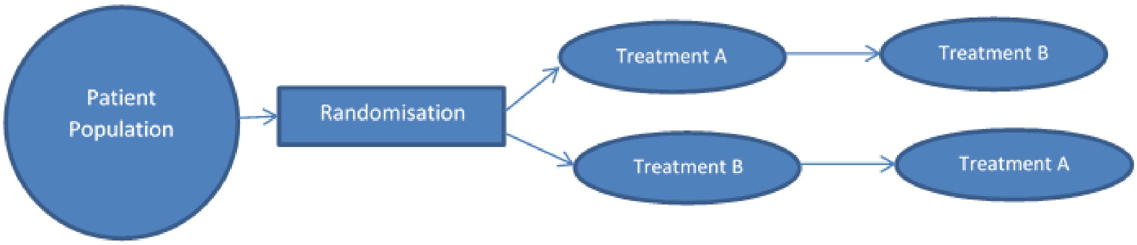


Figure 2.2: Illustration of a typical randomisation for a crossover trial

The key point for this trial is that all patients get all treatments; the order in which they are given is random. Once the first treatment has been given, the groups usually have a washout period in order for a patient to clinically return to baseline, therefore providing a constant base on which to test each of the subsequent treatments.

Crossover trials enable within-patient analysis since each patient has experienced both treatments. Once again we can consider an example to illustrate how the crossover trial is formed. The two period (AB/BA) design is the simplest type of crossover trial, which simply compares treatment A with treatment B. The subjects will be randomised to either have treatment A in period 1 followed by treatment B (AB sequence), or treatment B in period 1 followed by treatment A in period 2 (BA sequence). Subjects are randomly assigned to either sequence AB or BA, and as in parallel group trials, blocking can be used to ensure balance.

If we continue with the examples from earlier, with a block size of 4, instead of A and B being separately assigned to heads and tails, now the sequences AB and BA are assigned to heads or tails, which would look something like:

Block 1	AB	BA	AB	BA
Block 2	AB	AB	BA	BA
Block 3	AB	BA	BA	AB
Block 4	BA	AB	AB	BA

There are two main assumptions when performing a crossover trial, namely that the order the treatments are received does not affect the patients response to the treatment, and that all patients return to baseline prior to the second treatment. This type of trial is best suited to long-term stable conditions such as eczema or asthma and less so for conditions where the patient is likely to get worse over time (degenerative conditions).

Decide to	Null hypothesis is actually	
Reject Null Hypothesis	Correct Decision Power of test $(1-\beta)$	Type I Error (α)
Not Reject Null Hypothesis	Type II Error (β)	Correct Decision

Table 2.1: Type I and Type II Errors and Power

2.3 Continuous Outcomes

This section will focus on trials with continuous outcomes. Within each outcome type, the trial designs will be discussed separately.

2.3.1 Superiority Trials

Superiority trials focus on providing statistically significant evidence against the null hypothesis that the two treatments in question are the same with respect to the comparison of interest (i.e. mean response time). The null and alternate hypotheses are as follows:

- H_0: $\mu_A = \mu_B$ (The two treatments are the same)

- H_1: $\mu_A \neq \mu_B$ (The two treatments are different)

The two types of errors which can occur when testing the null hypothesis are Type I and Type II errors (Julious and Walters, 2014; Neyman and Pearson, 1928a,b, 1933a,b). These errors are shown in Table 2.1 in terms of the decision made at the end of the trial and whether the null hypothesis is really true or not, but are also summarised below.

- Type I Error : Rejecting H_0 when it is true
- Type II Error : Not rejecting H_0 when it is false

The sample size calculation takes these errors into consideration and aims to estimate the sample size whilst minimising them. The type I error is conventionally fixed at a two-sided level of 0.05 which is also the significance level of the trial. The type II error is usually fixed at 0.1 but can be as high as 0.2.

A type I error is deemed more serious in both medical and financial terms. From a medical perspective, a type I error would mean an ineffective treatment being shown to be effective when it is not the case. Giving a patient an ineffective treatment would be unethical, as it may be preventing the patient from receiving an effective treatment, and could result in the condition of the patient deteriorating. From a financial perspective, money may be spent to change treatments and medical practice unnecessarily if the new treatment were not more effective than the current treatment.

A type II error is less costly both clinically and financially. A type II error would not result in a change in medical practice, although the patients may be deprived of an effective intervention, and is therefore seen as the lesser of two evils. It is because a type II error is deemed not as costly as a type I error that it can be set at a higher level.

It is more common to think not in terms of the Type II error but in terms of the power of a trial (power$= 1 - \beta = 1-$Type II Error). The power of a trial can be understood as the probability of rejecting H_0 when a specific alternative is true, so getting the result correct. Therefore the higher the power of a trial is, the more likely that a significant result will be declared if there is truly a difference. It is the chance that the study will provide a conclusive result. If we consider the null hypothesis for a superiority trial, it can be seen that there are two ways in which the null can be rejected. This can occur if $\mu_A < \mu_B$ or if $\mu_A > \mu_B$. It is due to these two chances that this type of test is referred to as a two-tailed test. This simply means that the Type I error is split equally between the two instances, so each tail has a 0.025 probability of occurring under the null hypothesis. Superiority trials are normally used when comparing a treatment to a control. This control could be a placebo (negative control) or a current treatment (active control).

2.3.1.1 Parallel Group Trials

As mentioned earlier in the chapter, a parallel group trial consists of two groups with the sample size for group B being able to be written as a multiple of the sample size for group A ($rn_B = n_A$) where r is the allocation ratio. A preliminary sample size calculation for the population would be (Brush, 1988)

$$n_A = \frac{(r+1)(Z_{1-\beta} + Z_{1-\alpha/2})^2\sigma^2}{rd_S^2}, \tag{2.1}$$

where r is the allocation ratio, d_S is the target difference between the treatments, $Z_{1-\beta}$ and $Z_{1-\alpha/2}$ are the Normal values for power $(1-\beta)$ and significance (α), and σ^2 is the population variance.

For the variances of each treatment group A and B (σ_A^2, σ_B^2) respectively, the assumption is $\sigma_A^2 = \sigma_B^2 = \sigma^2$ (homoscedasticity). This will be the assumption throughout the chapter. Each part of the sample size formula is important; however there is one component to which it is most sensitive. This is the target difference d_S. It can be shown that if the target difference is halved, the required sample size will quadruple (Fayers and Machin, 1995)

$$n_A = \frac{2\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{(0.5d_S)^2} = \frac{2\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{0.5^2 d_S^2} = \frac{2\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{0.25 \times d_S^2}. \quad (2.2)$$

Rearranging this equation gives

$$4 \times n_A = \frac{2\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{d_S^2}, \quad (2.3)$$

This highlights the sensitivity of the sample size to d_S, as discussed in chapter 1. A consideration for parallel group trials is when the trial is cluster randomised and we have a cluster effect. This results in a slightly different sample size formula and calculation. This topic is discussed in further detail in Appendix A.5.

The sample size calculation is based on the population variance estimate; therefore the Normal distribution values can be used. However, when a trial has been conducted and the data collected, the population variance σ^2 is considered to be unknown and the sample variance estimate from the trial, s^2, is used instead. As a result of this, the Normal distribution and by extension the Z- values cannot be used so the t-distribution and t-values are used (Brush, 1988; Chow et al., 2002; Senn, 1993). This gives us the following equation, where n_A is the smallest integer value that satisfies it

$$n_A \geq \frac{(r+1)(Z_{1-\beta} + t_{1-\alpha/2, n_A(r+1)-2})^2\sigma^2}{r d_S^2}. \quad (2.4)$$

However, it is noticeable that this equation does not give a direct estimate since n_A appears on both sides of the equation. The best method to deal with this is to re-write the equation in terms of the power and solve using an iterative technique

$$1 - \beta = \Phi\left(\sqrt{\frac{r n_A d_S^2}{(r+1)\sigma^2}} - t_{1-\alpha/2, n_A(r+1)-2}\right). \quad (2.5)$$

Here, $\Phi(.)$ is the cumulative density of a Normal distribution. When the sample variance is being used instead of the population variance, Senn describes how instead of using the Normal distribution, the power should be estimated from the non-central t-distribution with $n_A(r+1) - 2$ degrees of freedom and a non-centrality parameter $\sqrt{\frac{r n_A}{(r+1)}}$ (Senn, 1993). This is due to the power being estimated under the alternative hypothesis, which states $d \neq 0$ therefore the corresponding t -distribution would be non-central. The two distributions (t- and $Normal$-) are very similar, with the t-distribution being slightly fatter than the Normal distribution (Julious, 2010b). The equation above can then therefore be rewritten as follows, using the non-central t-distribution

$$1-\beta = 1 - T^{-1}\left(t_{1-\frac{\alpha}{2},n_A(r+1)-2}, n_A(r+1)-2, \sqrt{\frac{rn_A d_S^2}{(r+1)\sigma^2}}\right), \tag{2.6}$$

where $T^{-1}(\dots)$ is the cumulative density function of a non-central t-distribution. To allow for the Normal approximation to the non-central t-distribution a small correction factor can be added to (2.1) so it better approximates to (2.6) as follows (Guenther, 1981; Julious et al., 1999):

$$n_A = \frac{(r+1)(Z_{1-\beta}+Z_{1-\alpha/2})^2\sigma^2}{rd_S^2} + \frac{Z_{1-\alpha/2}}{4}. \tag{2.7}$$

There are a number of quick results which are shown in Appendix A.2. These are conservative estimates for the sample size calculations for superiority, crossover, non-inferiority and equivalence trials.

Table 2.2 shows how the sample size changes as the standardised difference, $\frac{d}{\sigma} = d$, changes for a parallel group, two-armed, superiority trial. It can be seen that as the effect size δ increases, the sample size required decreases rapidly. If $\delta = 0.05$, the required sample size is 8407, compared to if $\delta = 0.10$ when the sample size is 2103. Again it is clear that the sample size has fallen four-fold when the effect size has doubled. This shows that the sample size used in a trial is extremely sensitive to slight changes in the target or target difference.

2.3.1.2 Crossover Trials

In order for us to be able to estimate a sample size for a crossover trial, we need to first estimate the within-subject standard deviation, σ_w. This can be extracted from the residual line of the analysis of variance (ANOVA) model; it evaluates the variation which occurs through repeated measures on the same patient. The within-subject variability from the ANOVA model is directly related to the variability around the difference, σ_d, from a paired t-test as $\sigma_d^2 = 2\sigma_w^2$. The sample size can be calculated using the within-group standard deviation and the effect size, the sample size calculation can be attained by a similar method as in the parallel group study (Guenther, 1981),

$$n = \frac{2(Z_{1-\beta}+Z_{1-\alpha/2})^2\sigma_w^2)}{d_S^2} + \frac{Z_{1-\alpha/2}}{2}, \tag{2.8}$$

	Allocation Ratio (r)			
δ	1	2	3	4
0.05	8407	6306	5605	5255
0.10	2103	1577	1402	1314
0.15	935	702	624	585
0.20	527	395	351	329
0.25	338	253	157	211
0.30	235	176	115	147
0.35	173	130	89	108
0.40	133	100	70	83
0.45	105	79	57	66
0.50	86	64	47	53

Table 2.2: Sample size requirements for one group, n_A (where $n_B = rn_A$), with various standardised differences ($\delta = \frac{d_S}{\sigma}$) and allocation ratios, r, for a parallel group trial. This table has been calculated for a 90% power and a two sided type I error rate of 5%. These sample sizes are calculated from the non-central t-distribution. (Julious and Campbell, 2012; Rothwell et al., 2018a)

where n is the total sample size. Similar to parallel group trials, an additional factor of $\frac{Z_{(1-\alpha/2)}}{2}$ can be added to allow for the approximation to the Normal distribution. The non-central t-distribution result with $n-2$ degrees of freedom and non-centrality parameter $\sqrt{\frac{nd_S^2}{2\sigma_w^2}}$ is given by (Senn, 1993)

$$1-\beta = 1 - T^{-1}\left(t_{1-\alpha/2,n-2}, n-2, \sqrt{\frac{nd_S^2}{2\sigma_w^2}}\right). \tag{2.9}$$

As in the superiority parallel group case, Table 2.3 shows how the sample size varies as the standardised difference changes for the superiority crossover case. If we compare Table 2.2 and Table 2.3 (Julious and Campbell, 2012; Rothwell et al., 2018a), we can see that the sample sizes are similar for $r = 1$ in Table 2.2. This confirms that the formulae and the results are similar. The results are similar irrespective of whether the trial is parallel group or crossover design. Tables 2.2 and 2.3 also show how rapidly the sample size decreases as the standardised difference increases, highlighting the sensitivity of d_S. As the target difference increases the sample size decreases.

δ	n
0.05	8408
0.10	2104
0.15	936
0.20	528
0.25	339
0.30	236
0.35	174
0.40	134
0.45	106
0.50	87

Table 2.3: The total sample size (n) for a crossover study for various standardised differences ($\delta = \frac{d_S}{\sigma_w}$) with 90% power and a two sided type I error rate of 5%. These sample sizes are calculated from the non-central t-distribution.

2.4 Binary Data and Other Trial Objectives

The sample size calculations for non-inferiority and equivalence trials are similar to those described earlier in this chapter, see Appendix A.1. Whilst for the superiority design there is a target difference which the trialists' are aiming to reach, for the non-inferiority and equivalence designs there is a non-inferiority or equivalence margin. This equivalence/non-inferiority margin is the limit which, by showing the treatment difference is less than, we can conclude the treatments are clinically the same. Often these limits are defined in reference to d_S.

The calculations for binary outcome measures for all the trial designs mentioned here are similar to those discussed in this chapter; they are all extremely sensitive to the target difference, therefore emphasising the importance of its careful determination. More details regarding the different trial designs and binary outcome measures can be found in the book "Handbook of statistical methods in randomised controlled trials"(Rothwell et al., 2018a).

2.5 Chapter Summary

This chapter has demonstrated the sensitivity of the target difference d_S for all trial designs and outcome measures. Due to the sensitivity of the target difference parameter, d_S, it is imperitive to have a good estimate. The sample size tables in this chapter have shown that as the target or estimated effect size increases, the sample size decreases dramatically which highlights the sensitivity of d_S once again. Further calculations can be found in the Appendices (A) for extensions from the parallel group, continuous case discussed in this chapter.

2.6 Discussion

Throughout the work completed for this chapter, it has been shown that there are many different sample size calculations and trial designs used. These calculations are all quite similar, extensions of the parallel group, superiority case. As a result of this, it was decided to focus on parallel group superiority trials. One reason for this is that crossover trial calculations are an extension of those for the parallel group case.

As mentioned earlier, σ_w^2 is the variance for the crossover trial, which as a function of the variance of a parallel group trial, σ^2, can be written as $\sigma_w^2 = \sigma^2(1 - \rho)$. In this expression, ρ is the correlation between the two measures in period 1 and period 2 (Julious, 2010b). Owing to this difference in the variance, and by extension the standard deviation, for the same mean difference the standardised difference will be quite different for parallel group and crossover trials. Due to this large difference, crossover trials are excluded from the review of randomised controlled trials in chapter 4. This is also justified by the nature of crossover trials themselves being restricted by therapeutic area; some areas such as oncology are not suitable for crossover trials. It is preferable to include as many clinical areas as possible.

Also, many Health Technology Assessment (HTA) publicly funded trials are parallel group superiority trials, therefore these shall be the focus of this thesis (Rothwell et al., 2018b).

Another aspect which needs discussion is estimands. These are similar to parameters of interest in a trial, however they also take the study population into account (Council, 2010). An addendum to the ICH E9 guidance was commissioned in 2014 (ICH-E9, 2014). These estimands will impact on the analysis population. Whilst this is an area of interest and does impact the research questions in this thesis, they will not be researched in the scope of this PhD.

As has been mentioned throughout the thesis thus far, the estimated difference d_S is the most sensitive part of the sample size calculation. As briefly discussed in chapter 1, there could be an inherent bias introduced when moving from one trial to the next, such as when moving from Phase II to Phase III trials in industry, or pilot/systematic review to main trial in academia. Chapter 3 will review the methods used to elicit the target effect size, d, and will move on to chapter 4 which will aim to investigate which of the methods discussed in chapter 3 are reported in the Health Technology Assessment reports.

3. Review of Methods for Quantifying the Target Difference

3.1 Introduction

As discussed in chapters 1 and 2, the target difference is the most sensitive part of a typical sample size calculation. The methods of choosing the target difference have been discussed for a number of years, recently becoming quite a popular topic amongst statisticians and trialists (Beaton et al., 2002; Cook et al., 2014; Wu et al., 2011). The importance of the target difference is due to the effect it has on the sample size calculation, which in turn impacts the cost and the elapsed time of a trial as the more people you need the more expensive a trial generally is to run. The purpose of this chapter is to set the scene for the next chapter, to provide definitions and background which is useful for subsequent chapters.

3.1.1 Chapter Aims

This chapter aims to discuss the various methods which are used to elicit the target effect size. A scoping review of the literature was undertaken, which shows the different methods, as well as the advantages and disadvantages of using each method of elicitation.

3.2 Background to Quantifying the Effect Size

Research has been done to determine how trialists and clinicians are choosing the clinically meaningful difference for their trials including the DELTA review published in 2014 which describes the various methods used to choose the minimum clinically important difference (MCID), as well as providing advice for the different methods (Cook et al., 2014). This review shall be discussed in further detail in this chapter. As discussed briefly in the previous chapter, the most sensitive part of the sample size calculation is the anticipated effect size. It is primarily at the discretion of the trialists what this value takes, with some choosing the actual target difference between the treatments, and some choosing the minimum clinically important difference (MCID), so the smallest change that will be worthwhile for the patients or lead to change in clinical practice (Jaeschke et al., 1989). Other similar names for this difference include the sufficiently important difference (SID) and the target difference or the minimum important difference (MID) (Barrett et al., 2005a). This section will focus on the target difference, denoted d. This is to be defined as the target effect size the trial is powered to detect.

A description of the various terms used to describe the target difference and their meanings can be found in section 1.3.1. An illustration of how the target difference, MCID and detectable difference are associated is shown in Figure 3.1. This diagram shows that the target difference, which is the difference the trial is powered to observe, is all-encompassing and the MCID can be the same or smaller than this. The detectable difference is smaller than both of these other differences as it is the smallest difference which will result in a statistically significant P-value.

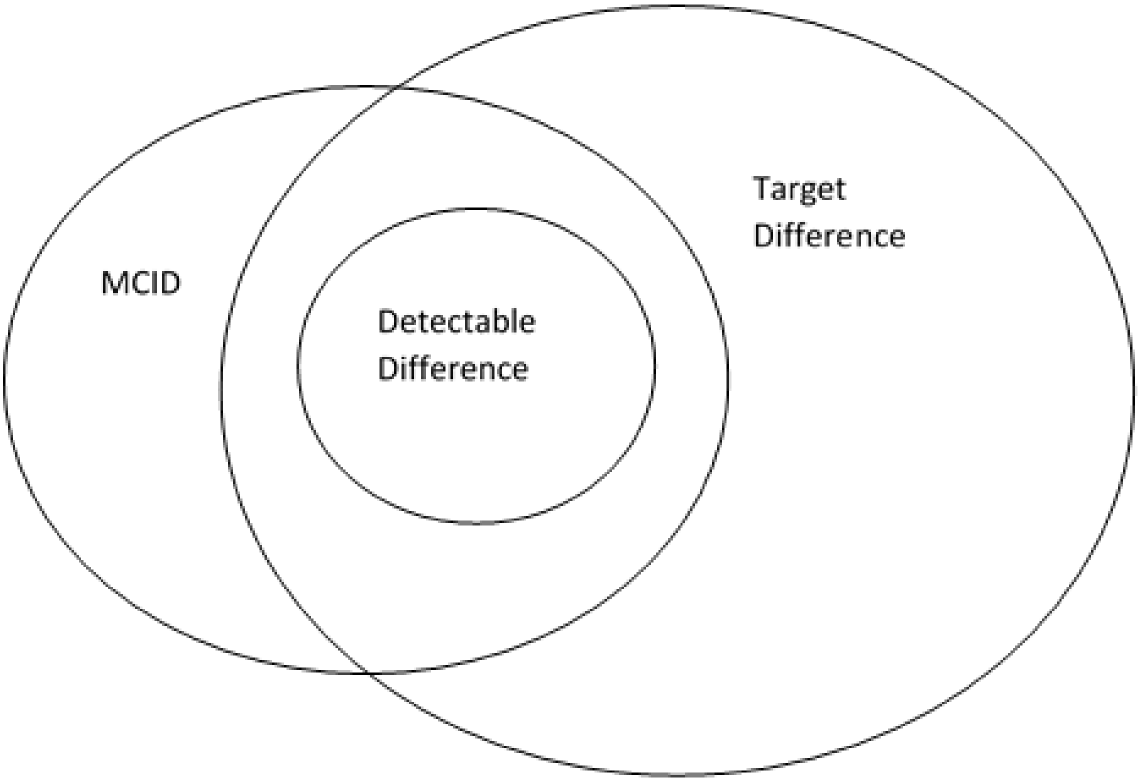


Figure 3.1: Illustration of the relationship between the target difference, the MCID and the detectable difference.

3.2.1 Quantifying an Effect Size

The estimated difference is commonly called the minimum clinically important difference (MCID), as this terminology encompasses not only the estimate or target difference for a new treatment, but also the difference which will cause an impact on the patient population; a difference that is deemed clinically meaningful. This was first described by Jaeschke, though the definitions have broadened since (Jaeschke et al., 1989). The original DELTA review identified seven commonly used methods for choosing an important and/or realistic difference. These are as follows: (Cook et al., 2014)

1. The anchor method,
2. The distribution method,
3. The health economic method,
4. The opinion-seeking method,
5. The pilot-study method,
6. The review of evidence base method,
7. The standardised effect size method.

The anchor, distribution, health economic and standardised effect size methods can be used to specify an important difference, whilst the pilot study method can be used for specifying a realistic difference. The opinion-seeking method and the review of evidence base can be used for specifying a realistic or important difference, or both simultaneously. Each of these methods has its own advantages and disadvantages, which shall be discussed briefly here.

A Medical Research Council (MRC) funded project is currently researching this in more detail, to which this PhD work has contributed. The DELTA^2 project is currently underway and these categories mentioned above shall form a recurring comparison throughout this thesis. More details about the DELTA^2 project and how this research fits in is given in Section 3.2.4.

3.2.1.1 Anchor Method

This method takes the form of two parts. The first part is establishing the anchor by calculating a mean change in score for patients who have expressed that a minimum clinically important difference or change has occurred in the context of quality-of-life measures (Jaeschke et al., 1989; Zhang et al., 2015). The patients who have expressed that they have experienced a change in their quality of life are then asked to quantify that change. This would not be asked of those patients who did not express that they had observed a change. This change in their quality of life measure can then be evaluated and used as a clinically important difference in future trials using the same outcome measure.

The second part is to implement the MCID found in the first part. The MCID will change depending on the measure being used. So for any subsequent study which is being performed on a similar population, all those patients who express a change in their quality of life greater than the MCID are shown to have improved (assuming the changes are positive). This MCID can also be used in sample size calculations when using that particular outcome measure in a trial. Another example is to consider the scenario where previous studies indicated if patients were able to walk 100 yards after a particular surgery, they would have reached the target change. This would be a clinical anchor based on retrospective data. The anchor questions can be posed solely to the patient or the clinician, or both if it is of interest to assess the agreement between patient and clinician.

Another variation of this method is to 'anchor' a new outcome measure to a previously used outcome measure, when both measures are correlated (DeRogatis et al., 2009; Khanna et al., 2009). This works by mapping a known validated measure to a new measure, enabling direct comparison between the current and new measures. An example of this would be trying to implement a new QoL measure or subscale, and anchoring it to a generic QoL questionnaire.

3.2.1.2 Distribution Method

The distribution method uses the imprecision value of the measurement in question (how reliable is the measurement) and results in the MCID being a value which is larger than this imprecision value, therefore being likely to represent a meaningful difference (Wyrwich et al., 1999). A common approach is to use test-retest data for an outcome (Cook et al., 2014). This can help specify the size of the difference due to random variation in the measurement of the outcome.

3.2.1.3 Health Economic Method

This method endeavours to take into account not only the MCID, but also the cost of the treatment and any other factors which are deemed to be important when deciding whether to run a trial. This method aims to establish a threshold value which is deemed acceptable for the cost per unit increase in health (Torgerson et al., 1995). It estimates the relative efficiency of the treatments which can then be compared directly. This method is not commonly used in practice, with the 13 papers which used this method to establish the MCID using hypothetical data sets (Cook et al., 2014). The focus is on the health economics, as the name implies, of a treatment rather than the MCID alone, including data on the costs in terms of harms (side-effects or adverse events) and financial aspects of the treatment (Torgerson et al., 1995). One trial, published in 2011 which investigated the cost-effectiveness of yoga for lower back pain did use a cost-effectiveness ratio from a previous trial to determine the sample size (Tilbrook et al., 2011).

3.2.1.4 Opinion-seeking

This method is perhaps one of the more intuitive methods, based on determining a value, or a range of values, for the clinically meaningful difference by asking clinicians or experts in the relevant fields to provide a professional opinion. From another angle, the experts could be patients who have been suffering from a particular illness, as they would have a better idea of what they deem a minimum clinically important (or realistic) difference (Allison et al., 2010; Barrett et al., 2005b). Another method is to use a combination of clinicians and patients (McAlister et al., 2000; Stone et al., 2004).

3.2.1.5 Pilot Study

A pilot study is a small version of the trial which is being planned (Hulley, 2013; Thabane et al., 2010a). It is conventionally used to assess the feasibility of the main trial, though information can be collected for use in the sample size calculation such as the effect size and population standard deviation (Julious, 2010b; Salter et al., 2006). The effect size observed in a pilot study can be used as a starting point to help determine the MCID (Cook et al., 2014). With regards to trials of complex interventions, pilot trials are normally used to make inferences about whether the true value of the effect size will lie, which in turn can be used in the development of stop/go criteria. The important point about these studies is that they are usually small and are not usually aiming to test a formal hypothesis. However, since there are only a small number of people involved, the estimates of the standard deviation and effect size will be imprecise (Samsa et al., 1999; Wang et al., 2006). This method is commonly used but not often reported (Cook et al., 2014).

3.2.1.6 Review of Evidence Base

This method collates and summarises all the existing evidence about the treatment or disease in question to allow researchers to choose an important or realistic difference based on previous trials and research (Johnston et al., 2009; Thomas et al., 1997; Woods et al., 2001). The optimum method used to do this is meta-analysis (Cook et al., 2014), however trialists should be wary of possible publication bias. Publication bias occurs when trials which are significant get published and therefore cited more, this results in a skew of positive results for published work. Another problem with this method is that the effect size from another study may not be applicable to the current trial due to differences in the study population or treatments

(Barbui et al., 2000; Cranney et al., 2001; Ravelo et al., 2009). It has been reported that these methods are commonly used, more recently in quality-of-life measures (Cocks et al., 2011).

3.2.1.7 Standardised Effect Size

The standardised effect size is scale-invariant, which means that it can be generalised across a variety of clinical areas, it has no units of measurement (Cook et al., 2014). For continuous outcomes, this is calculated by taking the difference in means and dividing by the pooled standard deviation (Cohen, 1988). If we let the difference between the two groups be d, and the pooled population standard deviation be σ, the standardised effect size (δ) can be calculated as $\delta = \frac{d}{\sigma}$.

The size of the standardised effect is used to establish whether an important difference has occurred, which are conventionally 0.2 for a small effect, 0.5 for a moderate effect and 0.8 for a large effect (Cohen, 1988). The benefits of this method are that it is simple to calculate and allows for comparisons across different outcomes, trials, populations and disease areas (Cook et al., 2014). The sample size formula can be rewritten in terms of δ as oppose to d.

3.2.2 Discussion of Methods for Determining Target Difference

The following table details some comments and possible problems which are associated with the various methods of elicitation of the target effect size (Cook et al., 2014). These are based on the DELTA review and are not absolute.

These comments have been raised in the DELTA document and further discussed at the DELTA2 workshop in Oxford, 2016. However, whilst some of these will be discussed in more detail throughout the thesis, not all will be covered in the scope of this research.

Method	**Comments**
Anchor Method	• Patient- or clinician-centric so subjective in nature, • Other factors like length of illness come into play, • Useful in quality of life studies (Crosby et al., 2003; Hays and Woolley, 2000; Wyrwich et al., 1999).
Distribution Method	• Not commonly used but can be used alongside anchor method (Copay et al., 2007; Oxberry et al., 2012; Wu et al., 2011), • Difficult to establish what is meant by 'important', which can result in underestimating the MID (Guyatt et al., 1987)
Health Economic Method	• Speculation that patients and clinicians are against modelling techniques involving cost as a factor (Cook et al., 2014), • These methods can require strong assumptions and are complex in terms of modelling (Cook et al., 2014) • The HTA reports health economics as co-primary
Opinion-seeking Method	• Surveys most common method but low response rates or missing data common (Cook et al., 2014) • This method is likely to produce realistic differences by clinicians and important differences by patients (Jaeschke et al., 1989)
Pilot Study Method	• Example of method could be using pilot to imitate trial • Focus primarily on feasibility of trial, not assessing intervention. • Could be used to assess response rates for a questionnaire or survey.

Method	**Comments**
Review of Evidence Based Method	• This method could be used to estimate target effect size based on prior information • This is investigated further in chapter 5
Standardised Effect Size Method	• Different combinations of values can result in the same standardised effect size (Cook et al., 2014) • One of the most common methods found in the DELTA review and can be observed in a number of different research areas (Hill et al., 2008; Kazis et al., 1989)

3.2.3 Further Discussion of Methods Determining and Reporting the Target Difference

One part of the original DELTA review involved sending out a survey to establish what methods were being used currently and what factors influenced the choice of method. The results confirmed that there was some degree of reverse-engineering in choosing the target or target difference (Cook et al., 2014). In these cases, there was already a set number of participants available and the sample size calculations were re-arranged to select a target effect size in order to result in that sample size. However, this is not disclosed in the trial protocols although a number of responders in the survey admitted to having done this (Cook et al., 2014). This could be due to the availability of funding, time or participants themselves if the outcome was rare. This method of determining the sample size is known as a convenience sample size calculation (Kraemer et al., 2006).

The initial review resulted in a second publication by the same authors (Cook et al., 2015) which formed a set of guidelines for researchers when choosing their target effect size. However, the article focused primarily on the improvement needed in reporting the justifications used to choose a target difference.

Relevant to this subject area is the difference between statistical significance and clinically meaningful changes. Statistical significance is used to describe the likelihood that the results are due to chance (Fethney, 2010; Goodman, 1999). Clinical significance describes the practicality of the results of a study, whether the results show a clinically important effect (Jacobson et al., 1984). Many studies in the past focused solely on statistical significance, without considering the difference in a clinical context (Cocks et al., 2008; Johnson et al., 2013), with the use of point estimates and confidence intervals being criticised for not including the clinically meaningful effect (Fethney, 2010; Kieser and Hauschke, 2005). The intense concentration of high-impact journals on statistically significant results with P-values less than 0.05 is well known and results in publication bias (Dickersin et al., 1987; Ioannidis et al., 2014; Tressoldi et al., 2012). However, P-values and the test statistics from which they arise are affected by the sample sizes being used in the study (Hedges, 2008). So P-values are useful in informing how reliable a difference between treatments is, but they do not provide any information on the size of the effect observed (van Tulder et al., 2007).

The issues which surround basing the choice of the target difference on clinical or patient opinion stem from the highly subjective nature of this method. The most important consideration of this method is the perspective of the group designated with setting the minimum clinically important difference (Beaton et al., 2002). Patients who have been experiencing discomfort or pain for a long period of time may have the opinion that a larger difference is needed to be clinically important compared to patients who have only been ill for a short period of time, or compared to the clinician (Beaton et al., 2002).

Clinicians will have differing opinions to patients, as patients will see the difference from an extremely personal perspective whereas clinicians are more likely to be pragmatic about it; there could be an issue with the opinion not being generalizable (Allison et al., 2010). An example of this method being subjective is if the outcome measure of interest is number of days in recovery from a cold, one would expect there would need to be a larger difference to be clinically significant for the patient compared to an outcome measure related to survival from cancer, where even the smallest improvement in survival time may be considered to be clinically significant from the perspective of the patient.

Using information from early phase trials in the sample size calculation for the current trial results in a biased estimate of the target effect size or target difference (Kraemer et al., 2006). This bias is the results from the theory that a test should not be based on the results from another test, so the fact that the second trial or confirmatory trial would not take place without a promising pilot or phase II trial implies that this will automatically lead to biased results. These issues will be built upon in chapter 5. The target effect size will also change depending on the type of trial being performed, whether it is a public health study or a pharmaceutical trial and the outcome of interest.

It has been debated that researchers are often overly optimistic when choosing a target difference for their sample size (Fayers and Machin, 1995; Friedman, 2010). If this is the case, the observed effect size would be vastly different from the predicted effect size used in the sample size calculation. This, in turn, leads to the study being under-powered (Friedman, 2010). However, consideration must be placed on establishing a method to estimate which null effects ($P > 0.05$) are truly non-effective interventions and which are due to the trial being underpowered. If a trial is underpowered then there still could be a true effect, and indeed the observed effect could be the true effect, where there just was not enough participants to get a statistically significant result. On the other hand, if the intervention is not effective, then any target effect size would be deemed overly-optimistic and the methods used to get that target would need to be investigated. Since one can never truly know what the true effect is, we need to develop a protocol to follow to estimate the proportion of true "null" effects.

Different methods are more appropriate for different contexts. Contexts, in this sense, are whether the difference is important or realistic. The box below shows which methods can be used to specify the differences by context, along with Figure 3.2.

- Important Difference
  - The anchor method
  - The distribution method
  - The health economic method
  - The standardised effect size method
- Realistic Difference
  - The pilot-study method
- Important or Realistic Difference
  - The review of evidence base method
  - The opinion-seeking method

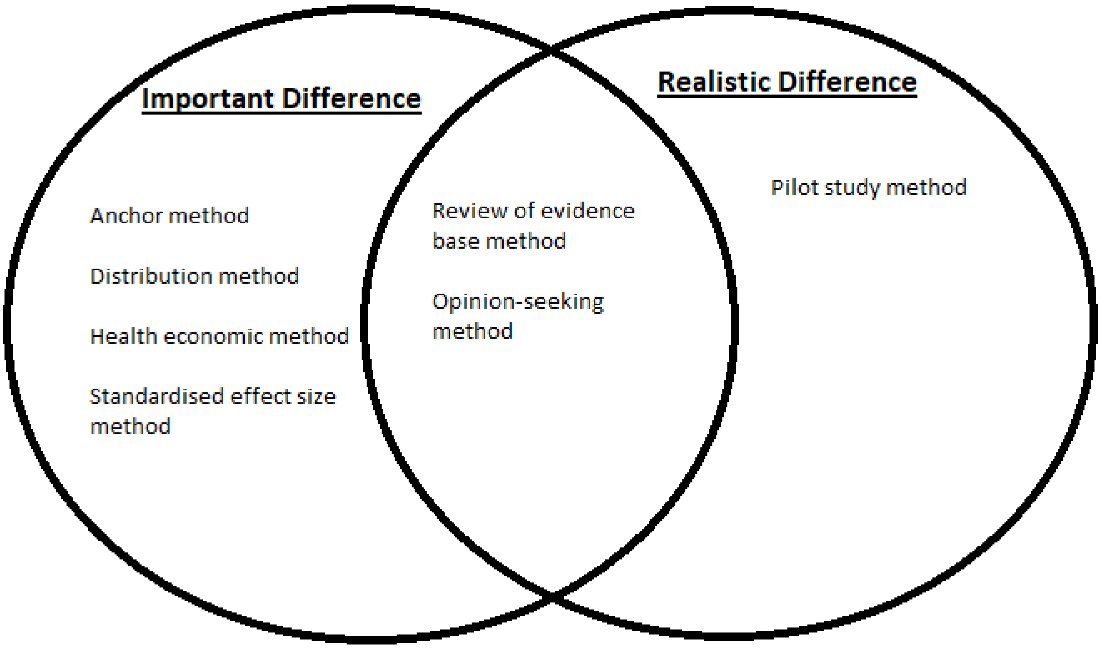


Figure 3.2: Illustration of methods of quantification of the target difference by context.

3.2.4 DELTA^2 Project

The DELTA^2 project is a research collaboration funded by the Medical Research Council. It concentrates on developing a guidance document for trialists, researchers and statisticians about eliciting the target difference. This project started after the commencement of the PhD and JCR became part of the research team. It included the opportunity to present work at conferences and workshops and allowed the collaboration with experts in this field of research from other institutions as well as industry. The work in this thesis forms part of the research used in this guidance document and JCR has assisted in writing the document. Being part of this reseach project has allowed JCR the opportunity to travel for conferences and the work to be presented in Oxford, Montreal, London and Baltimore (sadly JCR was unable to attend the conference in Baltimore) throughout the course of the PhD.

3.3 Chapter Summary

This chapter has described the most common methods for determining the estimated effect size, or target difference, d. All the methods have advantages and disadvantages, however only some of the methods are based on previous work. These include the anchor method, the distribution method, the pilot study method, the review of the evidence base method and it has been shown that the health economic method can be based on previous work (Tilbrook et al., 2011). However, the health economic method is a more specialised area of statistics which is outside the scope of this research. In the context of two powered up studies in sequence, if the current trial is being based on previous information this would imply that the previous work was shown to be promising. If trials were unbiased, the second trial would be performed regardless of the results of the previous trials. Therefore, it would seem that there could be a form of bias when basing one trial on the results of others, which will be further investigated in chapter 5. This leads us to question whether this bias is ever adjusted for or accounted for, which will be discussed in chapter 4. To investigate this further, a review of published randomised controlled trials from the Health Technology Assessment programme is completed and discussed in chapter 4. This review demonstrates the most commonly used and reported methods of elicitation.

4. Review of Reports by the Health Technology Assessment

4.1 Introduction

In previous chapters it was highlighted that there are a number of different ways to elicit a target difference for a sample size calculation. In chapter 2 it was demonstrated how sensitive a standard sample size calculation is to the target difference for a variety of trial designs. Due to this sensitivity, it is important to consider carefully the chosen target difference. This chapter will explore what are currently the most common methods used to estimate the target effect size in Health Technology Assessment (HTA) funded trials.

A major funder of research into clinical interventions in the UK and the most integrated clinical research system in the world is the National Institute of Health Research (NIHR) (of Health Research, 2010c), and the largest programme within that is the Health Technology Assessment (HTA) (of Health Research, 2010a). The HTA programme funds commissioned and researcher-led health related research including randomised controlled trials of clinical interventions in the UK (of Health Research, 2010a).

In order to answer the research questions described earlier in chapter 1, we should consider the current state of reporting of clinical trials. In particular, it would be interesting to establish what particular methods of elicitation for the target effect size are being used.

As mentioned in chapter 3, if trials are being designed and powered using observed effect sizes from previous research, there could be a bias introduced similar to regression to the mean. This would result in the observed effect size being considerably less than expected, even if the treatment worked. This bias occurs when using the results from one trial to design the next.

It was reported in the DELTA publication (Cook et al., 2014) that the methods of elicitation which were commonly used could be broadly categorised into seven distinct groups, which are described in brief detail in chapter 3. It would be of interest to explore whether the methods used do fall into these broad categories and to explore whether there are some of these methods which are more commonly applied.

The seven DELTA elicitation categories are

- Anchor method;
- Distribution method;
- Health-economic method;
- Opinion-seeking method;
- Pilot study method;
- Review of evidence base method;
- Standardised effect size (SES) method.

4.1.1 Chapter Aims

The aim of this chapter is to investigate, through a review of the Health Technology Assessment reports, what methods are most commonly being used and reported to quantify the target effect sizes. This will provide the groundwork for which methods will be best to focus on for the simulations of common trial designs. Chapter 6 will focus on those designs which are more often reported to be able to provide the most useful advice for trialists. It is also of interest to know whether particular methods are more common for various clinical areas or outcome measures.

As stated, the primary objective of this part of the research is to establish which methods of elicitation for the target difference are most commonly used. This is to show that if many reports indicated that previous research was used to estimate a target effect size, there may be a significant problem as in these cases the sample size calculation and resulting trial could be biased due to regression to the mean.

4.2 Methods

In order to fully assess the current methods used to elicit the target effect size, a review of the HTA journal reports (of Health Research, 2010b) was conducted. The HTA journal was chosen because it is one of the largest funders in the UK (of Health Research, 2010a). A criteria upon receiving funding from the HTA is that the trial is published in the HTA journal, which means that the HTA publish all trials it funds, irrespective of the statistical significance achieved. This reduces the likelihood of publication bias and justifies the use of a single journal in this review. Some of these trials are also published in other leading journals, such as the Lancet, the New England Journal of Medicine and the British Medical Journal. Strict inclusion criteria is implimented which is primarily based on the discussion from chapter 2, focusing on parallel group, superiority trials. Based on personal correspondence with a statistician in the clinical trials unit, 95% of trials in Sheffield are publicly funded by the NIHR so this research is of local importance.

4.2.1 Scoping Review

Initially a general scoping review was undertaken for a recent volume of the journal. This was done to assess the number of eligible trials and refine the inclusion and exclusion criteria. The scoping review used volume 18, published in 2014. The inclusion criteria for this review were that the trial was a randomised controlled trial. Reports were excluded if they focused on systematic review or pilot, observational or cohort studies. Details of the inclusion of reports can be seen in Figure 4.1. It can be seen that on the initial assessment of this volume there were 12 trials which were eligible for inclusion. It was assumed that if this was representative of the average number for each volume, it was estimated that there would be 216 reports in the final review. However, it was expected that due to time contraints this would not be feasible to complete within the time of the PhD, so it was decided to limit the inclusion to a 10 year period from 2006 to 2016, section 4.2.2 describes the criteria decisions in more detail. This resulted in a potential 120 reports meeting the eligibility criteria.

4.2.2 Final Inclusion and Exclusion Criteria for the Full Review

The inclusion criteria for the full review were as follows:

- Randomised controlled trial (RCT)
- Parallel-group, superiority trial design
- Not a cluster trial

Superiority trials with a parallel group design are the simplest and most common type of trial performed. The selection process consisted of a number of stages; the titles of journal reports were initially read to establish relevance and reports which

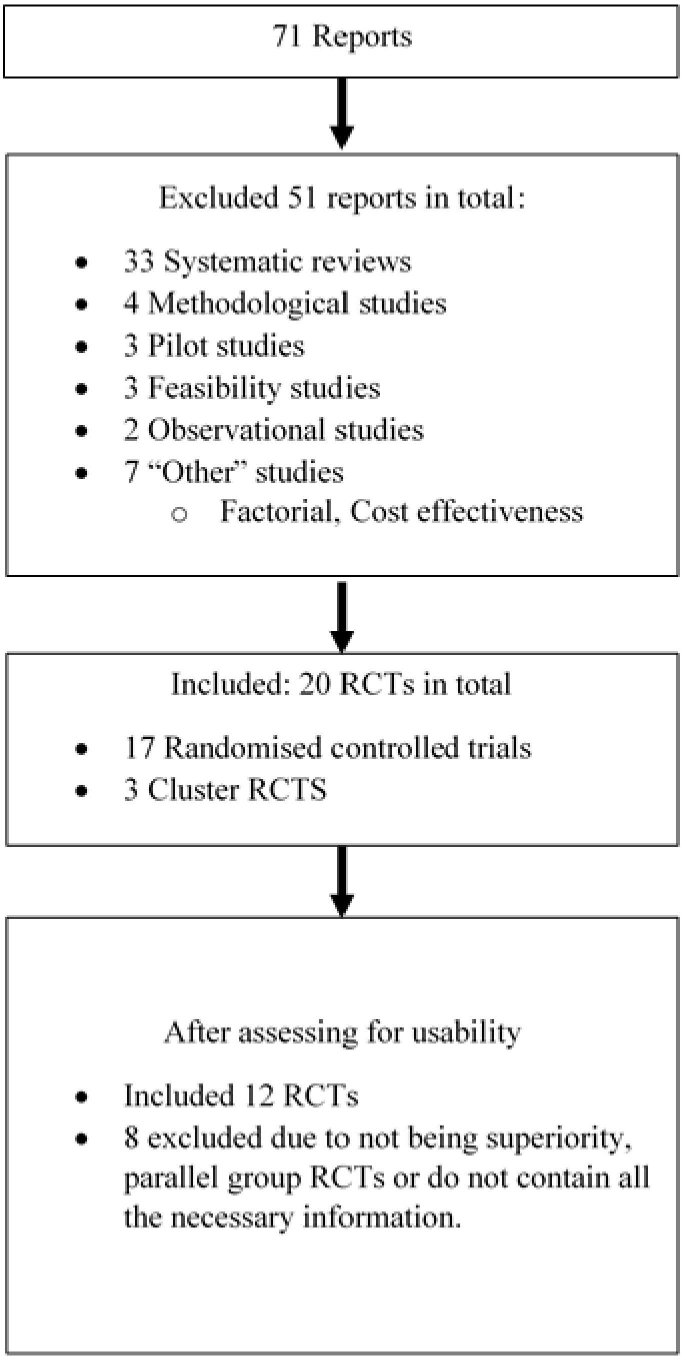

71 Reports
Excluded 51 reports in total:
33 Systematic reviews
4 Methodological studies
3 Pilot studies
3 Feasibility studies
2 Observational studies
7 "Other" studies
Factorial, Cost effectiveness
Included: 20 RCTs in total
17 Randomised controlled trials
3 Cluster RCTS
After assessing for usability
Included 12 RCTs
8 excluded due to not being superiority, parallel group RCTs or do not contain all the necessary information.

Figure 4.1: A flow chart to show inclusion of reports from Volume 18 in scoping review.

were not RCTs were excluded. The titles were read initially but if this was not sufficient to determine eligibility then the abstract was read as well to inform the decision.

The exclusion criteria consisted of excluding

- Pilot, feasibiliity or observational studies
- Systematic reviews
- Cost-effectiveness studies
- Methodological reports
- Vaccination trials

4.2.2.1 Reasons for Exclusion

Reports were excluded if, during the title and abstract reading phase, their focus was a systematic review, feasibility study, cost-effectiveness study, pilot study or the trial being observational as opposed to a RCT. Methodological reports were also excluded for this reason. During the data extraction phase it became apparent that for some trials which were RCTs it proved difficult to extract the correct information. An example of this is vaccination trials, which had multiple primary end-points, making data extraction challenging owing to the number of primary end-points and results provided.

Trials with more than 3 arms tended to be more complex due to the number of comparisons being made. It was difficult to decipher the sample size justifications for these trials due to missing information or there being more than one primary outcome of interest. This finding resulted in these types of trials becoming exclusion criteria.

Trials with non-inferiority or equivalence objectives were also excluded due to the target difference not being specified for these trials. Instead, a non-inferiority or equivalence margin is given to determine the required sample size, as described in chapter 2. The hypothesis for these types of trial are slightly different to that for superiority trials as they are not aiming to show a particular size of difference, rather that they are aiming to show the difference is the same, almost the opposite. Hence, these trials were excluded as well as crossover trials. Whilst the crossover sample size calculation is similar to the parallel group sample size calculation, as discussed in chapter 2, the observed effect sizes which are reported can be different. Also discussed in chapter 2 that the standardised effects are larger for crossover trials since the variance is the pooled variance σ_w^2, not the individual, σ^2. Due to time constraints it was decided to exclude these reports. Finally, trials which implemented a factorial design were excluded due to the sample size justification being too complex for extraction and the information required for extraction being incomplete or missing. The trials which did contain the relevant information tended to have co-primary endpoints which made extraction more difficult as there were multiple target effect sizes, and the sample size justifications were unclear which one was used as the initial sample size target effect size.

It was determined that a time frame should be included in the selection criteria due to the consuming nature of a review this large, as well as there being only one data extractor (JCR). Since the CONSORT guidelines were implemented in 2010 for HTA reports, it was discussed that it may be useful to compare the reports from pre- and post-CONSORT. Therefore, reports were excluded if they were published prior to 2006 and after 2016, which left 10 years' of reports to review. This was confirmed to be a justified choice at the start of extraction since there were very few RCTs published in early volumes, and those which were published did not contain enough information for full extraction.

4.2.3 Data Extraction

The data was extracted using a series of Microsoft Excel spreadsheets. This was first completed for title and abstract reading, then a new workbook was used for the data obtained from reading the full report. All categorical variables were coded prior to the completion of data extraction and a record of the coding kept separately to the extraction sheet.

The data extraction was completed by one person (JCR) over a period of 9 months. A full list of variables which were extracted can be found in Appendix B.1, which includes:

- Study title, corresponding and lead author
- Publication year, Volume, Issue and ISRCTN numbers
- Trial design detail (including multicentre, randomised, type of trial)
- Clinical area
- Trial population of interest and setting
- Detail about the primary end-point, intervention and control
- Target, achieved and evaluable sample sizes
- Target power and significance level
- Target difference, elicitation method, MCID?

- Observed treatment differences and effect size
- P-value, 95% Confidence interval
- Further comments

Any particularly interesting phrases or comments about the results of the trials were documented in the extraction sheet as free text; good and bad examples of reporting were also recorded for the possibility of discussion. Free text was also used for any further details about each variable. The interesting case studies have been discussed later in the chapter.

4.2.4 Categorisation

A number of variables could be classified into several categories. In the event of a categorical variable which was subjective in nature, such that it was unclear to JCR which of the categories they fitted into best, advice was sought. This occured for two of the variables, the elicitation method used and the clinical area. Each of these were managed in a different way.

4.2.4.1 Elicitation Categories

As discussed in chapter 3, the original DELTA project formed seven broad categories for methods frequently used to determine the target difference which are

- Anchor method
- Distribution method
- Health economic method
- Opinion-seeking
- Pilot study
- Review of evidence base
- Standardised effect size (SES).

Whilst these categories are not absolute, they do condense the wide variety of methods into manageable units. The various methods of elicitation could have been categorised in two different ways. The first being in categories designed by JCR. These categories were based on JCR's experience of the whole range of approaches identified up to that point and provided more detail about the previous research which had been used in elicitation. During data extraction, JCR attended a DELTA2 consensus workshop in Oxford ($27^{th} - 28^{th}$ September 2016) and presented the intial results. In this presentation of the results, JCR used the elicitation categories she had devised herself (as mentioned above). The workshop delegates advised that it would be useful for JCR to develop the categorisation process so that her categories were reduced in number and mapped onto the DELTA categories to enable direct comparison. The two sets of categories are shown in Table 4.1.

Dissertation Defined Categories	**DELTA Categories**
Previous research - (excluding pilot study, systematic review, meta-analysis and Cochrane review) Systematic Review Cochrane Review Meta-analysis	Review of Evidence
Pilot	Pilot
Expert consensus	Opinion-Seeking
SES	SES
Other (including Anchor, Distribution, Health Economic methods)	Anchor Distribution Health Economic Other
No mention	No mention

Table 4.1: The two sets of elicitation categories, demonstrating how they match up.

The original categories provided more detail about the review of evidence and previous research, which was needed for another aspect of the research. However, from an analytical perspective it was decided to use the DELTA categories. These categories are useful since they allow direct comparison with the findings of the DELTA survey, where trialists were asked which methods they commonly used to elicit the target effect size. Therefore, both were included in the final data extraction.

It was also discussed in the DELTA report that a mixture of methods was commonly used (Cook et al., 2014). This resulted in the development of a Mixed category as well. One category which needs further explanation is the Previous Research category in the dissertation defined categories. This category included the use of previous published research which was not consistent with the other categories of systematic review, meta-analysis, Cochrane review or pilot studies. This included observational or cohort studies, or one previous trial which the trialists' used as a reference.

4.2.4.2 Clinical Categorisation

The opinion of a clinical doctor, Dr Zhe Hui Hoo, was sought to assist with putting the clinical areas documented in the reports into categories. The initial categories were set by JCR, however as she has no clinical background she struggled to identify and categorise some of the conditions into clinical areas. The clinical areas themselves are based on the disease or condition being assessed. Prior to clinical guidance from Dr Hoo, there were 15 categories in total with a considerably large "Other" category. After clinical guidance this had increased by 5 categories, resulting in a comprehensive list of clinical categories all of which are commonly recognised clinical areas and can easily be differentiated.

	Categories	
Initially Used	**Medically-Advised**	**Post-hoc Condensed**
Mental Health	Mental Health	Mental Health
Oncology	Oncology	Oncology
Orthopaedics	Orthopaedics	Orthopaedics
Obstetrics and Gyn.	Obstetrics and Gyn.	Obstetrics and Gyn.
Cardio-vascular	Cardio-vascular	Cardio-vascular
Gastro-intestinal	Gastro-intestinal	Gastro-intestinal
Respiratory	Respiratory	Respiratory
Stroke	Stroke	Stroke
Diabetes	Diabetes	Diabetes
Dermatology	Dermatology	Dermatology
Immunology	Immunology	Immunology
Neurology	Neurology	Neurology
Paediatrics	Paediatrics	Paediatrics
Haematology	Haematology	Other
Primary Care	Primary Care	
Other	Emergency Care	
	Renal/Urology	Renal/Urology
	Geriatrics	Geriatrics
	Critical Care	Critical Care
	Lifestyle	Lifestyle

Table 4.2: The two sets of clinical categories, demonstrating the changes at each stage of extraction.

Table 4.2 shows the various clinical categories used at each point of extraction. The clinical advice was useful for cases where the reviewer (JCR) was unsure which category to put some trials in if they had more than one possible category. For example, for a trial investigating childhood asthma, it was advised to categorise this as paediatrics as opposed to respiratory. One category which could constitute a multitude of other clinical areas is Primary Care. This category was not used frequently, however it specifically referred to trials which directly involved GP practices and care-pathways involving the GP.

Once extraction had been completed it became apparent that three clinical categories were only used for one reported trial, so these were condensed back down post-hoc to form a considerably smaller "Other" category. This is detailed in Table 4.2.

4.2.5 Analysis

Descriptive statistics and plots were used, with a variety of summary statistics being used dependent on distribution shape. Means and standard deviations were used for data which were Normally distributed, whilst median and inter-quartile ranges were used for any variables which were skewed or count data.

4.2.5.1 Standardisation of Effect Sizes

It was anticipated that a wide variety of measures would be reported, and these would depend on the outcome of interest and method of analysis used in the trial. It was assumed that these may include mean differences, regression coefficients, odds ratios, relative risks or difference in proportions. A discussion was held between JCR and her supervisors on how to make the results comparable. It was agreed that the results should all be presented on the same scale. One way to do this is to use a scale-independent measure of effect, such as the standardised effect size. Based on the variables which were to be extracted, some manipulation of equations could be completed in order to standardised all the effect sizes, both the observed and target.

The observed effect sizes were standardised using two methods to confirm accuracy, the first method by using the extracted observed effect size and confidence interval to calculate the test statistic, as shown in Table 4.3. The second method simply took the inverse of the P-value as shown in Equation 4.1. If only a P-value was provided then the second method was used alone.

$$\delta_{observed} = \Phi^{-1}\big(P - value\big) \times \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \tag{4.1}$$

where $\delta_{observed}$ is the standardised observed effect size, $\Phi^{-1}(\dots)$ is the Normal inverse of the P-value and n_i is the sample size in arm i.

These results were compared and then used to assess the observed effect sizes in various clinical areas, as well as allow a direct comparison of observed and target standardised effect sizes.

The target effect sizes were calculated by re-arranging the sample size formula (as shown in Equation 4.2) using the target sample size, the target power and significance level, all provided in the report.

Observed Effect Type	**Z-statistic Calculation**	**Re-arrangement for Standardised Observed Effect**
Mean Difference	$Z = \frac{d}{SE(d)}$	$\delta_{observed} = Z \times \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$
Difference in Proportions		
Regression Coefficient		
Absolute Risk Reduction		
ANOVA/ANCOVA coefficients		
Odds Ratio	$Z = \frac{ln[OR]}{SE(ln[OR])}$	
Risk Ratio	$Z = \frac{ln[RR]}{SE(ln[RR])}$	
Hazard Ratio	$Z = \frac{ln[HR]}{SE(ln[HR])}$	

Table 4.3: The methods used to estimate the standardised observed effect size.

$$\delta_{target} = \sqrt{\frac{2 \times \left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2}{n_A}} \tag{4.2}$$

where δ_{target} is the standardised target effect size, $Z_{1-\beta}$ and $Z_{1-\alpha/2}$ are the Normal values for the power and significance of the trial, respectively, and n_A is the sample size per arm.

These standardised values were assessed using simple summary statistics and plots to provide a visual display of the results.

4.2.5.2 Minimum Detectable Difference

Another aspect of the standardised differences which can be analysed is the *minimum detectable difference* (Wu et al., 2011). This is further detailed and utilised in chapter 7. The minimum detectable difference is the difference that is the smallest difference under the set parameters of the trial which is detectable. This can be calculated by setting the power to be 50%, such that $Z_{1-\beta} = 0$ in the sample size equation 2.1. Doing this shows the minimum value that the confidence interval around the point estimate will exclude the null value.

The minimum detectable difference (MDD) can be used to show that if a trial recruits the required target sample size, thereby achieving the planned power, the difference which is needed to observe a P-value of 0.05 is $0.6d$ for 90% power and

Year	Volume	Frequency
2016	20	20
2015	19	19
2014	18	12
2013	17	11
2012	16	8
2011	15	6
2010	14	8
2009	13	10
2008	12	2
2007	11	3
2006	10	8
Total		107

Table 4.4: Frequencies of included trials by year and volume.

$0.7d$ for 80% power. These results are used in sub-analysis (See section 4.3.5.1) to determine whether the observed effect size is larger than the detectable difference.

4.3 Results

4.3.1 Report Selection

In total there were 994 reports available between 1998 and 2016, after restricting the publication dates to 2006 – 2016 this became 684. Once the initial exclusion criteria were imposed (excluding systematic reviews, feasibility trials, observational trials and methodology reports) this was reduced down to 175 reports which were taken forward to full reading.

Of the 177 reports which were taken forward, 75 were excluded which resulted in 102 being included in the analysis, corresponding to 107 randomised controlled trials. This was due to there being 5 reports which documented 2 trials. Two of these trials consisted of 3 arms, where comparisons were made between control and each of the intervention arms. The remaining three reports had two separate 2 armed studies in the same clinical area but with different interventions. Since the reporting of each trial was separate, it was decided to consider these trials as separate also. Therefore, for the remainder of this chapter the results will be presented by trial as opposed to by report. The reasons for excluding 73 of the reports was that they were of a complex trial design. These included factorial, non-inferiority, equivalence or cluster trials, as well as those reports which were not RCTs. It is of note that 11 of the reports were excluded due to not containing enough information or data. Table 4.4 shows the number of RCTs included for each of the volumes between 2006 and 2016 and Figure 4.2 shows the inclusion of trials at each point of assessment.

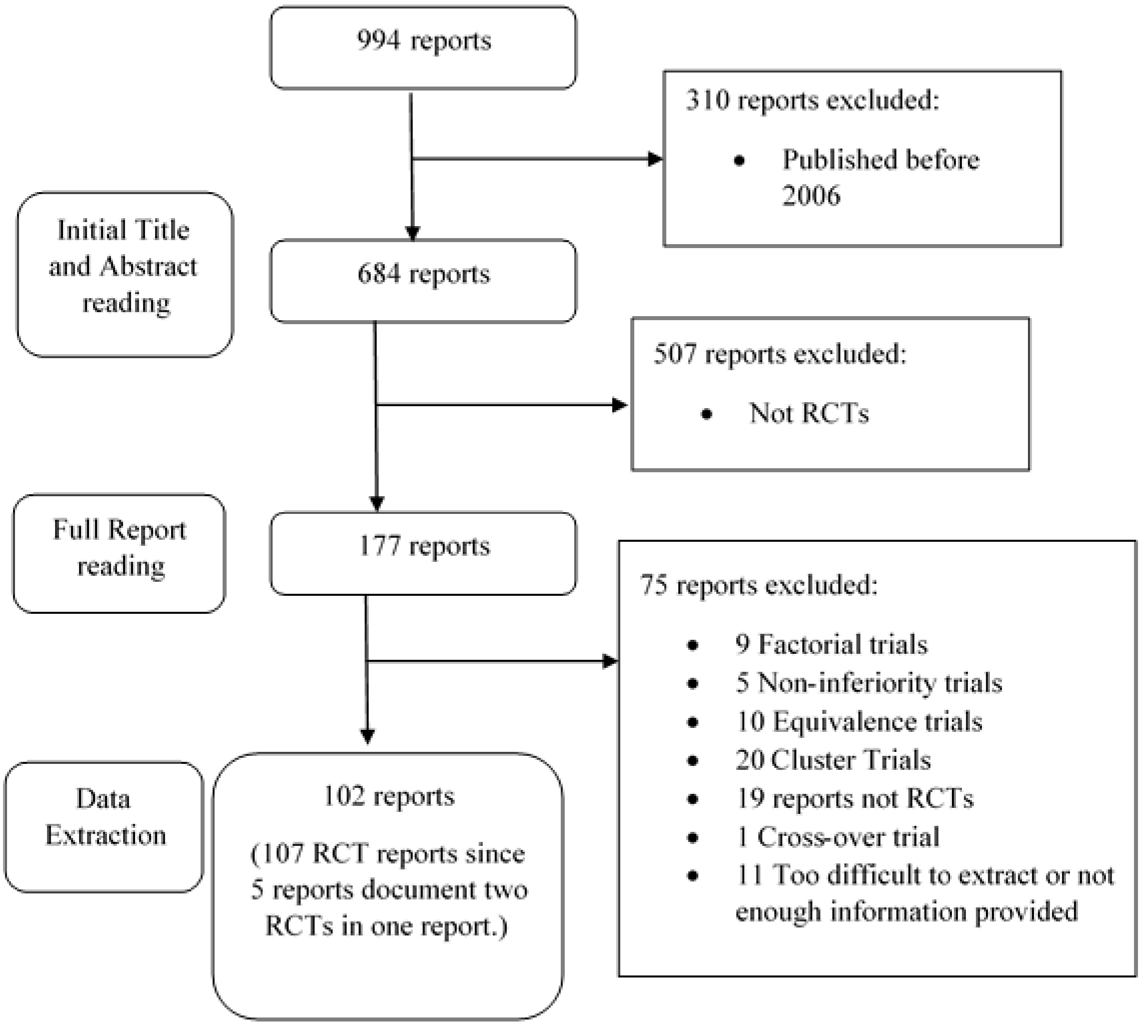


Figure 4.2: A flow chart to show inclusion of reports.

4.3.2 Study Characteristics

The reports which were included were all superiority parallel group trials. As stated there were 5 reports which documented multiple RCTs, each of the RCTs were documented separately if it was possible to distinguish clearly between the trials. Figure 4.3 shows the various clinical areas which were selected. These were broadly assigned during data extraction with additional relevant details noted. The categories were confirmed by a clinician once extraction was complete.

For example, if considering a trial investigating asthma treatments, it was categorised as respiratory and confirmation was sought once data extraction was complete. This was particularly useful for disease areas like sleep apnoea, which were intially categorised as “Other” but subsequently categorised by the clinician as respiratory.

Table 4.5 shows the frequencies and percentages of study characteristics. There are a large number of trials reported in the area of mental health in the past 10 years by the HTA, with 16% of included reports being in mental health. This was followed closely by cardiovascular (9.5%) and paediatrics (8.6%). 49% of reports were categorised as trials taking place in hospital settings, with 23% in primary or secondary care. This category included whole NHS trusts and specialist clinics. The majority of included reports were two-arm trials (78%) and had standard care (79%). Further plots can be found in Appendix B.

4.3.3 Elicitation of Target Effect Size

Figure 4.4 shows there are nearly 47% of reports used a review of evidence to estimate the target effect size. It was noted in a free text entry that all trials which reported using mixed elicitation methods all used review of evidence as part of the combination of methods. The reports which were categorised as “Mixed” methods all used the review of evidence in combination with other methods. Combining review of evidence methods, mixed methods and pilot methods demonstrate that any use of previous research resulted in 57.2% of trials within this combined category. This shows that since the frequency of trials using previous research to base their sample size calculations or justifications is so high, without an adjustment being implemented then a large proportion of trials could be inherently under-powered and observe an effect size which is less than expected, even if there is a true treatment effect present.

We can see the relationship between the elicitation methods used and the clinical area, as shown in Figure 4.6. Mental health and cardiovascular show the greatest proportion of reports which use the review of evidence (8.3% and 6.5% respectively).

Characteristic		**Count**	%
Clinical Area	Cardiovascular	11	10.3%
	Critical Care	2	1.9%
	Dermatology	9	8.4%
	Diabetes	3	2.8%
	Gastrointestinal	9	8.4%
	Geriatrics	2	1.9%
	Immunology	2	1.9%
	Lifestyle	5	4.7%
	Mental Health	18	16.8%
	Neurology	4	3.7%
	Obstetrics and Gynaecology	2	1.9%
	Oncology	4	3.7%
	Orthopaedics	6	5.6%
	Paediatrics	9	8.4%
	Renal/Urology	6	5.6%
	Respiratory	7	6.5%
	Stroke	5	4.7%
	Other	3	2.8%
Number of Arms	2	84	78.5%
	3	23	21.5%
Setting	Hospital	51	47.7%
	General Practice	16	15.0%
	Mixed	7	6.5%
	Community	6	5.6%
	Primary/Secondary Care	25	23.4%
	Other	1	0.9%
	Missing	1	0.9%
Intervention	Drug	20	28.0%
	Therapy	41	38.3%
	Surgical	11	10.3%
	Education	2	1.9%
	Complex	5	4.7%
	Other	18	16.8%
Control Type	Active	85	79.4%
	Placebo	22	20.6%
Primary Endpoint Measure	Continuous	49	45.8%
	Proportion	41	38.3%
	Time to Event	10	9.3%
	Count	4	3.7%
	Other	3	2.8%

Table 4.5: Frequencies for study characteristics.

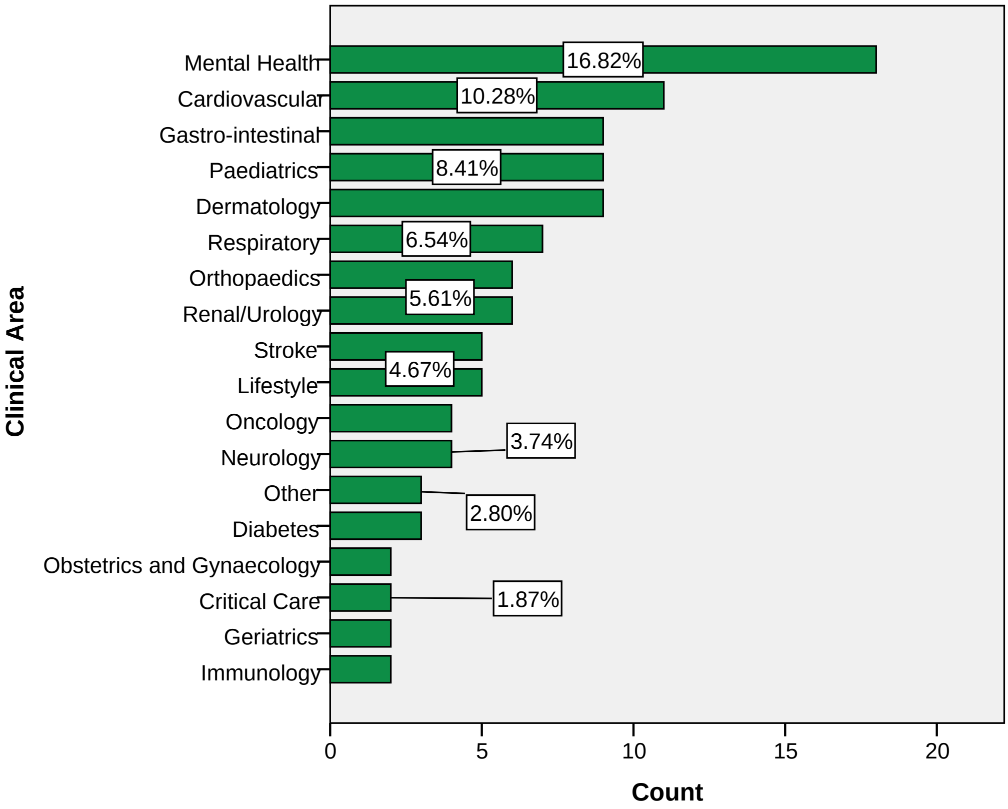


Figure 4.3: Frequency of RCTs categorised by clinical area.

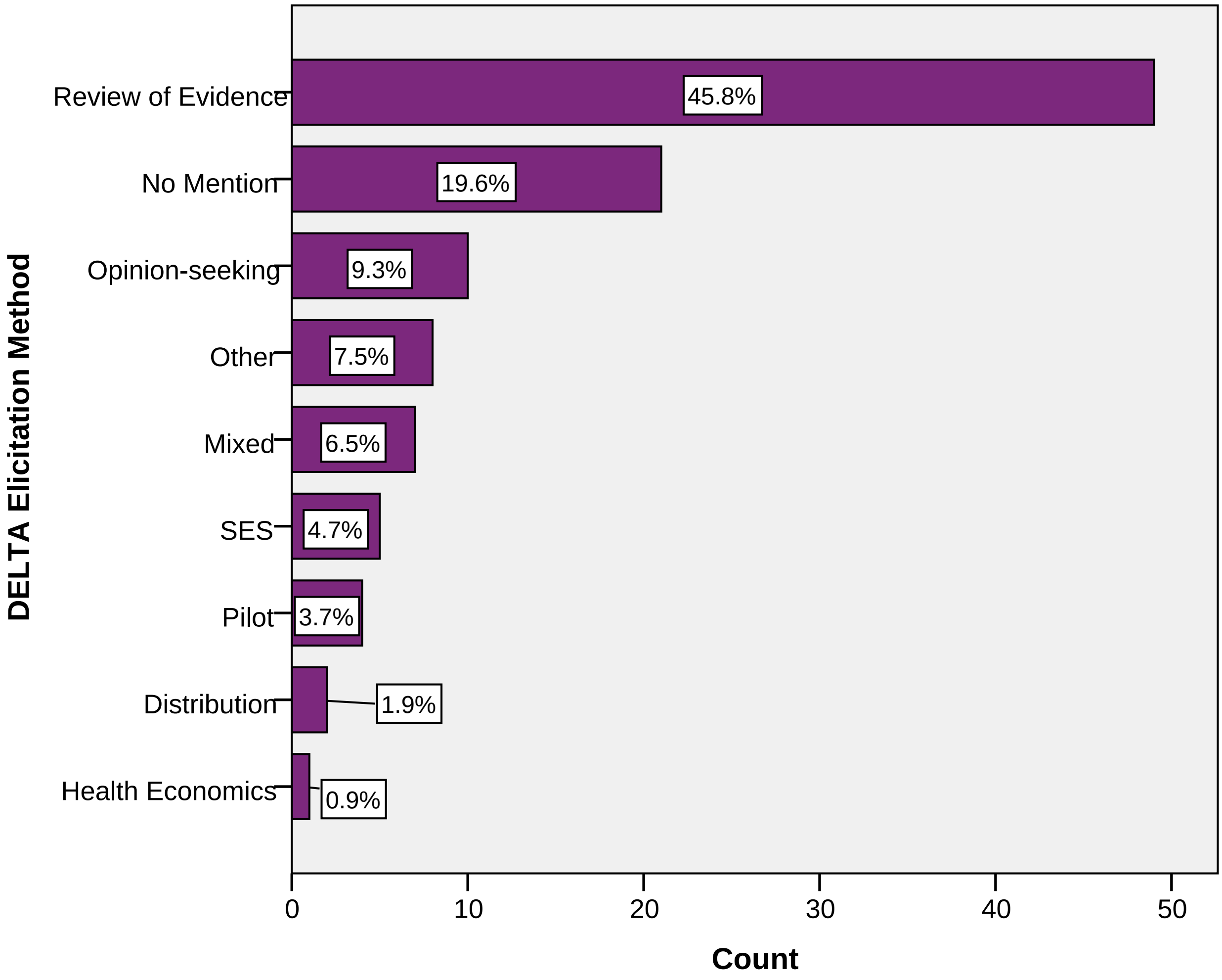


Figure 4.4: Frequencies of trials by elicitation methods.

		DELTA Elicitation Method									
		Anchor	**Distribution**	**Health Economics**	**Opinion Seeking**	**Pilot**	**Review of Evidence**	**SES**	**Mixed**	**No Mention**	**Other**
Clinical Area	Mental Health				2		**9**	1		4	2
	Oncology		**1**			**1**	**1**			**1**	
	Orthopaedics						**2**	**2**	1	1	
	Obs. and Gyn.				**1**			**1**			
	Cardiovascular		1				**7**			2	1
	Gastrointestinal				1		3		1	**4**	
	Respiratory				2		**3**			2	
	Stroke						1		1	**2**	1
	Diabetes				1		**2**				
	Dermatology			1	**2**		**2**		**2**	1	1
	Immunology						**2**				
	Neurology					1	**3**				
	Lifestyle					**2**	**2**		1		
	Paediatrics				1		**5**	1		1	1
	Renal/Urology						**4**			2	
	Geriatrics						**1**		**1**		
	Critical Care						**1**				**1**
	Other						**1**			**1**	**1**

Table 4.6: Elicitation methods for each clinical area with the **most commonly used** highlighted. Empty cells were zero counts.

4.3.4 Target and Observed Effect Sizes

Another aspect of this research was to establish whether particular clinical areas were over- or under-estimating their target effect size. If the estimated effect sizes are standardised, we are able to see the variation across the different clinical areas. Whilst this variation could be attributed to the different interventions and outcome measures, it is still of interest. The estimated effect sizes were standardised by rearranging the sample size formula to give the standardised effect size using the target power, sample size and significance levels.

Table 4.7 shows the average estimated and observed standardised effect sizes for each clinical area and overall, both for all trials and the statistically significant trials. These were calculated as described in section 4.2.5. It can be noticed that for all trials the average estimated effect size is around 0.30, which indicates that those trialists' who seek to observe a "moderate-large" effect size (using the definitions by Cohen (Cohen, 1988)) of greater than 0.6 are being unrealistic in their trial design. Recall that the boundaries set by Cohen are 0.2 for a small effect, 0.5 for a moderate effect and 0.8 for a large effect. For the statistically significant trials ($N = 35, 32.7\%$), the results are still slightly skewed as the mean and median are not similar. The median target standardised effect size is 0.31 compared with the median observed effect size of 0.34. This is as expected, since the observed effect size would by similar to the target in order to reach statistical significance.

	All Results					Statistically Significant Trials Only				
		Target SES		Observed SES			Target SES		Observed SES	
Clinical Area	**Count**	**Mean**	**Median***	**Mean**	**Median***	**Count**	**Mean**	**Median***	**Mean**	**Median***
Cardiovascular	11	0.228	0.171	0.133	0.050	2	0.113	0.113	0.236	0.236
Critical Care	2	0.151	0.151	0.016	0.016	0				
Dermatology	9	0.343	0.368	0.111	0.061	0				
Diabetes	3	0.308	0.316	0.220	0.166	1	0.431	0.431	0.446	0.446
Gastro-intestinal	9	0.306	0.295	0.311	0.343	5	0.363	0.360	0.487	0.410
Geriatrics	2	0.290	0.290	0.331	0.331	1	0.261	0.261	0.220	0.220
Immunology	2	0.509	0.509	0.432	0.432	1	0.258	0.258	0.432	0.432
Lifestyle	5	0.295	0.300	0.118	0.065	1	0.433	0.433	0.243	0.243
Mental Health	18	0.360	0.332	0.227	0.165	6	0.430	0.386	0.346	0.358
Neurology	4	0.330	0.270	0.177	0.056	1	0.577	0.577	0.596	0.596
Obstetrics and Gyn.	2	0.252	0.252	0.341	0.341	1	0.299	0.299	0.628	0.628
Oncology	4	0.256	0.255	0.139	0.143	1	0.212	0.212	0.273	0.273
Orthopaedics	6	0.349	0.331	0.160	0.164	4	0.359	0.344	0.220	0.237
Other	3	0.180	0.180	0.041	0.041	1	0.051	0.051	0.041	0.041
Paediatrics	9	0.338	0.362	0.359	0.230	5	0.414	0.410	0.584	0.543
Renal/Urology	6	0.271	0.296	0.095	0.019	1	0.114	0.114	0.070	0.070
Respiratory	7	0.248	0.229	0.128	0.099	3	0.263	0.251	0.247	0.241
Stroke	5	0.263	0.284	0.109	0.133	1	0.145	0.145	0.028	0.028
Overall	107	0.302	0.300	0.190	0.112	35	0.336	0.309	0.363	0.343

Table 4.7: Summary statistics for the standardised estimated and observed effect sizes. **SES = Standardised Effect Size.** * Median used due to skewed observed effect size distribution.

4.3.4.1 Minimum Clinically Important Difference

Only 20 reports made a direct comment about either using or not using the minimum clinically important difference (MCID), with 25% of those reports stating they were using the MCID as the target effect size. Figure 4.5 shows that whilst there appears to be a linear trend between the target and observed effect sizes, there seems to be no association between whether the target effect size was based on the MCID and the size of the target and observed effect sizes.

There is also no obvious association between the use of the MCID and the significance of the trial, as seen in Figure 4.6. However, since there were so few trials which reported this detail the initial conclusions are limited.

4.3.4.2 Clinical Areas

It can be seen in Figure 4.7 shows the large variation which occurs, as well as the median target effect size being 0.300. There is considerable grouping around certain values for different clinical areas. There appears to be a cluster around 0.4 for dermatology, whereas the mental health cluster is slightly lower at around 0.30. Cardiovascular is also exhibiting a clustering around the lower end of the scale. The scale of this plot ends at 0.8 which indicates that none of the trials expected an effect size larger than that. If we compare this to the Cohen boundaries (Cohen, 1988), then none of the trials expected a large effect size.

Figure 4.8 shows the standardised observed effect sizes for each clinical area along with the average for each clinical area. There are differing extents of variation between clinical areas.

The difference between the standardised target and observed effect sizes can be seen in Figure 4.9. This plot shows that there are some extreme differences, yet the majority lie around 0.1 which is a small difference.

In Table 4.8 it can be observed that the pre- and post-2010 standardised effect sizes are slightly different both overall and when split by clinical area.

4.3.4.3 Standardised Target and Observed Effect Sizes

If these standardised target effect sizes are compared to those observed, as shown in Figure 4.10, it is clear to see that it is rarely the case where the target and observed are similar. The significant results can also be seen on this plot. The line on the plot indicates the point where the target and observed effect sizes are equal. The blue points to the left of the line are trials where the observed effect size was larger than the target, yet the trial remained not significant. The green points to the right

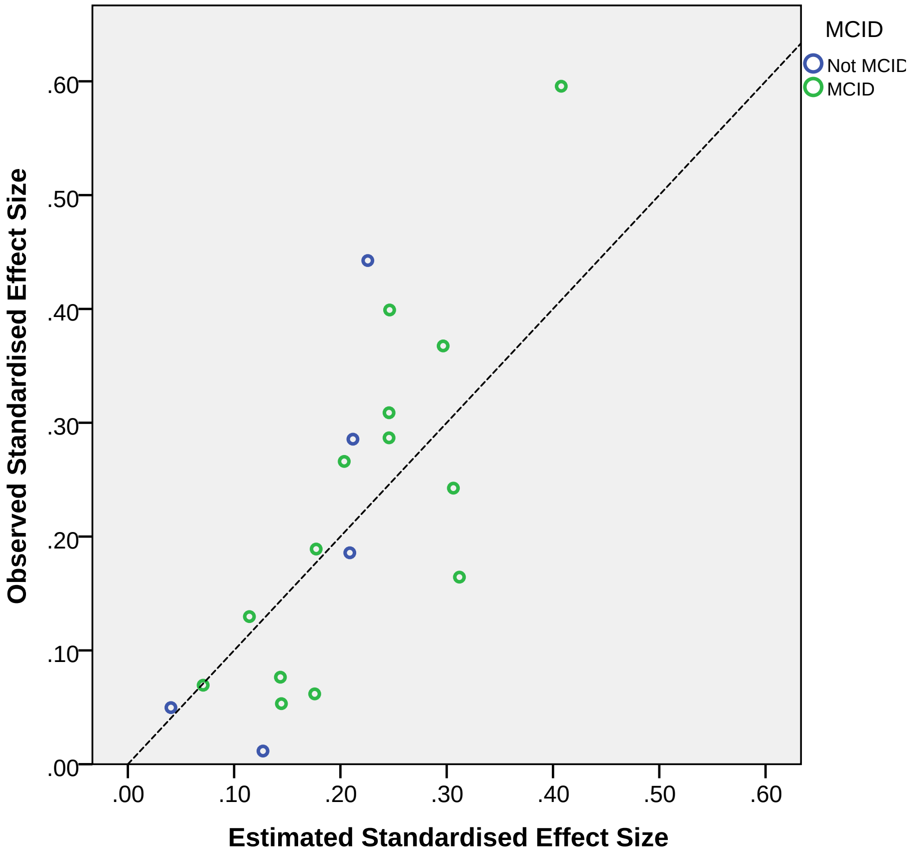


Figure 4.5: The standardised estimated and observed effect sizes split by whether the estimated was the MCID or not.

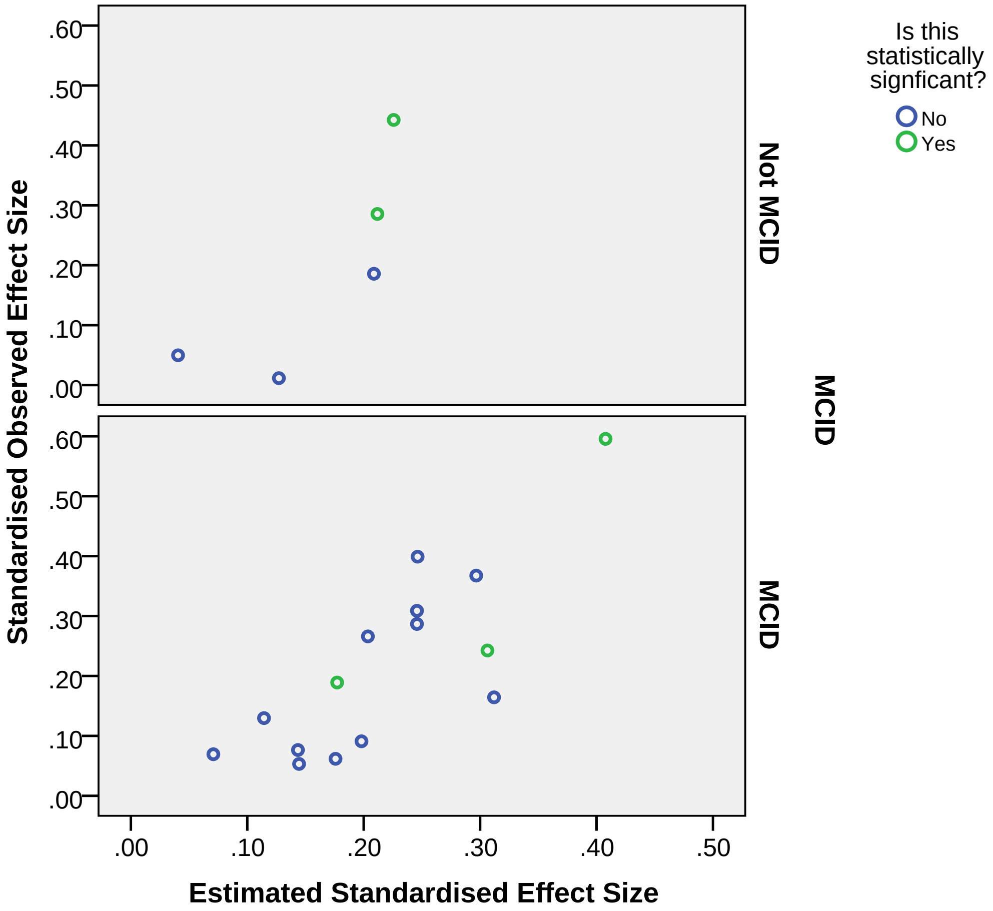


Figure 4.6: The standardised observed effect sizes split by whether the estimated was the MCID or not and significance of the trial.

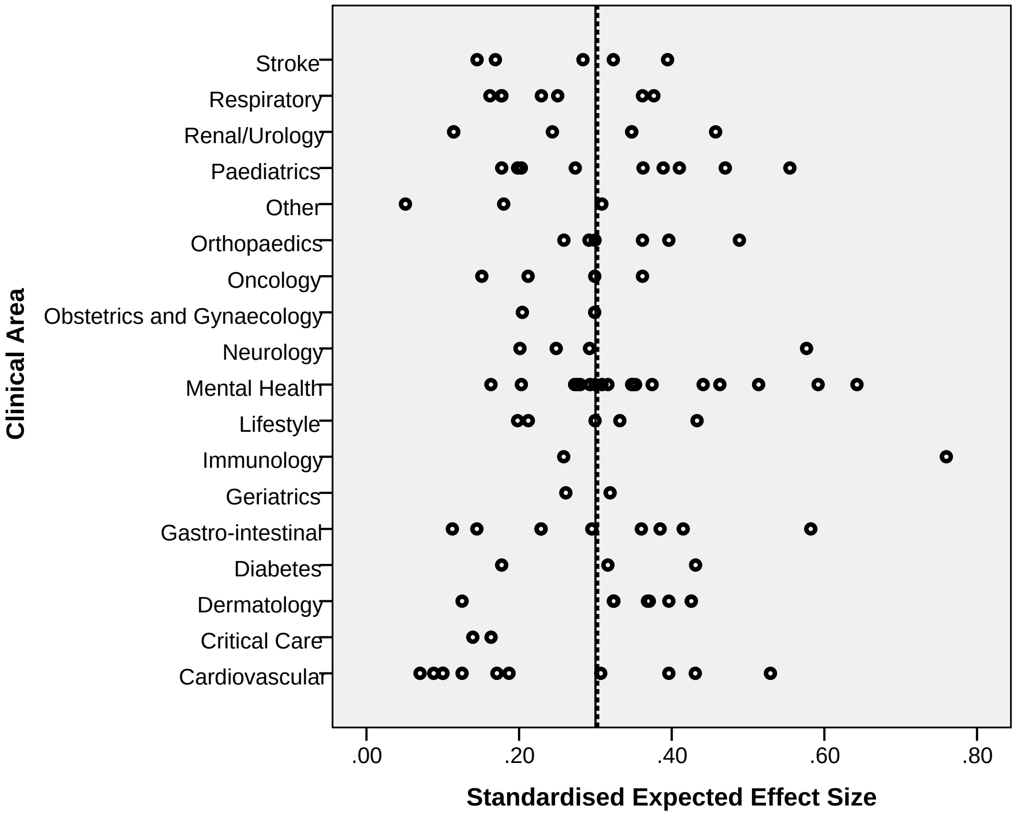


Figure 4.7: The standardised estimated effect size used in the sample size calculation across various clinical areas. The dashed line indicates the mean estimated effect size (0.302) and the solid line represents the median (0.300).

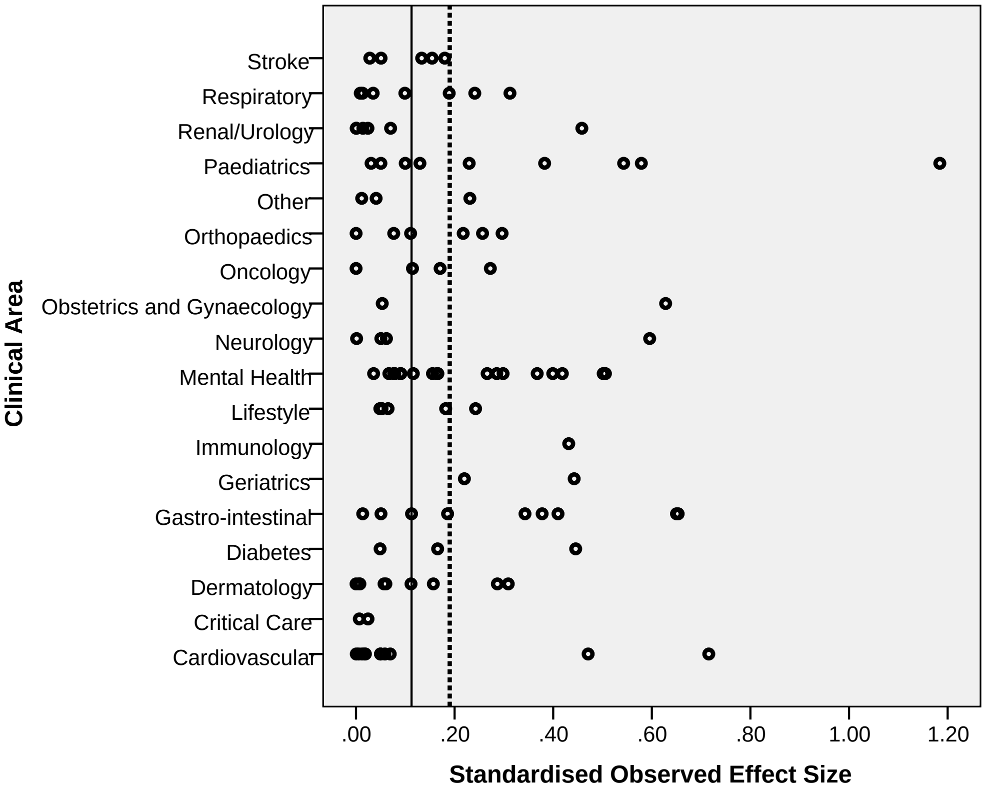


Figure 4.8: The standardised **observed** effect size across various clinical areas. The dashed line indicates the mean estimated effect size (0.191) and the solid line represents the median (0.113).

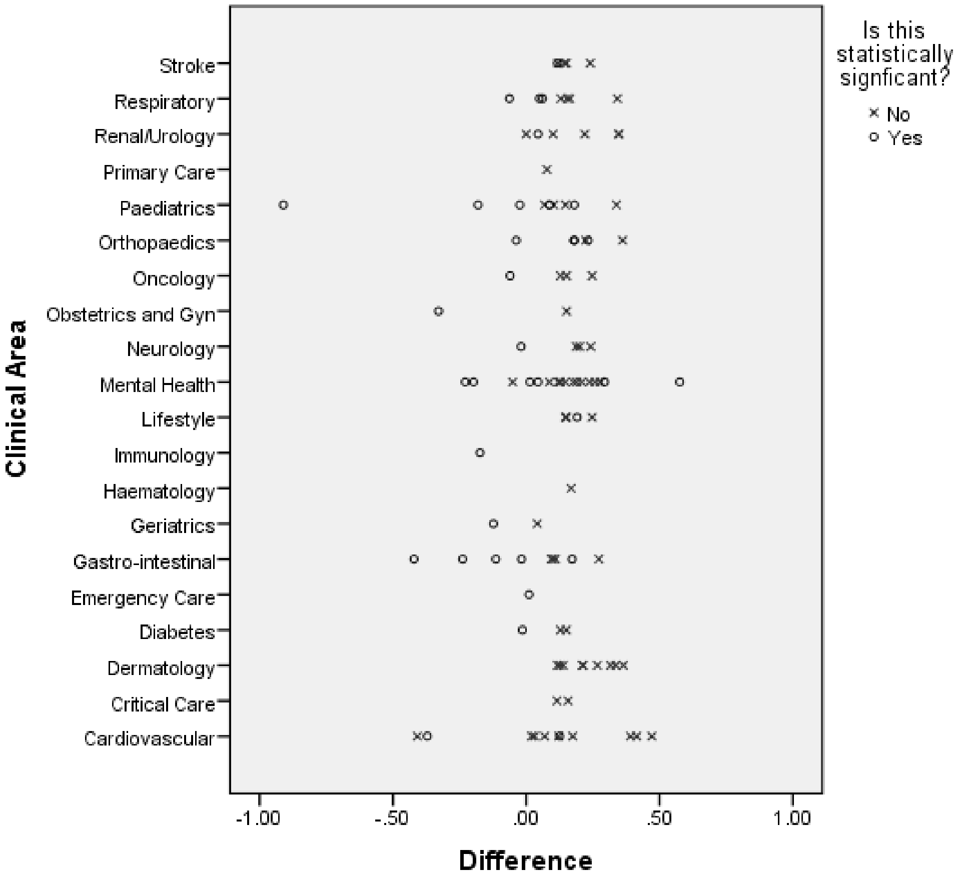


Figure 4.9: The difference in standardised target and observed effect size by clinical area and statistical significance.

Clinical Area	**Pre-2010**			**Post-2010**		
	Target Effect Size			**Observed Effect Size**		
	Count	**Mean**	**Median**	**Count**	**Mean**	**Median**
Cardiovascular	1	0.100	0.069	10	0.179	0.034
Critical Care	0	0.000	0.000	2	0.151	0.016
Dermatology	4	0.324	0.059	5	0.371	0.157
Diabetes	1	0.316	0.166	2	0.304	0.247
Gastro-intestinal	3	0.229	0.410	6	0.328	0.264
Geriatrics	1	0.261	0.220	1	0.319	0.443
Immunology	2	0.509	0.432	0	0.000	0.000
Lifestyle	0	0.000	0.000	5	0.300	0.065
Mental Health	5	0.463	0.166	13	0.309	0.164
Neurology	0	0.000	0.000	4	0.270	0.056
Obstetrics and Gynaecology	0	0.000	0.000	2	0.252	0.341
Oncology	1	0.362	0.115	3	0.212	0.171
Orthopaedics	0	0.000	0.000	6	0.331	0.164
Other	0	0.000	0.000	3	0.180	0.041
Paediatrics	3	0.274	0.579	6	0.376	0.180
Renal/Urology	0	0.000	0.000	6	0.296	0.019
Respiratory	1	0.162	0.009	6	0.240	0.144
Stroke	1	0.169	0.051	6	0.240	0.144
Overall	23	0.316	0.135	84	0.299	0.105

Table 4.8: Pre- and Post-2010 summary statistics for the observed and target effect size. * Median used due to skewed **observed** effect size distribution.

of the line are the trials which observed a smaller effect size than the target, yet were statistically significant.

Table 4.9 shows that continuous endpoints are most common, followed by proportions. The three endpoints in the 'Other' category are two trials which used area under the curve and one which had an ordinal endpoint.

Table 4.10 shows the movement between Cohen categories. The red cells indicate trials where a moderate or large effect size was estimated and a smaller effect size was observed. There are 32 trials which had a target effect size greater than the observed effect size, compared to 57 trials which were similar in terms of the target and observed effect sizes, with either a small or moderate effect size. A total of

Primary Endpoint Measure	**Count**	**Standardised Target Effect Size**		**Standardised Observed Effect Size**	
		Mean	**Median**	**Mean**	**Median**
Continuous	49	0.375	0.353	0.277	0.219
Proportion	41	0.224	0.198	0.115	0.048
Time to Event	10	0.291	0.312	0.147	0.065
Count	4	0.250	0.245	0.045	0.048
Other	3	0.295	0.295	0.169	0.186

Table 4.9: Standardised effect sizes by type of primary endpoint.

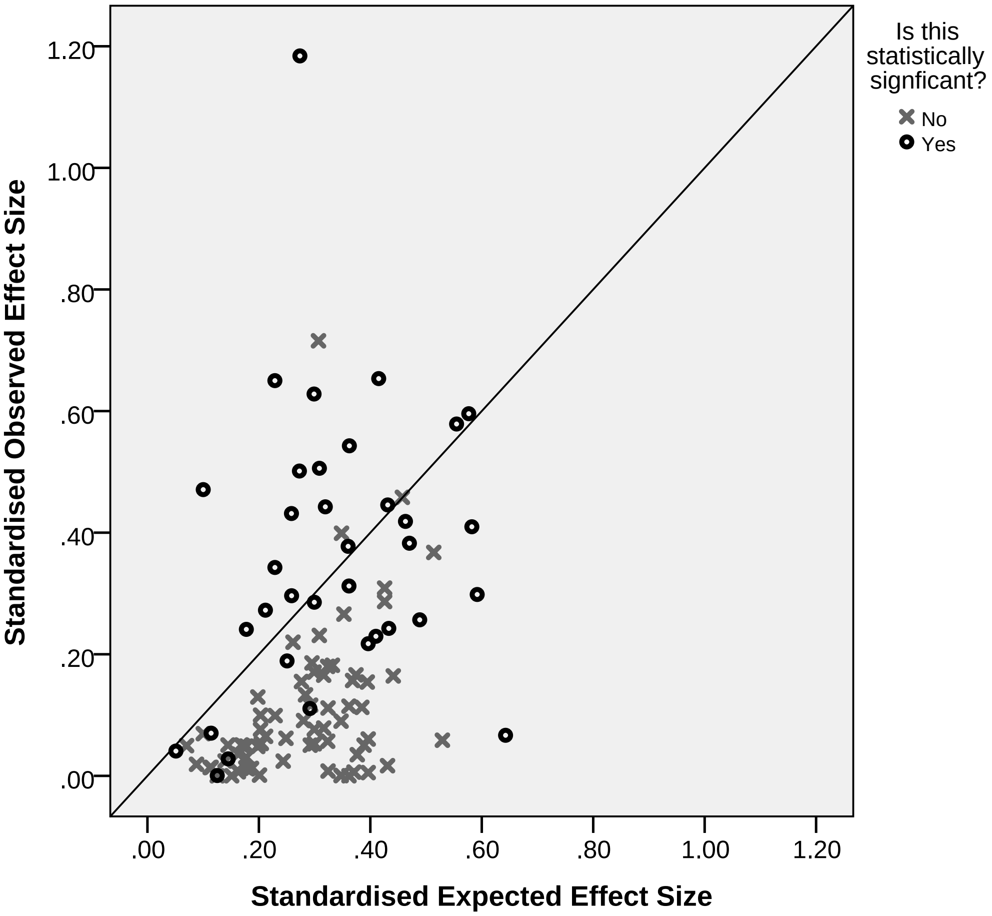


Figure 4.10: The standardised observed and estimated effect sizes, categorised by statistical significance.

		Standardised Target Effect Size				Total
		Small	Moderate	Large	Very Large	
Standardised Observed Effect Size	Small	37	31	1	0	69
	Moderate	7	20	0	0	27
	Large	2	7	0	0	9
	Very Large	0	1	0	0	1
Total		46	59	1	0	106

Table 4.10: The standardised estimated and observed effect sizes, categorised by Cohen's values.

17 trials had smaller target effect sizes than observed effect sizes. One trial was excluded due to not reporting an observed effect size.

A plot was created (Figure 4.11), similar to a Bland-Altman plot, which shows the average of the standardised observed and estimated effect sizes against the difference between them. This plot also shows whether the results were significant or not. This plot should show a random scatter if there is no relationship or association between the difference and the average of the estimated and observed effect sizes. As seen in the figure, there appears to be a relationship, as the average of the observed and target effect sizes gets larger, the difference between the two also gets larger. The points above the black line are over-estimated effect sizes and the points below the black line are under-estimations. As expected, the over-estimated effect sizes are mostly non-significant whereas the majority of under-estimations are significant. The limits of agreement are the standard 95% confidence interval around the mean difference. The plot is split by those trials which under-recruited by more than 10% and those which achieved a sample size within 10% of their target.

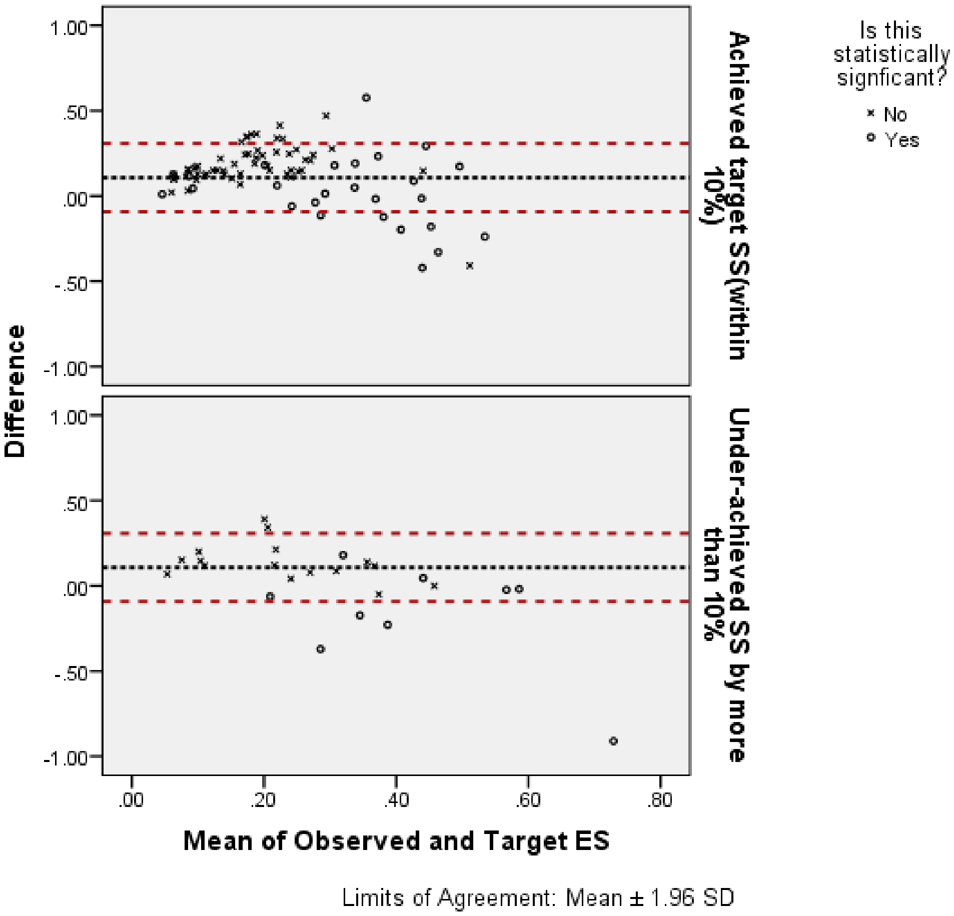


Figure 4.11: A Bland-Altman plot to show the differences between the target and observed standardised effect sizes.

4.3.5 Subgroup Analyses

Some considerations for subgroup analysis are detailed in this section. From the standardisation methods detailed in section 4.2.5.1, one is able to calculate the standardised observed 95% confidence interval and the detectable difference.

The standardised 95% confidence interval will allow further investigation to establish how many non-significant trials found the target difference in the observed 95% confidence interval. There were 7 trials which did not include all the required information to fully answer this research question.

Of the 100 included trials, 84% of them overall found the standardised target difference within the limits of the standardised observed 95% confidence interval. As discussed earlier in the chapter, the propotion of trials which were non-significant is around two thirds (71/106, 67%). The proportion of trials which are both non-significant and saw the standardised target within the bounds of the standardised observed 95% confidence interval is 56% ($N = 100$).

4.3.5.1 Detectable Difference

As discussed in section 4.2.5.2, the minimum detectable difference is calculated for the target power of each study. These values are then multiplied by the original standardised target effect size to investigate whether the observed effect sizes are detectable.

The results show that 33.96% of studies had a standardised observed effect size larger than the MDD. It should be noted that one study had data missing, so the total number of studies in this subgroup analysis is 106. This value is similar to the proportion of studies which reached statistical significance (32.7%), though 18.9% of the trials had opposing results, such that they were either statistically significant but did not reach the MDD or vice versa.

4.4 Case Studies

There were a number of reports which were noted as having well-explained examples of the target difference elicitation or examples of other methods used.

4.4.1 Example of Well-Justified Target Effect Size

4.4.1.1 TITRe2 Trial

One report which was published in 2016 by Reeves et al. provided a thorough justification of the target effect size (Reeves et al., 2016). It used a variety of methods to gather as much information as possible in order to estimate the target effect size, including using observational data to estimate the prevalence of eligible participants, as well as to estimate conservatively the transfusion rates for each group in the trial.

> “The trial was designed to answer superiority questions. The following steps were taken to calculate the sample size.
>
> - From observational data, we assumed that approximately 65% of patients would breach the threshold of 9g/dl and 20% would breach the 7.5g/dl threshold. Therefore, with complete adherence to the transfusion protocol, we assumed that transfusion rates should be 100% in the liberal group and $\approx$ 30% (0.20/0.65) in the restrictive group.
> - In the observational analysis, 63% of patients with a nadir haematocrit between 22.5% and 27%, and 93% of patients with a nadir haematocrit below 22.5%, were transfused. Therefore, in combination with the proportions of patients expected to breach the liberal and restrictive thresholds, these figures were used to estimate conservative transfusion rates of 74% for the liberal group and $\leq$ 35% for the restrictive group. These percentages reflected the rates of transfusion documented in the observational study (Figure 1) and assumed non-adherence with the transfusion protocol of approximately 26% in the liberal group and 5% in the restrictive group.
> - The observational frequencies of infectious and ischaemic events for transfused and non-transfused patients were adjusted to reflect the estimated transfusion rates in the two groups (i.e. 74% and $\leq$ 35%), giving event rates for the proposed composite outcome of 17% in the liberal threshold group and 11% in the restrictive threshold group. A sample size of 1468 was required to detect this risk difference of 6% with 90% power and 5%

> significance (two-sided test), using a sample size estimate for a chi-squred test comparing two independent proportions (applying a normal approximation correction for continuity) in Stata version 9.
>
> - The target sample size was inflated to 2000 participants (i.e. 1000 in each group) to allow for uncertainty about non-adherence and the estimated proportions of participants experiencing the primary outcome. We regarded these parameter estimates as uncertain because (1) they were estimated from observational data, (2) they were based on the red blood cell transfusion rate only in Bristol, (3) they were based on routinely collected data, using definitions for elements of the composite primary outcome which are not identical to those proposed for the trial, and (4) they were based on any compared with no red blood cell transfusion, rather than on the number of units of red blood cells likely to be transfused in participants who breach the liberal threshold. No adjustment was made for withdrawals or loss to follow-up, as both rates were expected to be very low.
>
> We expected approximately two-thirds of participants to breach the haemoglobin threshold for eligibility. Therefore, we predicted that we needed to register approximately 3000 participants into the study as a whole to allow 2000 participants to be randomised into the main study.
>
> The main outcome measure for the economic evaluation was quality-adjusted life-years (QALYs), which are derived from EQ-5D-3L utilities measured on a continuous scale and time under observation. The analysis of QALYs required baseline utility to be modelled as a covariate; the correlation between baseline and 3-month EQ-5D-3L utilities was assumed to be ≥ 0.3 With a total sample size of 2000, the trial had more than 95% power to detect a standardised difference in continuous outcomes between groups of 0.2 with 1% significance (two-sided test). This magnitude of difference is conventionally considered to be 'small'."

This discussion of the sample size calculation is extremely detailed, including the process the trialists implemented to use previous research, along with possible cautions and limitations of the calculation. The sample size was inflated to allow for uncertainties, which included that the estimates were based on observational data and there were differences in the composite end points for the observational data and the proposed trial (Reeves et al., 2016).

Following personal correspondence with the primary author, he stated that the process was truly “prospective”, such that they spent a lot of time working up the trial and trying to think how it would work - in particular, agonising over whether we could recruit and randomise before the operation and only include in the analysis population the patients who breached 9.0 g/dl.

> “What you have described is comprehensive, I think. I can also confirm that the process was truly “prospective” - we spent a lot of time working up the trial and trying to think how it would work - in particular, agonising over whether we could recruit and randomise before the operation and only include in the analysis population the patients who breached 9.0 g/dl. I am just so glad we stuck to randomising at the time of breaching the liberal threshold (even though there were considerable logistical challenges) - since there were instances of deviations to do with randomisation that, otherwise, we would have spent hours discussing!
>
> I find that ”target difference” is an alien concept to many clinicians - and I often fall back on my original education as a experimental psychologist, employing a ”bracketing” method to home in on the ”threshold” difference that a clinician believes to matter. In an application (currently under review), we wrote in defence of a non-inferiority margin: ”Instead, we estimated the non-inferiority margin using a bracketing technique. The trials unit investigators asked the surgeon investigators to consider how they would act on differences of different magnitudes; they were first asked to consider a large difference (expected to be unanimously judged as importantly different, namely (absolute difference of 3%, e.g. 7% vs. 10%, RR= 1.43). Surgeons were then asked to consider a small difference, expected to be judged non-inferior, namely 0.5% (e.g. 7% vs. 7.5%, RR= 1.07). Intermediate differences were postulated, alternating large then small, converging on a threshold difference where the surgeons were equivocal (1.5%).”
>
> One aspect of this sample size calculation was my decision to inflate the sample size - by quite an arbitrary amount, really - to take some account of the uncertainty. This probably reads as a better ”story” than it really was. It reflects an unquantifiable degree of uneasiness I felt (mainly because I knew the observational dataset inside out) and I would be hard pressed to defend or ”operationalise” the decision in other contexts. I guess I like target sample sizes that are round numbers (because I don’t believe that sample size calculations warrant the precision that are usually described as having). It is also interesting to consider whether a NIHR funding board would accept this rather broadbrush reasoning now - some older trialists might but I am less sure about younger

> statisticians who are often responsible for the very precise calculations that are often reported."

4.4.2 Examples of target effect sizes larger than those previously observed

There were three useful examples of an apparent over-estimation of the target effect size. These are the PD REHAB trial by Clarke et al., NIC-PIP trial by Tickle et al. and the BoTULS trial by Shaw et al.

4.4.2.1 PD REHAB trial

For the Clarke study, the trial investigated the effect of physiotherapy and occupational therapy on patients with Parkinsons' disease (PD) (Clarke et al., 2016).

> "In stroke patients, the MCIC in NEADL is 1-2 points. However, such a small change may be of only little benefit to PD patients; a clinically meaningful change in NEADL for PD patients is likely to be around double this, at 2.5 points. **A 2-point change on the NEADL represents becoming independent in one item (e.g. stair climbing, crossing roads or feeding onself) or improvement in two items (e.g. from being dependent on another person with help to being fully independent).** To detect a 2.5-point difference in NEADL at 3 months (using the observed SD from the PD OT trial of 10.1 points; $p < 0.05$ two-tailed; 90% power) required 340 patients in each group, increased to 750 participants (375 per group) to allow for around 10% non-complaince and dropout."

They aimed for a difference of 2.5 on the NEADL score which was the primary outcome, but observed changes of 1 and 1.5 in each of the groups in the trial. The target was elicitated based on what was deemed to be clinically meaningful to the population of interest (PD patients), since the original basis was that the minimum clinically important change (MCIC) of 1-2 points on the NEADL scale was based on stroke patients. There was no further discussion in the report regarding the choosing of 2.5 and whether or not this was overly-ambitious. However, it was discussed that the majority of the trial population had "mild disease with near normal NEADL scores. This may have led to a floor effect, as the NEADL score could not improve much from such a good baseline score." (Clarke et al., 2016).

Correspondance with the trial statistician resulted in some discussion that the highlighted section in the above extract implied that the method of elicitation was based on opinion-seeking from experts (both clinicians and patients).

It should be noted that whilst the target effect size was not met, this does not mean that it was an over-estimate of the effect size. It could be that the intervention was not successful in meeting the effect size, or, more likely in this case, that the population selected were too healthy so large improvements were not possible.

4.4.2.2 NIC-PIP trial

The Tickle study aimed to measure effects of a dental regime for young children(Tickle et al., 2016). This study based the target effect size on a previous similar study which observed a difference in proportions of 0.08; they expected to see a larger difference in their trial so used 0.10 in the sample size calculation.

> “The principal outcome measure is conversion from a caries-free state to caries-active state in the primary dentition. The sample size is therefore based on measuring an absolute difference between the intervention and control groups in the proportion of children who are free of the disease at 3 years. In the sample size calculation, we expected to see an absolute difference in the proportion of children with caries after 3 years of 0.1 between intervention and control groups. This expectation was based on the findings of a public health trial of toothpaste containing 1450 p.p.m. of flouride, on preschool children in the north-west of England, which reported 0.08 absolute difference in the proportion of children with caries active between the intervention and control groups. In this proposal, as fluoride-containing toothpaste was supplemented with biannual applications of fluoride-containing varnish, we expected to see a larger effect size and, therefore, a 0.1 absolute difference in proportions.
>
> The best data on the event rate for the practice-based population in Northern Ireland came from the Business Services Organisation database rather than epidemiological studies on other populations. Business Services Organisation data collected in 2008, at the time of planning the study, showed that 75% of 2− and 3−year olds in Northern Ireland who were registered with a dentist were caries free at first attendance. over a 3−year period, this reduced to 40% of 5− to 7−year-old children being caries free. Therefore, a further 35% of children were expected to develop caries active over a 3−year period. Based on these data and selecting caries-free children for inclusion in the trial, it was estimated that 47% would develop caries active over the 3 years. A two-group chi-squared test with a 0.05 two-sided significance level would have 90% power to detect the difference

> between a proportion of 0.47 and a proportion of 0.37 [odds ratio (OR) of 0.662], if the sample size in each group is 510."

The observed proportion was actually 0.05 which did not reach statistical significance. This is an example of an over-estimation of the effect size. In the discussion of this report it was mentioned that the estimate was also based on what would change clinical practice, as well as the previous data. They performed post-hoc analyses to confirm that the study was not underpowered to detect the original estimate of 0.10, but they clearly state that "one could argue that we set an ambitiously high effect size" (Tickle et al., 2016).

After discussion with the primary author, some further information regarding the journey to estimate the target effect size was provided. The target effect size was based on 3 factors. These were as follows

1. "a large trial of fluoride toothpaste sent to children in the same target age group conducted previous to the trial reported an effect size of 0.8. NIC PIP was a trial of a composite intervention comprising of providing fluoride toothpaste and fluoride varnish to the children. As fluoride has a well documented dose response relationship to caries we expected that the two interventions would have an additive effect.

2. we took soundings from primary care dentists about what would be a clinically important difference which would convince them to change their practice or would convince policy makers that an investment in the intervention was worthwhile - and there was consensus that a 10% or more absolute reduction would influence clinicians behaviour and policy. In hindsight it would have also been helpful ask what threshold would convince policy makers or clinicians that the intervention was not justifiable in terms of continuing NHS funding.

3. sample size - a sample size larger than 1500 - would have increased costs dramatically and the trial may not have been funded by the HTA. "

They also stated that

> "Out of the 3 factors I think the issue of clinical and policy importance is the most important factor in determining effect size - but this should be informed by current evidence (1) and take into account logistical and financial aspects of running a trial (3). "

4.4.2.3 BoTULS trial

Another report which stated they were overly ambitious in the design of the trial was by Shaw et al. (Shaw et al., 2010). This trial aimed to asses the treatment of upper limb spasticity casued by a stoke with a drug.

> “A power calculation was performed at the start of the study using prognosis based methodology. A clinically important treatment effect was defined as a difference in good outcomes between intervention and control groups of 15% where a good outcome was defined as listed above for each ARAT group; it was expected to see 20% of the control group achieve good outcomes and 35% of the intervention group achieve good outcomes. Using Fleiss’s method for a binary outcome and inflating the sample size by 10% to allow for attrition, we needed to recruit a total sample of 332 participants to give 80% power to detect a 15% difference in good outcomes assuming a two-tailed test and a significance level of 5%.”

The study observed a difference of 5.6%, compared to the 15% which was originally set as the target. In the discussion they stated

> “The study achieved the prespecified sample size of 332 participants so we can be confident that we have not missed an important treatment effect upon out primary outcome measure. However, it could be argued that our prespecified level of successful treatment was too ambitious.” (Shaw et al., 2010).

This honesty and transparency to the possibility that the original estimate was overly ambitious is not frequently seen, though has become more common in recent years. It can be calculated that the 95% confidence interval for the difference in proportions is −3.41% to 14.75% which also doesn’t include the target difference of 15%, therefore the trialists’ were possibly overly ambitious with their target estimate.

Following contact with the lead author and sending the 95% confidence intervals calculated above, further details about the power calculation were highlighted.

> “Our power calculation was looking for a 15% difference in good outcomes and we observed a 5.6% difference.
>
> When we presented our results to patient groups and study participants, their view was that any improvement in function was potentially an important difference.

> Thank you for sending the 95% CI for difference in proportions, on reflection an absolute difference of 15% was perhaps ambitious.
>
> The scientific basis on which clinically important differences are defined is an interesting area, and perhaps in the future there will be consensus for how this is decided"

This correspondance emphasises the importance of this work alongside the DELTA^2 project.

4.4.3 Example of Using a Pilot Study to Aid Elicitation

4.4.3.1 CADET Trial

One trial which reported using a pilot study to aid the elitication of the target effect size was by Richards et al. (Richards et al., 2016). This study was a cluster trial, therefore it was excluded from the full review. However, initially cluster trials were being included since they are an extension of individual RCTs so data extraction was completed on this report. The trial was investigating the effectiveness of collaborative care for depression in primary care.

> "We powered the trial at 90% (alpha= 0.05) to detect an effect size of 0.4, which we regarded as a clinically meaningful difference between interventions. This figure was within the 95% confidence interval (CI) of the effect predicted from data collected during our pilot work (effect size 0.63, 95% CI 0.18 to 1.07). To detect this difference would have required 132 participants per group in a two-armed participant-randomised trial. For our cluster trial, with 12 participants per primary care cluster and an intracluster correlation (ICC) of 0.06 from our pilot trial, the design effect was 1.65 leading to a sample size of 440. To follow up 440 participants, we aimed to randomised 550 participants (anticipating 20% attrition)."

The trial observed an effect size of 0.26 but reached statistical significance ($p = 0.009$). The discussion section in the paper details that whilst the observed effect size was less than the one which the study was powered on, the 95% confidence interval around the observed effect size included the target effect size. It also discussed that the observed effect size was also within the confidence interval of the smallest meaningful difference in a recent meta-analysis.

Conversation with the trial statistician and corresponding author, the trial was based on a clinically meaningful effect size of 0.4 which was independently identified. This was demonstrated in the trial protocol (Richards et al., 2009) which references two trials, a review and clinical opinion to elicit the target.

> "We did not use merely one rationale. I think this sort of triangulation of trial, review and clinical opinion data makes these decisions more secure."

The pilot study was used to demonstrate that a UK version of collaborative care might be likely to acheive such an effect, in line with collaborative care interventions elsewhere, mainly in the USA.

It was also clarified that the discussion about the observed effect size being in line with other published work was mainly to comment about the level of uncertainty around the observed effect size and the meta-analysis of other trials.

4.4.4 Conservative estimate

A number of reports discussed choosing a conservative estimate for the target effect size. The initial estimate could be based on previous research or various expert opinions, then the trialists' have made a slight informal adjustment to this estimate in order to be more conservative.

4.4.4.1 3CPO Trial

The 3CPO trial in particular used a conservative estimate and explained it clearly. It was a trial published in Volume 13 of the HTA journal which reported a trial investigating two methods of ventilation for sleep apnoea (Gray et al., 2009). The sample size justification was split into two parts, each referring to a separate research question.

> "The trial addresses two distinct questions:
>
> 1. **Is non-invasive ventilation superior to standard oxygen therapy?**
>
> The primary end point was 7-day mortality. Seven previous studies of acute cardiogenic pulmonary oedema ($n = 11 - 50$ per treatment group) at the time of protocol development had assessed standard facial oxygen therapy in comparison to CPAP ventilation, with only two further available studies assessing NIPPV ventilation. The pooled data shows a mortality rate of 21% (38/181) in patients receiving standard facial oxygen and 9% (16/173) in those receiving CPAP ventilation.

> In this trial we aimed to be able to detect a 6% absolute difference in mortality, which is half the effect size previously reported. To have an 80% chance of detecting a 6% difference (9% versus 15%) using a two-sided signficance level of 0.05 we needed approximately 400 patients to be randomised to standard facial oxygen therapy and 800 patients randomised to either CPAP or NIPPV.
>
> 2. **Which form of non-invasive ventilation is the most efficacious?**
>
> ...With 400 patients in each of the CPAP and NIPPV arms the trial aimed to have 80% power using a two-sided significance level of 0.05 to detect an absolute difference of approximately 7% in the composite end point (18% versus 11%) and of approximately 6% in mortality (12% versus 6%). "

This was just one example of trialists using an adjustment to result in a smaller target effect size. The results of the trial indicated that there was no significant difference in the primary end-point of 7-day mortality nor the primary composite end-point. The discussion included reference to these results being contradictory to recent meta-analyses, which indicated that patients treated with non-invasive ventilation reported up to a 47% reduction in mortality.

Communication with the author of this paper was achieved but due to the amount of time that has elapsed since the trial, he was uncertain of the various methods used to choose the target difference.

4.4.4.2 Ulceration in Diabetes

Another trial which used a conservative estimate was a trial by Jeffcoate et al. (Jeffcoate et al., 2009). They observed data from previous trials which ranged from 24% to 89% healing rates, but chose to power based on a 20% difference after reviewing the evidence. Both the Intention to Treat (ITT) and Per Protocol (PP) analyses at 24 weeks showed no statistically significant difference between the groups. There was no discussion about the target effect size or the observed effect size in relation to each other.

> As healing was the primary objective, this was the basis for the calculation of sample size. Calculation of sample size was difficult because of the paucity of data on the healing rate of different types of ulcer, and although data are available for neuropathic ulcers on the plantar surface, they are inconsistent. Thus, Katz *et al.* reported $61-89\%$ healing of plantar neuropathic ulcers within 12 weeks, while an earlier meta-analysis of the control arm of published trials

> of similar (but not all identical) ulcers reported only 24% healing with accepted good clinical practice by 12 weeks, and 31% at 20 weeks. Moreover, neuropathic ulcers with good vascular supply form a minority of ulcers cared for in the UK and, despite the lack of much published information, it is accepted that they heal more quickly than other types. The experience at the City Hospital, Nottingham, was that of all 449 individuals referred in the 4 years between January 2000 and December 2003, only 55% of index ulcers healed without amputation within 6 months of referral. It is on these bases that we calculated that in order to demonstrate a 20% difference in healing between groups, with 80% power, and with alpha= 0.05, and allowing for 25% dropout, 300 recruits were required. This was based on equal distribution of the sample to the three arms of the study, with an anticipated healing rate of 30%. The size was powered to indicate a 20% increase in healing for those in the Inadine group (50% healed at 24 weeks), and a 25% increase for those receiving Aquacel (55% healed at 24 weeks).

4.4.5 BeST Trial

One report by Lamb *et al.* provided an excellent discussion about choosing the Minimum Clinically Important Difference (MCID) based on previous trials (Lamb et al., 2010). This trial was investigating a primary care-based treatment for low back pain using a cognitive behavioural programme. The primary outcomes were commonly used continuous clinical measures.

> “Deciding the minimal clinically important difference (MCID) between groups was problematic, particularly for the RMQ. Previous trials (including the UK BEAM, Oxfordshire Low Back Pain Trial and York Back Pain and Exercise Trial) adopted a clinically significant difference between groups of 2.5 RMQ points, based on the views of an expert group of clinicians and researchers. This equates to a large standardised effect size of 0.65, assuming an SD of 4.0. Differences of this magnitude had not been observed in several large trials (effect sizes were 0.35 for BEAM and 0.36 for the York Low Back Pain Trial). Careful back tracking through trials (reviewed by Bombardier et al.) suggested that the MCID had been derived from a few studies of short-term benefits (< 8 weeks) of therapies in LBP. This is the stage at which one would expect to see the largest differences between the groups because of the natural history of LBP. Powering a trial on the short-term clinical benefit was unlikely to be sufficient to monitor longer-term impacts of public-health significance. The majority of outcomes reported for CBA suggest moderate benefits at 1 year, with a between-group effect size of

> approximately 0.35 for the majority of outcomes reported in efficacy trials. This equates to a between-group difference of approximately 1.4 change points on the RMQ disability score (i.e. new treatment approaches are approximately half as good again as the comparative treatment at reducing disability). We therefore considered that an effect size of 0.35 would be a suitable target for the CBA to be worthwhile."

The sample size discussion focused on the difficulties with choosing a MCID for one of the measures, the RMQ. It also detailed the calculation of the sample size to adjust for cluster effects and economic analyses. The observed treatment effects were 1.1, 1.4 and 1.3 at 3, 6 and 12 months respectively, all of which reached statistical significance.

After contact with the lead author of this paper, it was discussed that the main issue faced during the recruitment of this trial, along with many other pragmatic trials, is the issue of non-compliance. However, since this is not the primary focus of this research, it is discussed as further work in chapter 8.

4.5 Discussion

This chapter has presented the results from a large review of the HTA reports. It has been demonstrated that the most commonly reported method of target difference elicitation is the review of evidence method (45.8% when reported as the primary method, 52.3% when including use of multiple methods including review of evidence). There was no clear difference between which endpoint types used this method more commonly, nor clinical areas.

The median standardised estimated effect size was 0.300, whilst the median standardised observed effect size was 0.112. These both corresponded to "small" effects when compared to the Cohen categories (Cohen, 1988). This was not unexpected, as 67.3% of the trials did not reach statistical significance. The largest estimated effect size was 0.76 and the largest standardised observed effect size was 1.18, though this was only one of two trials which observed values greater than 0.66.

As part of the results of this review, several case studies were extracted to observe examples of good practice in reporting the target effect size elicitation. It was noted that many of the examples had used previous research also (part of the review of evidence method) and some had used multiple methods of elicitation to estimate a reasonable target difference.

One example of good practice is the TITRe2 trial (Reeves et al., 2016) in section 4.4.1.1. This report demonstrated the many aspects which this complex trial team had to consider and showed transparency in their reporting of it.

Another comment about the case studies was the number of reports which used a conservative estimate to base the target difference on. The included trials are a snapshot of well-reported trials or trials which used the MCID as part of the elicitation. It is known that secondary trials are not performed unless the results of the previous trial are promising, otherwise there would not be a need for the secondary trial. Using previous research for estimating the target difference is therefore difficult since the previous research available is mainly the trials which were published or reached statistical significance. This would result in an inflated estimate of the difference if based on the observed difference. This issue would result in subsequent trials possibly observing smaller differences, which sounds similar to a phenomenon observed in before-after trials called regression to the mean. This is discussed further in chapter 5.

4.5.1 Limitations

There were a number of challenges with this research. There is a large quantity of information reported in the documents which resulted in a large number of variables and data to extract. There was only one reviewer (JCR) which is a limitation of this work as it could not be quality-assured along the way.

There were some intial problems with the categorisation of the elicitation methods. The use of the seven DELTA categories was the product of a face-to-face discussion at a DELTA^2 workshop which JCR attended; the initial set of categories (based on experience of JCR) were added to as the research progressed. These initial categories provided more detail and as understanding increased as to how trialists were choosing their effect sizes, this allowed the design of more representative simulations further down the line. The use of the seven DELTA categories allowed better comparison to the original DELTA document (Cook et al., 2014) and ensured the categories were not unmanageable.

There were a considerable number of reports which didn't mention any method of elicitation (19.6%). These reports were evenly distributed across the volumes, not just from those pre-2010 as would be expected based on the requirements of the CONSORT document which outlined the requirements for publication, including a sample size justification (CONSORT, 2010).

4.6 Conclusions

The overall conclusions of this chapter are that previous research is the most common method which is reported for eliciting the target effect size. Trialists' are using this method alone or in tandem with other methods to get the best, most appropriate target effect size possible for their study.

The most common type of endpoint is a continuous endpoint, though proportion follows closely behind. Based on the findings in chapter 2, these are similar in their sample size calculation and therefore their use of the target effect size.

The case studies demonstrate some examples of good practice with regards to reporting the elicitation of the target difference. Transparency of the methods used to elicit the target difference is important, for trialists', clinicians and funders. The use of multiple methods to establish the best possible estimate appears to be best practice, as recommended in the DELTA2 guidance for target difference elicitation (Cook et al., 2018).

Since previous research is the most commonly used method of elicitation, chapter 5 will investigate further methods used to adjust for the possible bias which is introduced when using previous trial results or research to design a new trial. This will lead into chapter 6 where the findings of chapter 5 are tested through simulations.

5. Regression to the Mean

5.1 Introduction

As discussed in chapter 1, d is the most sensitive part of the conventional sample size calculation. Chapters 2 and 3 described how there could be a bias introduced when using the result from one trial to design the next, since the second trial would not occur unless the first trial had a significant result (for Phase II trials) or promising results (for pilot studies). During general scoping of the literature, whilst trying to determine if this bias is recognised, the term "regression to the mean" arose (Fayers and Hays, 2014; Novack and Crockett, 2009). In order to get more information about this phenomenon, a systematic searching of the literature was performed. Regression to the mean (RTM) "refers to the likelihood that an outcome variable will show a significant change depending upon how much baseline values depart from the mean", as defined by McCall et al. (McCall et al., 2011). This definition will be demonstrated in chapter 6. If only the 'promising' trials are taken forward to Phase III from Phase II, the average of the Phase III results will be less than the average of the 'promising' Phase II trials, due to an expected truncated distribution. This is caused by trials in Phase II having to exceed a pre-specified criteria to move to Phase III. The left-truncated Normal distribution results in a higher mean difference for that group, which subsequently dropped back to the average mean difference in study 2 which would not be truncated.

There are two points where regression to the mean could occur; those are when multiple measurements are being taken on the same patients (this can be defined as within-study) and when there are similar trials being conducted (this can be defined as between-study). This literature review will also try to investigate whether both these types are being considered when making adjustments to trials to reduce the effect of RTM.

5.1.1 Chapter Aims

The primary purpose of this systematic review is to improve understanding of regression to the mean and investigate what methods are used to adjust for it, both generally and in the context of trial design. A secondary interest is for which disease area or intervention types are adjustments for the effect more frequently used. It is also of interest to establish whether there are any articles stating that the same issue could occur for trials in sequence, compared to using the literature.

This review will detail what regression to the mean is, along with when it occurs. Some details have been placed in Appendix C as they were part of the initial systematic review, however as the thesis progressed, the focus of the review narrowed to a method more targeted with investigating the current adjustments used in the context of clinical trials. This was primarily researched using a pearl-growing method for reviewing the literature.

5.2 Background of Clinical Trials

In drug trials, randomised controlled trials are used to establish the efficacy and safety of new treatments before they enter the medical market. The process for a new drug or treatment to reach the open market is extremely long and expensive, on average lasting over 10 years from the compound discovery to product approval (Research and of America, 2015) and costing in excess of $18 million a year (Holland, 2013). This process is made up of multiple sections called phases, ranging from preclinical phases to Phase 4. Each phase has a particular purpose and the drug or treatment is unable to pass to the subsequent phase without being deemed successful in the previous phase. The phases are as follows (Friedman, 1985; Meinert, 1986; Schwartz, 1980)

- Preclinical phases: the drug under investigation is tested on non-human subjects, both in-vitro and in-vivo. The aim is to gather information on efficacy, toxicity and pharmacokinetics.
- Phase 0: this phase is sometimes omitted. The aim is to investigate the pharmacokinetics and pharmacodynamics of the drug. This would be the first test on humans, so a very low dose is used and only on a small number of participants. This phase cannot impart information on safety and efficacy due to the dosage being so low.
- Phase I: this phase is the real start of the clinical phases, with the drug being tested on healthy human participants to investigate the range of doses which are safe to test in future phases. It also investigates the tolerability, as well as the pharmacokinetics and pharmacodynamics.
- Phase II: this is when the drug is tested on a larger number of people (up to 300 typically) to test the efficacy and safety after the dosing ranges have

been established in the previous phase. This is the turning point in the trials process, if a drug is not showing efficacy the trials are stopped. The same safety assessments from Phase I trials are continued throughout the remaining phases.

- Phase III: this phase is the largest phase before drug-approval (typically recruiting 300-3000 participants). It is tested on ill patients to further confirm efficacy, effectiveness, safety and consumer acceptability. If the drug is shown to be effective, approval is sought.
- Phase IV: this phase is a continual process of post-marketing surveillance. It is to detect the long term effects of the drug and any previously unknown side-effects.

The most important phases in terms of this research are Phase II and Phase III, since a promising result in Phase II will lead to a large Phase III confirmatory study. Since the progression through the process is predicated by the drug being successful at each stage, it becomes clear that there is some form of bias which can occur, as shown in chapter 6. In general, bias occurs when there are conditions on performing a particular experiment or test. If all phases were unbiased, they would all be performed irrespective of the results from the other phases. Since this is not the case, a bias must occur when moving from one phase to the next.

As seen in chapter 1, there are two common methods of eliciting the target effect size, δ, which depend on prior information. These are the pilot study method and the review of the evidence base method. Publicly funded trials commonly get the initial results from previous work, such as a pilot study or a Cochrane review (of Health Research, 2010b). This is closely mirrored by the moving from Phase II to Phase III in the drug trial pathway, where the phase II trial can be seen as a pilot study (Wang et al., 2006). Therefore if this research can lead to adjustments for moving from Phase II to Phase III trials then it could potentially be applied to publicly funded trials as well, if the pilot study or Cochrane review is set to be the Phase II study. The information gathered from these previous trials is then used to design the main trial, including the choosing of the target effect size. Since these two methods are the commonly used and are based on prior information, these will be the methods considered to help answer the research question. As discussed in the clinical area of heart failure, (Krum and Tonkin, 2003) a large number of Phase III trials fail even though the Phase II trials indicated a positive clinical effect. When performing a general scoping of the literature for anything which resembles this change towards the true mean difference, the phenomenon called "regression to the mean" was highlighted which appears to explain this bias (Krum and Tonkin, 2003; Morton and Torgerson, 2005; Novack and Crockett, 2009). In order to understand this phenomenon further, a systematic review was performed.

5.2.1 Introduction to Regression to the Mean

According to the paper by Morton and Torgerson, regression to the mean occurs when “an extreme group is selected from a population based on the measurement of a particular variable.” (Morton and Torgerson, 2005). The paper continues to comment that if another measurement of the same variable is taken from this same group, the mean of the second measurement will be “closer to the population mean than the first measurement.” This definition could be thought of as similar to what happens when moving from a Phase II to Phase III trial, or when using results from a Cochrane review or meta-analysis to design a large publicly funded trial. The first result must be “encouraging”, the definition of which will be discussed later, if the second trial is to commence. So in this situation, the extreme group contains all the trials which are “encouraging” though the results observed will likely decrease towards the population mean at the next trial. This highlights the importance of investigating current methods for adjusting for regression to the mean which could be applied to the case of moving from one trial to the next. Along with current methods, it would also be of interest in this review to research other definitions of regression to the mean and establish whether it has been linked with moving from one trial to another in sequence.

5.3 Searching Methods

5.3.1 Inclusion and Exclusion Criteria

It is important for any systematic review to clearly state the inclusion and exclusion criteria. An article would be included in the review if the title and/or abstract specifically mentioned “regression to the mean”. Variations of this phrase were acceptable at the discretion of the reviewer, such as “regression toward mediocrity” or “regression towards the mean”. This could be either in an explanatory context or in an analytical context, so the search could capture not only trials where adjustments have been made, but also papers which provide methodologies and explanations of the phenomenon itself. The term of truncated normal distribution was included as a synonym for regression to the mean, as that is the resulting distribution when regression to the mean occurs.

As discussed previously, the main focus was on methods used to adjust for regression to the mean in randomised controlled trials, but another aim for the review was to improve knowledge of the phenomenon itself and how it impacts trials. If any titles or abstracts were vague, or the reviewer was unsure as to their relevance, the full article was retained and read to establish its relevance. Upon further reading of the full articles, they were deemed to be acceptable if they did any of the following:

- Explained or illustrated what regression to the mean is;
- Adjusted for it in a randomised controlled trial;
- Described a method for adjusting for it.

Articles would be excluded if, upon further reading, they did not contain any reference to the phenomenon (either directly or indirectly), if they were not written in English, if regression to the mean was used as a justification for their results without any further investigation. Articles were also excluded if they were a conference abstract due to too little information being provided or a letter to the editor as these are not peer-reviewed. Articles which are not related to clinical trials were also excluded.

5.3.2 Systematic Searching

There were not a large number of search terms used in this review, and synonyms were not of interest as the main focus was to locate randomised controlled trials only, and investigate the effect of regression to the mean. The databases of PubMED, StarPlus and the Cochrane Library were searched using key terms as follows:

1. Regression to the mean
2. Truncated normal distribution
3. Randomised controlled trial$
4. (1 OR 2) AND 3
5. AND
   (a) Phase 2 to phase 3
   (b) Phase II to phase III
   (c) Trials in sequence
   (d) Clinical development
   (e) Development plan
   (f) Sequential meta-analys$
   (g) Early phase trial$
   (h) Pilot study

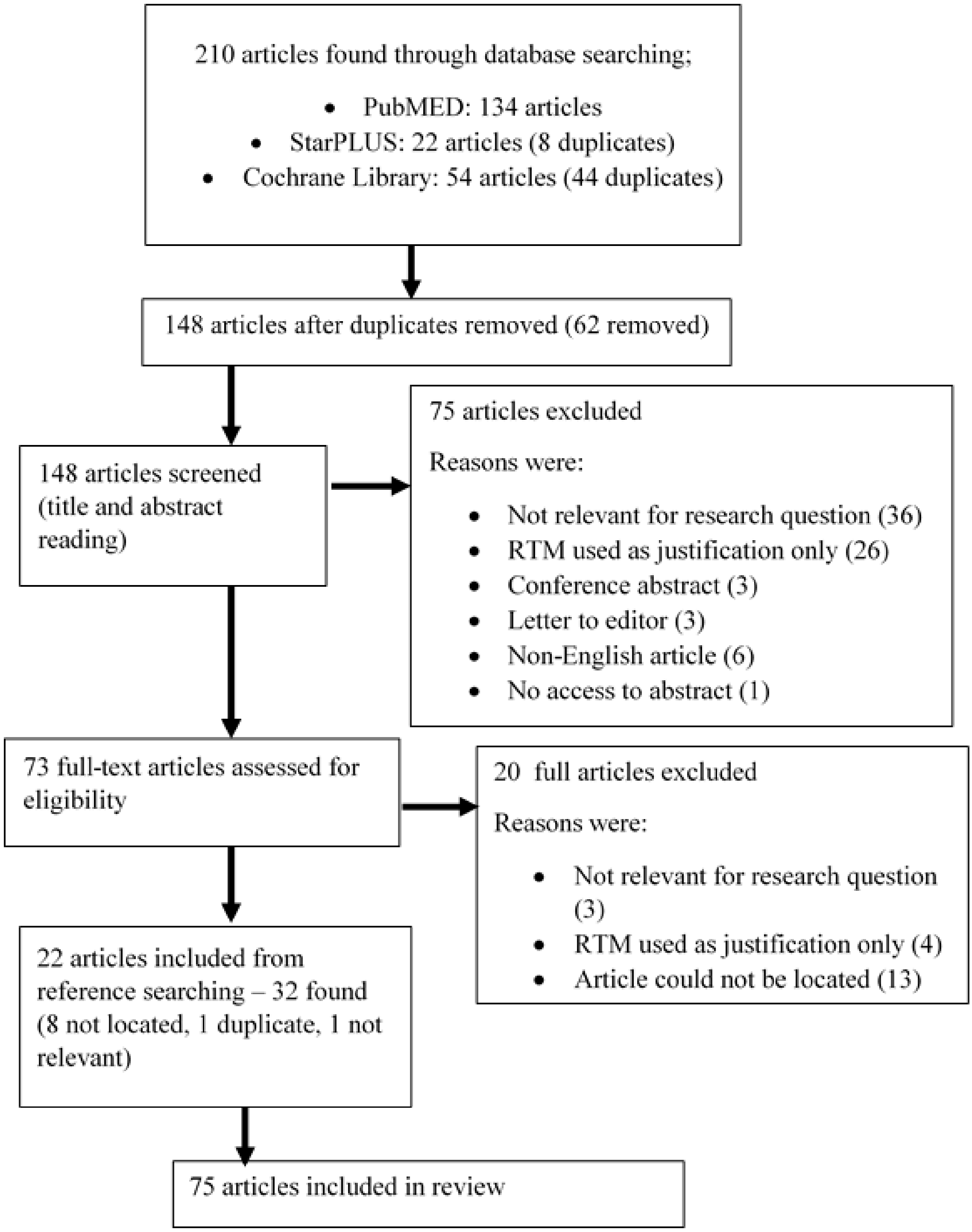

210 articles found through database searching;
PubMED: 134 articles
StarPLUS: 22 articles (8 duplicates)
Cochrane Library: 54 articles (44 duplicates)
148 articles after duplicates removed (62 removed)
148 articles screened (title and abstract reading)
75 articles excluded
Reasons were:
Not relevant for research question (36)
RTM used as justification only (26)
Conference abstract (3)
Letter to editor (3)
Non-English article (6)
No access to abstract (1)
73 full-text articles assessed for eligibility
20 full articles excluded
Reasons were:
Not relevant for research question (3)
RTM used as justification only (4)
Article could not be located (13)
22 articles included from reference searching – 32 found (8 not located, 1 duplicate, 1 not relevant)
75 articles included in review

Figure 5.1: Systematic Searching Results.

The initial searches used were 4 AND each of 5 (a-h), however these yielded very few results. The results of this search have been illustrated in Figure 5.1.

This produced 148 titles which were read, along with the abstracts, to determine their relevance to the research question. Many articles mentioned regression to the mean as a justification for their results (Brand et al., 2001; Burnham et al., 1994; Cooper et al., 1988; Elton et al., 1994) but did not attempt to account for it in the analysis. In total, 73 articles were taken forward for full reading, as well as 22 through reference searching. Unfortunately, 13 articles in the systematic searching and 7 articles in the reference searching could not be located so were omitted. This resulted in 75 articles being included in the review.

Regression to the mean was used as a justification of results in 30 articles without discussing possible adjustments which can be made in the analysis stage of the study. A few articles appear to be focused on the 'placebo effect' which one article has dissected to be a combination of regression to the mean, the Hawthorn effect and expectancy (McCall et al., 2011). Other explanations of the placebo effect were the natural history of the disease (or the cyclic nature of the disease), which can also contribute to regression to the mean (Burneo et al., 2002; Conboy et al., 2006; Enck and Klosterhalfen, 2005). Some of these placebo effect-related articles were not relevant to the research question of investigating regression to the mean (Burneo et al., 2002; Klosterhalfen and Enck, 2006).

5.3.3 Pearl-Growing Method

Another method which was used to enhance the review is the pearl-growing method. This method uses one useful resource and searches citations of and in that resource for other relevant resources. Citation pearl-growing is useful because some papers were not found through the systematic searching, instead they were discovered throughout the course of this research at conferences and through other relevant reading of articles.

5.4 Regression to the Mean

5.4.1 What is regression to the mean?

As mentioned previously, the problem of regression to the mean has been around for a long time, first being recognised by Francis Galton during the study of humans (Galton, 1886). He found that whilst studying the offspring of various heights, that the offspring did not tend to grow to the mean height of the two parents, but to actually be more similar to the population mean height. If the parents were tall, the offspring tended to be smaller, and the converse was also true. He subsequently called this phenomenon "regression towards mediocrity" which is now known as the more familiar term of regression to the mean. He also realised through his experiments that the further the parents were from the population mean, the more likely the child would be closer to that population mean (so would be smaller if the parents were extremely tall compared to the population mean).

More recently, the phenomenon is commonly seen with measurements like blood pressure, cholesterol levels and test scores in before-after designs. Consider the example of cholesterol, patients with high cholesterol are deemed high risk, and they are more likely to be recruited to a study in the reduction of cholesterol levels due to having higher levels. Regression to the mean would be observed if there was a lowering in the mean cholesterol level of the population of interest without any intervention occurring. The converse is also true, if a group of participants in a trial were selected based on low values of a biological marker, regression to the mean would result in the mean level increasing without intervention or treatment (Yudkin and Stratton, 1996).

The focus of this chapter is regression to the mean in randomised controlled trials, though this effect is commonly observed in observational trials over time (Cummings et al., 2004; Heather, 2014; Martinez-Yelamos et al., 2006; McCambridge et al., 2014; Victora et al., 1998).

One particularly interesting example of this was an observational study by McCambridge et al. which consisted of 976 college students in New Zealand being given an alcohol consumption questionnaire at baseline and again 6 months later. The purpose of the study was to show that regression to the mean was occurring, since in many alcohol intervention studies a reduction in alcohol consumed is observed in both intervention and control groups, implying that perhaps simply getting the participants to think about their drinking results in a reduction in consumption. It used an AUDIT questionnaire to assess how much alcohol was being consumed on average by the students, and tested different cut-off levels which have been used previously in alcohol intervention studies. The results of this study showed that the higher the cut-off level, the larger average reduction was observed in AUDIT score. This was useful as these cut-off levels have been used to recruit participants to studies in the past, so it shows that the studies with higher recruitment threshold levels would show more of a reduction simply through regression to the mean.

Other situations where regression to the mean can occur have been found in this systematic review. However, their relevance to the context of the research question, which is focused on moving from one trial to the next, is limited. Due to this reason, the details have been placed in Appendix C.

5.4.2 In what clinical areas is it common?

Whilst this question was posed at the start of the systematic review in the protocol, as the research has developed and the research question has become more refined, this question has become less relevant to the context of the research. Therefore, the details of this section have been moved to Appendix C. The main outcome of this question was that there is no set clinical area which appears to observe regression to the mean. There was no literature which indicated that regression to the mean occurs more frequently for particular conditions or clinical areas. It seems to be rather dependent on the design of the experiment.

5.4.3 What trial designs does regression to the mean commonly occur?

Regression to the mean is most commonly observed in randomised controlled trials. Since observational studies cannot have control groups in the regular sense as in randomised controlled trials, the observed effect in these studies is a combination of the treatment or intervention effect and other statistical factors like regression to the mean. However, Wolfe et al. states that the regression to the mean effect may be less in observational studies than randomised controlled trials (Wolfe et al., 2004).

One article states that regression to the mean is not just related to within-trial measurements, but can occur in meta-analyses and moving from phase II to phase III trials (Finney, 2008). This is similar to the definition formed at the beginning of this section for between-study regression to the mean. This article is extremely important as it is the only article to discuss that regression to the mean could occur in the context of moving from trial to trial, which is the primary focus of this PhD.

5.4.4 In what areas is regression to the mean adjusted for?

As discussed in Section 5.4.2, there was not a specific area or set of clinical areas which observed regression to the mean. This also applied to the adjustment of the phenomenon. There is no clinical area found in this systematic review which regularly adjusts for regression to the mean. It appears to be dependent on the design of the trial. Further details of this can be found in Appendix C.

5.5 Vanishing Treatment Effect

It has been reported that large treatment effects are often not replicated in future trials (Pereira et al., 2012). There are a number of plausible explanations for this, for example the trial may have been on a small sample size and when replicated on a larger sample the effect decreases in size (Ioannidis, 2008). Small or early trials which report observing very large treatment effects are then progressed to larger trials or later stage trials. These trials are then "failing" (Krum and Tonkin, 2003) when the main trial does not demonstrate an effect size close to that estimated in the smaller, early phase trial. Regression to the mean has been discussed as being a possible reason for this (ChuangStein and Kirby, 2014; Julious, 2010a; Krum and Tonkin, 2003), however, it could also be attributed to problems like publication bias and poor early phase trial design. The impact of publication bias would be that only the "positive" trials would get published, and are published more quickly, which leads to the possibility of a lot of seemingly positive evidence when actually there is a lag of the negative trials published work. With regards to the design of the phase 2 or pilot studies, if a phase 2 trial is well-designed, this could reduce the impact of regression to the mean, however it is unlikely to eradicate the issue completely. The occurrance of regression to the mean in the context of trial design and moving from one trial to the next needs careful management.

5.5.1 Bias in Sequential Trials

A paper by Kirby *et al.* discusses two sources of bias which can occur in phase III trials when using the results from phase II (Kirby et al., 2012). The first source is based around the populations used in phase II compared to those used in phase III. The populations used in phase II, a confirmatory phase, are usually more heterogeneous than those in phase III. The second source of bias comes from the selection of promising or optimistic results at phase II, using these results as the basis of treatment effect in phase III trial design (Kirby et al., 2012). The general advice from Kirby *et al.* is to reduce the observed effect size for a phase II trial by at least 10% to give a more accurate estimation of the treatment effect which will be observed in phase III.

5.6 Methods of Adjustment for Trials in Sequence

According to Zhang *et al.*, proposed methods of adjusting for overestimation of the treatment effect are not regularly implemented (Zhang et al., 2012). There have been adjustment methods developed as early as 1990 for group sequential designs (Emerson and Fleming, 1990; Pinheiro and Demets, 1997), yet as described in chapter 4, they are either not being reported or not being used in the context of previous research.

Some papers found in the review discuss methods of adjustment for data from early phase trials to use in later phase trials. These articles were further investigated and pearl-growing was used to collect the relevant articles.

The paper by Wang *et al.* (Wang et al., 2006) proposes a method of adaptation for the sample size calculation when using data from phase II trials. The context of this adaptation is industry-based (Phase II to Phase III, or early-phase to late-phase) and considers the use of surrogate endpoints in phase II trials to be one of the causes for a high failure rate of phase III trials. This could be, for example, the use of tumour shrinkage as the end-point for the phase II trial when the primary outcome is survival in the phase III trial. Another example could be using a 1 month outcome in the early phase trial when the main trial requires a 6-12 month outcome. This paper is focused on the calculation of the sample size for phase III trials based on either the point estimate from the phase II trial or the lower confidence limit. It recommends, based on simulation results, to have a bias adjustment of

$$\hat{\Delta} - 1 \times s.e(\hat{\Delta}) \tag{5.1}$$

where $s.e$ is the standard error and $\hat{\Delta}$ is the point estimate from the phase II trial. This result can lead to very small estimated effect sizes and therefore not many phase III trials being started.

Following from this result, Kirby *et al.* developed an adjustment method which was tested on the scenarios used by Wang. This method is a multiplicative adjustment (Kirby et al., 2012) which is based on the concept of assurance. The general adjustment is

$$\hat{\Delta} \times 0.9. \tag{5.2}$$

These methods of adjustment could be applied to the context of previous research to main trial which, as found in chapter 4, is the most common method of target difference elicitation. However, these methods currently are not being applied to this context. They are also unlikely to be generalisable to other scenarios such as the use of pilot data.

5.7 Discussion

This chapter has defined the phenomenon known as regression to the mean. It has investigated the areas which this can occur and how it arises. There are methods of adjustment developed for phase II to phase III trials, which adjust the observed treatment effect from the phase II trial to use in the design of the phase III trial. This is analogous for publicly funded trials moving from pilot study to main trial. However, chapter 4 highlighted that there are no formal methods of adjustment being implemented or reported for trials which use previous research to estimate the target difference. Some trials did report that they used a conservative estimate, though these appeared to be rough adjustments applied to the previously observed effect sizes.

Reviewing the literature has highlighted that whilst there are adjustment methods for phase II to phase III data (Kirby et al., 2012; Wang et al., 2006), these methods are generic to aid ease of use. It could be argued that more specific methods could be developed to be used in specific scenarios such as the use of previous research like systematic reviews, meta-analyses or previously published trials in a similar area, or pilot studies.

The adjustment proposed by Wang could have sensitivity issues, since it is based on the standard error which is influenced by the sample size. As the sample size increases for larger powers and effect sizes, the standard error will decrease, causing the adjustment to get less strict. However, for lower powers the adjustment will be too severe. The adjustment proposed by Kirby is a flat rule which does not account for the differences in trial designs or power.

Both Wang *et al.* and Kirby *et al.* use simulations to assess their adjustments. Chapter 6 will also use simulation methods to show the bias which results when moving from one trial to the next, as well as the scenario of moving from a pilot study to a main trial. These scenarios are commonly occurring, with the first scenario being similar to the use of previous research to elicit a target difference for a current trial.

The two methods of adjustment discussed in this chapter will be compared with the method developed in chapter 7.

5.7.1 Limitations

The limitations of this systematic review are based primarily on only having one reviewer. This opens up the potential for certain articles to be missed. There could have been more databases searched, however this would have taken up a considerable amount of time. Another limitation is the exclusion criteria of articles being in languages other than English, as there were 6 articles in the initial search which were in other languages, two of which were potentially of interest based on the abstracts.

5.8 Conclusions

There were a number of randomised controlled trials which adjusted for the statistical artefact of regression to the mean within a trial. However, many of them appeared to do this after the results had been published, as an additional consideration. Since this phenomenon occurs frequently, it is rather surprising that there are not more trials which adjust for it, particularly in outcome variables which are known for causing the problem more frequently such as cholesterol levels or blood pressure measurements. It is clear that there is not one main research area which regression to the mean occurs in, it appears to occur in a wide variety of fields not limited to medical or clinical trials; it occurs in education as well. There does not appear to be one method which is deemed superior to others, it is very much subjective to the authors preferences, whether they prefer methods like ANCOVA or multiple regression to make the adjustment or if they feel more comfortable performing simulations.

A considerable number of papers used the issue of regression to the mean to justify why the control groups in their trials exhibited a response, but made no attempt to adjust for it. This could be down to factors such as time or cost, however it should really have been considered at the design stage to establish whether it could potentially be an issue further down the line.

The focus of this research is the occurance of regression to the mean *between* trials in sequence. A few papers discuss adjustments for regression to the mean in the context of phase II to phase III trials, but they are more focused on industry-based trials. Publicly funded trials tend to use previous research then perform the main trial, or use a pilot-to-main trial design. As seen in chapter 4 there are no formal adjustments being implemented and reported, yet informal adjustments of previously observed treatment effects are being used. Since it is still a common problem which is being overlooked, this prompts the need for further research particularly in this context. The adjustments by Wang and Kirby both have their merits, but there appears to be scope for a new adjustment, which will be developed in the following chapters.

From this systematic review, the effect of regression to the mean appears to be one explanation for the bias which can occur when using the results of previous trials to design the next. Therefore, this research shall also focus on trying to establish an adjustment method which can be applied to the context of trials in sequence for commonly occurring scenarios found in chapter 4.

6. Simulations of Commonly Used Trial Designs

6.1 Introduction

As seen in chapter 2, the target difference d is the most sensitive part of a standard sample size calculation. Chapter 2 also stated that the focus for the research is based on superiority, parallel group trials, with exntensions to other trial designs discussed in chapter 8. Chapter 3 demonstrated that there are a number of methods which can be used to elicit the target difference for designing a trial. The commonly used methods were the review of the evidence base, seeking expert opinion and pilot studies (Cook et al., 2014). However, as seen in chapter 4 it is the review of evidence base or previous research method which is reported around 50% of the time in Health Technology Assessment journal reports. Chapter 5 demonstrated the existence of a phenomenon called “regression to the mean” which could occur when using results from previous research to design a new trial, as with the review of evidence method. It also presented some general adjustments for the bias, however these adjustments are constant. This chapter will investigate the effect of this bias on the results of simulated trials, both for using previous research such as a Phase II trial and using a pilot study to inform the future trial. It will also begin to develop an adjustment method specific to using previous data. This will lead on to chapter 7 where a possible method of adjustment is discussed and compared with current methods.

6.1.1 Chapter Aims

This chapter will develop simulation methods for some common sequential trial designs. The context for the simulations is of vital importance, it is key to establishing commonly occuring elicitation methods for the target difference and base the context of the simulations on these. The simulations need to be context specific, as this allows more focused results and will lead to various adjustment methods in chapter 7. The reason for performing simulations is to emulate a scenario under some set conditions to observe what occurs when moving from trial to trial. These simulations are performed under ideal conditions, starting with the simplest case and moving on to a more complex elicitation method. It aims to assess the distributions which arise from trials in sequence where the second design is based on the first result.

6.2 Initial Simulations

Based on the research in chapter 5, it has been indicated that a similar effect to regression to the mean could be occurring when moving from early-phase to late-phase trials. One way this could be investigated is through simulating a large number of trials and investigating the effect of this regression.

Consider the simplest trial design of Study 1 followed by Study 2 (Julious et al., 2007). This can be illustrated as shown in Figure 6.1. From this figure it can be seen that Study 2 would not occur without a "positive' result in Study 1. This example could relate to many different scenarios, including but not limited to Phase II to Phase III studies, pilot studies to main studies, or systematic reviews to a new study. All these scenarios are using information gained in a previous study to aid with the next one. In the context of trials in sequence, the continuation criteria would be that there was a significant result in Study 1 ($P < 0.05$) in order to proceed to Study 2. The mean of the first study will, on average, be higher than the "true" mean, towards which the second study mean will regress. This will be further demonstrated in this section. In the context of a pilot study to a main trial, the continuation criteria can vary and should not be based on the P-value of the pilot study. An example of the continuation criteria for a pilot study to main trial context could be that the mean treatment difference is above zero and a pre-specified confidence interval includes (or is greater than) the MCID (Lee et al., 2014a), or that the one-sided confidence interval includes the target point estimate which is to be used in the main trial (Cocks and Torgerson, 2013).

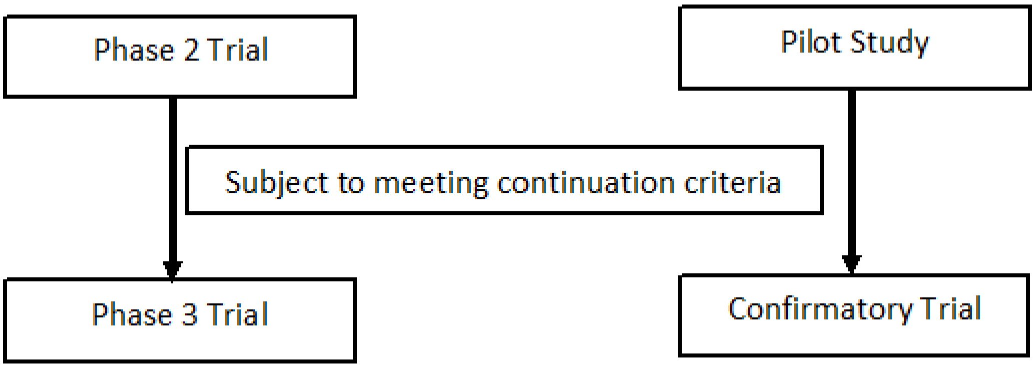


Figure 6.1: Illustration of the simplest scenario of moving from one trial to the next. The continuation criteria will differ depending on context.

The aim for the initial simulations is to investigate whether a bias is occurring when only the significant studies are taken forward to the second trial.

6.2.1 Methods for Initial Simulations

A study will usually not occur unless there have been “promising” or statistically significant results in earlier work. This automatically introduces a bias of the results from the first study; if there were no bias then the second study would occur irrespective of the result of the first. This bias is introduced through the selection of significant studies as predecessors of the new study, for example, Phase II to Phase III trials as described in chapter 5. Figure 6.2 shows that the significant Phase II studies will continue onto Phase III, they will have a P-value less than 0.05.

There are two scenarios which could occur in this section, the first being that all results are included irrespective of the result of the first trial, whilst the second being that only trials with a significant result in trial 1 (T1) are taken forward to trial 2 (T2). Both of these scenarios are considered in this section to investigate and compare the bias level and the distributions which arise from these two scenarios.

A trial will be deemed to be encouraging if the results are statistically significant, as shown in Figure 6.2. As discussed in earlier chapters, this is not always the case in reality, particularly when considering pilot studies which do not test a formal hypothesis therefore do not result in a P-value. The simulations which follow in this section are focusing on the situation where the results of the initial trial are used to design the following trial. This situation can occur not just with moving from Phase 2 to Phase 3 trials, but when results are used from systematic reviews or previously published clinical trials in a similar area. The simulations are performed in *R* (Version 3.1.2).

In order to perform the simulations, the design of the simulated trials needed to be considered. As discussed in chapter 2, the focus of this thesis is on parallel group superiority trials, further confirmed by those being the most common design found in the review in chapter 4. This design will be used in the simulations, for a continuous outcome as the results from chapter 4 indicate that these are more commonly used than other outcome types (45.8%) (Rothwell et al., 2018b). The more simulations performed, the less error that occurs in the results and the more accurate the results are in reflecting the “true” outcome.

The null and alternate hypotheses are as follows (with $d = \mu_1 - \mu_2$)

- $H_0 : d = 0$
- $H_A : d \neq 0$

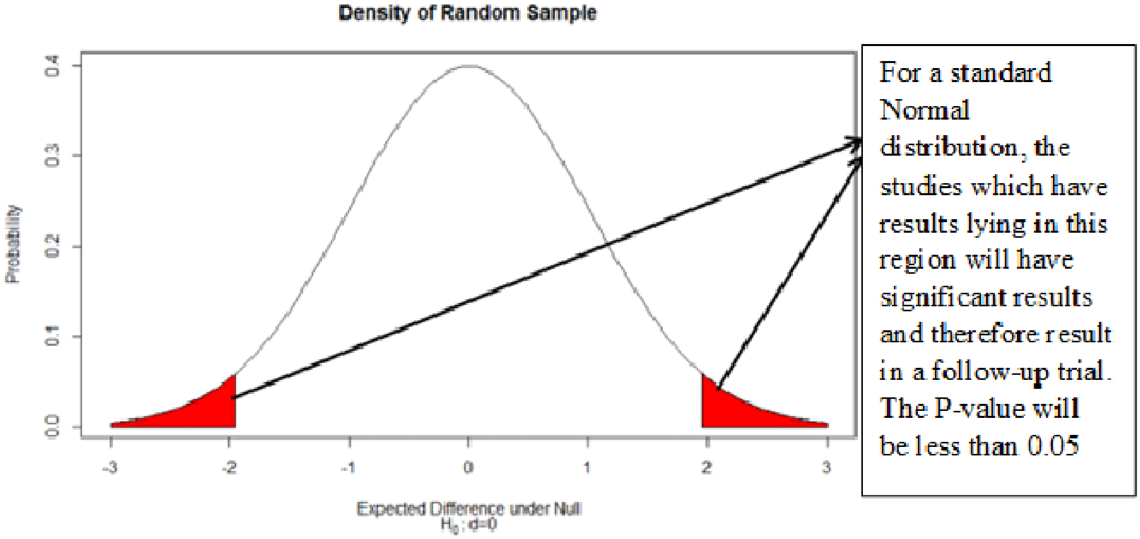


Figure 6.2: Illustration of a standard Normal distribution

The methods for this initial set of simulations is shown in Figure 6.3. Each simulation set represents a trial consisting of $2 \times n$ participants, where n is the sample size per arm. The sample size is calculated using Equation 2.6 from chapter 2. This set of simulations were performed twice, once with all results from the first study being included in the second study (the unbiased case) and once with only the 'encouraging' or significant results from the first study resulting in a second study (the plausible case). The power levels were then varied followed by the mean difference being varied to change the observed effect size.

The assumption that the patients follow a standard Normal distribution is implemented. Therefore, for both study 1 and study 2 the patients' results will come from a Normal distribution with means m_1 and m_2 respectively, and a common standard deviation. These simulations will be performed under the assumption of the alternative hypothesis, so the "true" mean for group 2 is equal to m_2.

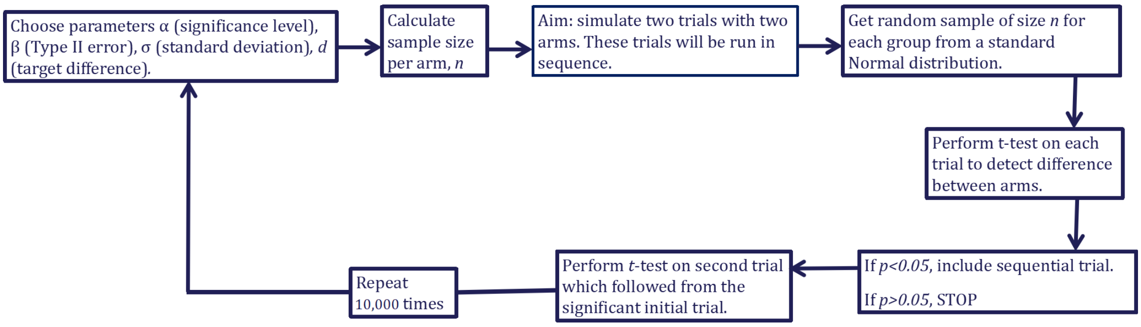

Choose parameters α (significance level), β (Type II error), σ (standard deviation), d (target difference).
Calculate sample size per arm, n
Aim: simulate two trials with two arms. These trials will be run in sequence.
Get random sample of size n for each group from a standard Normal distribution.
Perform t-test on each trial to detect difference between arms.
If p<0.05, include sequential trial.
If p>0.05, STOP
Perform t-test on second trial which followed from the significant initial trial.
Repeat 10,000 times

Figure 6.3: Flow chart of the methods used in the initial simulations.

Once the two studies have been simulated from the appropriate Normal distributions, the next step is to perform t-tests between the two groups of independent patients for each simulation set. This will give us a mean difference and P-value for each simulation set.

For example, there were 100 simulation sets, or simulations performed, then there would be 100 mean differences and P-values. From these results we are interested in the simulations where study 1 has a P-value of $P < 0.05$. These studies would have been statistically significant and would have lead on to a second trial; the simulations with $P \geq 0.05$ would not have resulted in a second study. The simulations which have a significant P-value for the first study can be extracted from the results and the original results for those significant simulations can be presented. These significant results are then plotted and compared with the results if these simulations were not removed.

This method results in a visual difference in the distributions of the simulations and trial data, dependent on whether the significant results were used or whether all results were used. It is of interest whether the mean difference for the Phase II and Phase III studies in these two different scenarios are similar or whether an effect mimicking regression to the mean is evident.

These initial simulations demonstrate the bias in the most simple case; using the sample target difference for both T1 and T2. Further simulations in section 6.4.1 will consider the case where the observed effect size in T1 is used to design T2. Section 6.4.2 will investigate the scenario more commonly used in publicly funded trials of a pilot study to a main trial.

6.2.1.1 Phase II to Phase III - Part 1

Two studies were simulated $10,000$ times, with one ‘preceding’ the other. Each simulation is a trial. First, using a standard continuous end-point sample size calculation was used to determine the number of patients needed for each study. This was calculated using a mean difference of 10, population standard deviation of 50, power of 80% and significance level of 5%, which resulted in the sample size being 393 patients per arm per simulation.

This sample size was used as the sample size for each arm in each simulation for each of the two studies, resulting in an extremely large data set of randomly sampled data from a Normal distribution using $m_1 = 0$ for arm 1 and $m_2 = 10$ for arm 2 for each of the two studies, and a standard deviation of 50 for both studies. This results in a random set of data for two arms and two studies. From this point forward, the assumption will be that the Phase II and Phase III trials are the same size.

These simulations were repeated for various powers and effect sizes, the results of which are shown in section 6.2.2.

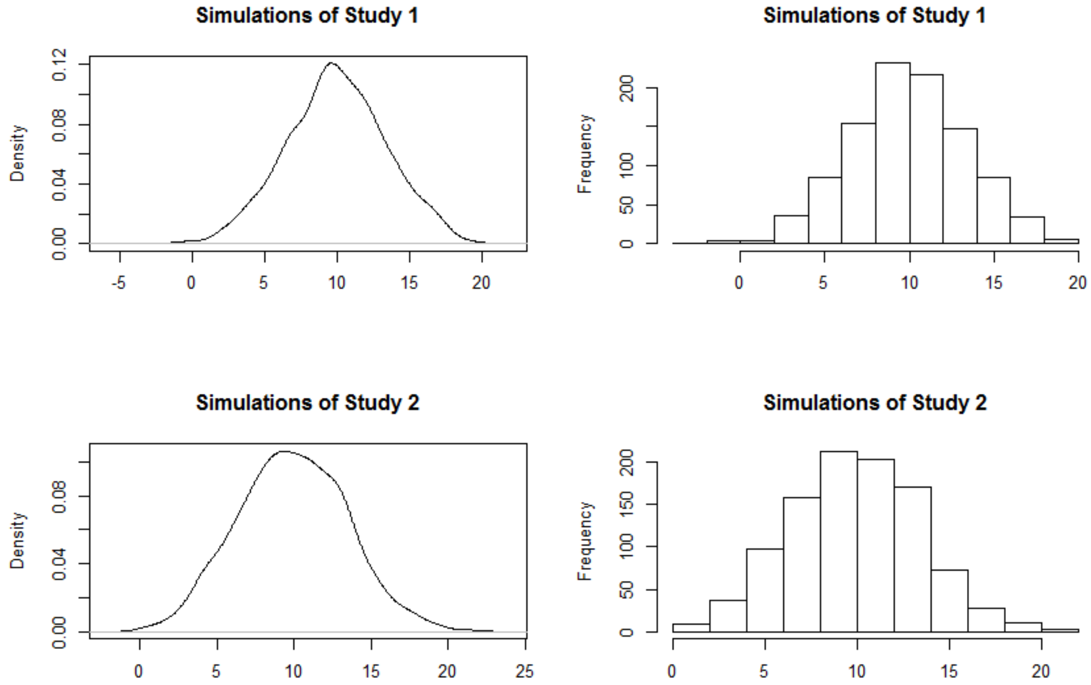


Figure 6.4: Distribution of the two studies with 80% power if all results included - The unbiased case.

		Unbiased Case		**Biased Case**		
Power	**Sample Size (N)**	Mean Difference (SE) Trial 1	Trial 2	Mean Difference (SE) Trial 1	Trial 2	Ratio of Mean Differences
80%	393	10.02 (3.57)	10.03 (3.56)	11.24 (2.75)	10.03 (3.56)	1.121
85%	450	10.04 (3.36)	9.94 (3.34)	10.97 (2.70)	9.96 (3.33)	1.101
90%	526	10.00 (3.09)	9.95 (3.08)	10.63 (2.58)	9.95 (3.09)	1.068
95%	651	10.02 (2.75)	10.01 (2.78)	10.31 (2.49)	10.00 (2.78)	1.031
99%	920	10.00 (2.33)	10.01 (2.34)	10.06 (2.26)	10.01 (2.34)	1.005

Table 6.1: The results from simulations with **constant effect size (0.2)** and variable power.

6.2.2 Results for Initial Simulations

Figure 6.4 highlights the assumption of no bias if all the results from the first study lead to a following study. The two distributions are similar, symmetrical and Normally distributed.

Figure 6.5 shows the distributions when the significant results from study 1 are taken forward to study 2. This can be considered because the Phase III trials would not usually occur if there were not signficant results at the Phase II stage. From Figure 6.5 it can be seen that when the significant results are taken forward, the resulting distribution for the first study appears to resemble a truncated Normal distribution.

From Table 6.1 it can be seen that as the power increases, the sample size increases. It is also clear that as the power increases the difference between the two trials for

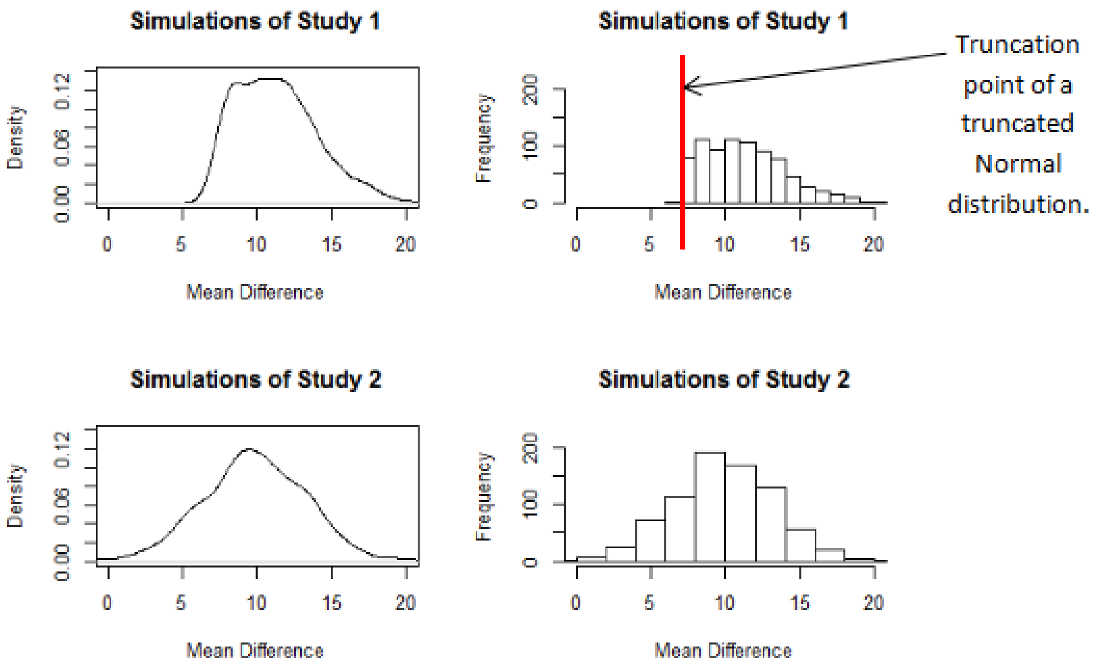


Figure 6.5: Distribution of the two studies with 80% power if only the significant trials continue to next phase - the biased case.

		Unbiased Case		**Biased Case**		
Effect Size	**Sample Size**	Mean Difference (SE)		Mean Difference (SE)		Ratio of Mean
	(N)	Trial 1	Trial 2	Trial 1	Trial 2	Differences
0.2	393	10.02 (3.57)	10.03 (3.56)	11.24 (2.75)	10.03 (3.56)	1.121
0.3	175	14.92 (5.37)	14.98 (5.43)	16.83 (4.10)	15.03 (5.45)	1.120
0.4	99	19.95 (7.19)	19.96 (7.03)	22.49 (5.50)	20.00 (7.03)	1.125
0.5	64	24.85 (8.75)	24.99 (8.84)	27.89 (6.68)	24.97 (8.85)	1.117
0.6	45	29.92 (10.65)	29.87 (10.63)	33.58 (8.24)	29.77 (10.58)	1.128
0.8	26	40.01 (13.79)	40.00 (13.79)	44.58 (10.72)	40.01 (14.08)	1.112

Table 6.2: The results from **constant power (80%)** and varying effect size.

the significant simulations results decreases, indicating that the truncation point is becoming smaller as the power increases. This prompts an interesting problem: since the minimum aim for power is 80%, if trialists are overestimating their effect sizes when calculating their sample sizes and recruiting for that target sample size, their study will not reach 80% power. If the power does decrease when the trial is completed and that particular trial is an early phase trial then the distribution will be extremely truncated and the potential regression to the mean effect will be very large when the following trial is performed, if the results from the current trial are used to design the future one. The relative mean difference is the ratio of the two mean differences, which decreases as the power increases.

Table 6.2 shows the impact of changing the effect size on the results of the simulations. Since the effect size was adjusted by altering the target means for each arm in the trials, the values of the mean differences themselves in Table 6.2 are not comparable; however the relative difference can be calculated, which is the ratio of the two mean differences. When the relative mean difference is calculated, we can see that it remains approximately constant as the effect size increases. This shows there is no change in bias. Since observed effect sizes can be quite small (Siontis and Ioannidis, 2011) the truncation would not be so extreme although it would still be there and regression to the mean would still occur. This would result in seeing lower mean differences between the groups of the second trial compared to the first trial.

6.2.3 Conclusions of Initial Simulations

Based on the simulation work in this section, it has been demonstrated that when moving from one trial to the next, there is a bias in the results which possibly leads to the phenomenon known as regression to the mean. This is where the mean of the second study is lower than that of the first. This bias only occurs when the encouraging results from study 1 are taken forward to study 2. It is shown by the distribution of the significant results for study 1 forming a truncated Normal distribution. This section has also shown that as the power of a trial increases, the truncation point of the distribution for Trial 1 becomes less extreme, resulting in the distributions for Trial 1 and Trial 2 becoming more analogous. As the effect size increases, the truncation point becomes more extreme, although this could also be affected by the significant decrease in the number of participants required as the effect size increases.

This effect, which resembles regression to the mean, could cause problems when designing a trial based on another trial. It also highlights further that the effect size is the most sensitive part of the sample size calculation. The work in this section is only the simplest type of trial design with two separate studies, not considering other potential factors such as multiple end-points or time points, or surrogate end-points which complicate the simulations further. These situations are common in a pilot setting, which is one of the situations being concentrated on in this research (Lee et al., 2014b). If an adjustment could be made when considering a biomarker for example, instead of the primary end-point in the first study, then perhaps money and resources could be preserved further during the early phase development. This work also compounds the need for the third and fourth research questions described in chapter 1

- Are there more optimal methods for quantifying the effect size?
- Are there more optimal methods to adjust for the bias of moving from one trial to the next?

If this bias is occurring when moving from one trial to the next trial based on encouraging results in the first trial, there should be an adjustment made to counter this issue and reduce the occurrence of regression to the mean. The term "encouraging results" is described by Lee et al. as a result where the 95% confidence interval includes the minimum clinically important difference (Lee et al., 2014b). This potential adjustment would not only reduce the propensity of regression to the mean, but provide a more accurate target difference for use when designing the second trial. Another consideration would be similar to using a Cochrane review to base the estimated effect size on, but using the results from a systematic review or meta-analysis, though this is outside the realms of this thesis.

The initial simulations shown thus far emulate moving from one trial to the next (Phase II to Phase III), as illustrated in Figure 6.1. This work forms the foundation from which the other simulations are built. These initial simulations were performed under the alternate hypothesis and were based on changing the power of a trial, and changing the effect size used in the sample size calculation. They were performed based on the theory that if a result is deemed non-significant at the Phase II level, the trial will not be progressed to the Phase III level. This results in a bias being introduced, if there were no bias then all studies from Phase II would progress to Phase III. This is why it has been documented that Phase III trials are failing to see target effect sizes close to that from previous early phase trials (Chuang-Stein and Kirby, 2014; Krum and Tonkin, 2003). It can be extended further to interim analyses, which will be discussed later in the chapter (Counsell et al., 2017). The effect of regression to the mean would also impact on surrogate end-points, this is discussed further in chapter 9.

When increasing the power of a trial, the sample size increased. The initial simulations showed that as the power increases, the truncation point of the distribution reduced. Taking the relative mean difference for the significant results only, it was seen that for an 80% powered trial the relative mean difference was 1.14 compared to 1.00 for a 99% powered trial (Table 6.1). This shows that as the power increases, the effect of regression to the mean became less pronounced.

As the target effect size increases, the sample size decreases dramatically. However, whilst it appeared as though the regression to the mean effect was getting larger, this was countered by the mean effects themselves increasing as well. When the relative mean difference between the significant trials was considered, it was around 1.12, indicating that there is an over-estimation of around 12% if the mean from trial 1 is used to design trial 2 (as seen in Table 6.2).

This highlights that changing the effect size does not have a drastic impact on the relative mean difference, the mean difference observed in trial 2 is consistently around 12% lower than the observed mean difference in trial 1. This allows us to begin to develop an adjustment for this over-estimation and test it through simulations. This adjustment can then be checked for the commonly used powers and adapted where necessary.

The target difference used in the sample size calculation for T2 is the same as that for T1. In section 6.4.1, the case is considered where the average observed effect size in T1 is used to design T2.

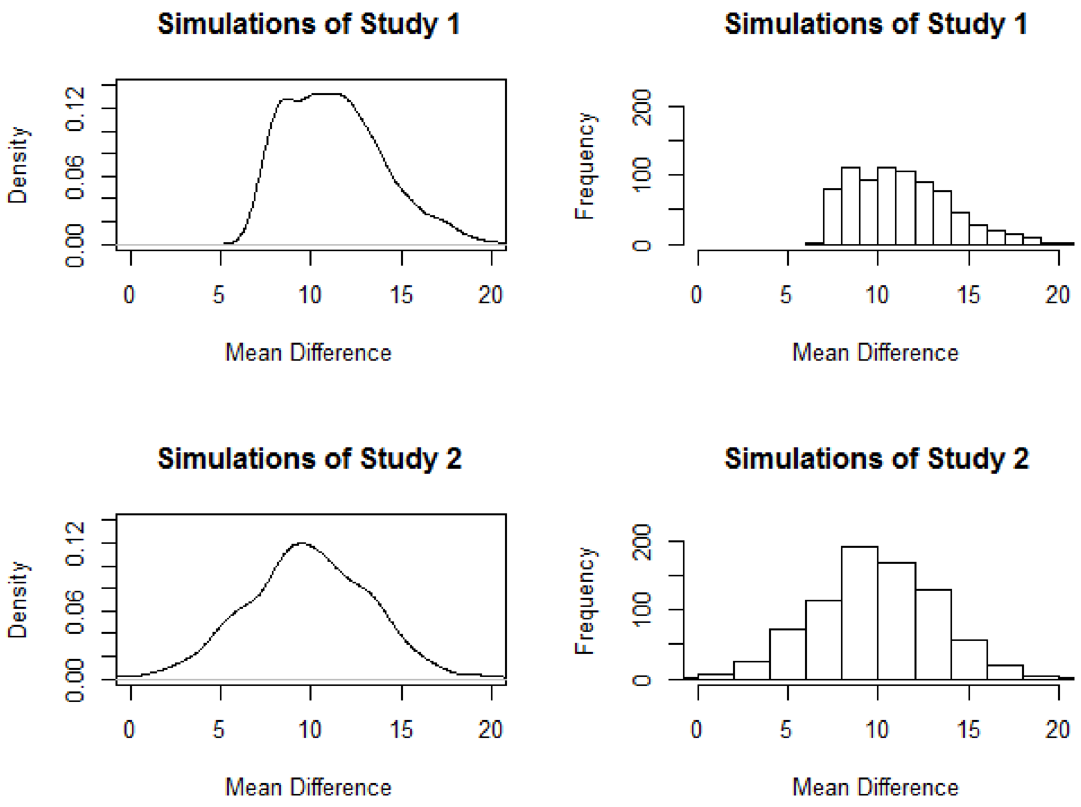


Figure 6.6: Distribution of the two studies with **80% power** if only the significant trials continue.

6.2.4 Truncated Normal Distribution

The distribution observed in Figure 6.6 appears to follow a truncated Normal distribution. This will be further defined and utilised in chapter 7, however, it can be noted that as the power increases to 99% in Figure D.2 from the 80% shown in Figure 6.6, the point of truncation moves further away from the mean of the distribution, resulting in a less extreme truncated Normal distribution. For the 99% power plot, the distribution is tending towards the full non-truncated Normal distribution.

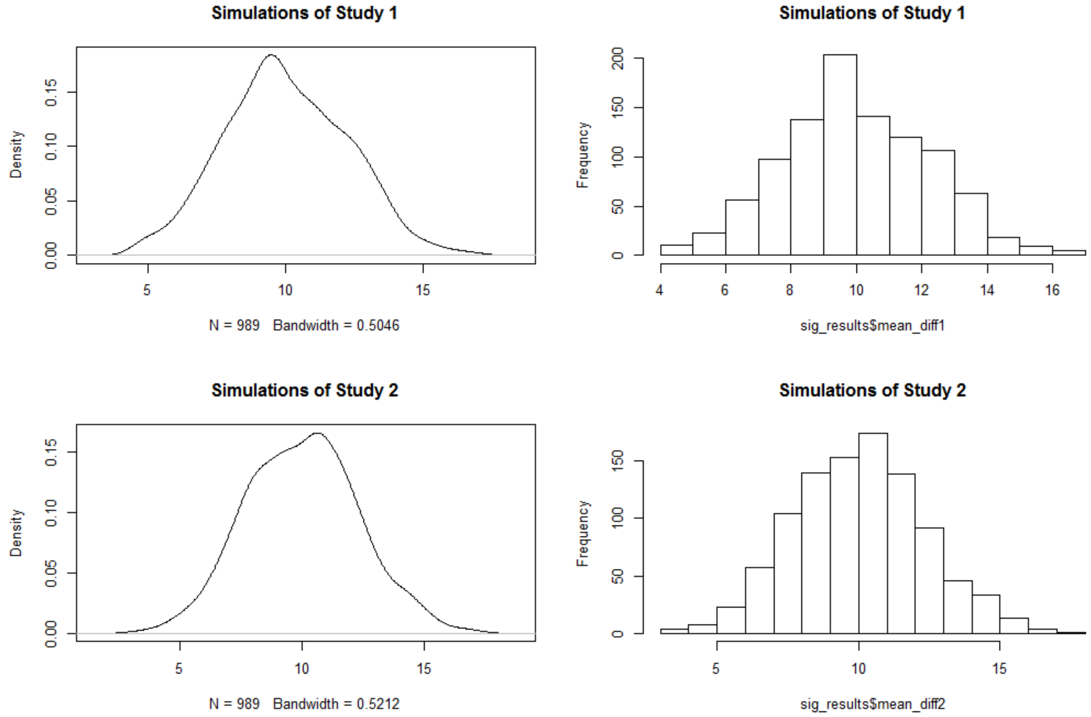


Figure 6.7: The distributions for trials with **99% power** if only the significant trials move on to study 2.

6.3 HTA Review Results

As seen in chapter 4, the most common method for elicitation of the target difference is using previous results and evidence. This could be from a variety of different study types, including pilot studies, meta-analyses and systematic reviews, and previous trials in similar populations. Since nearly 50% of the reports included in the review stated the use of previous studies or evidence, it is logical to base the simulations on this elicitation method to allow full exploration of the topic area rather than brief research on a large number of methods. The most common end-point type found in chapter 4 is a continuous end-point, therefore the main distribution used for the simulations shall be a Normal distribution.

As chapter 4 showed, the category of "previous research" could be broken down into the various categories of trials, including trial in sequence (for example, Phase II to Phase III) or pilot studies. These are the main categories which have been chosen for simulation due to their application to current practice and to the Clinical Trials Research Unit in Sheffield, who partially fund this research.

6.3.1 Simulation Plan

The work completed for the initial simulations demonstrated the scenario of a Phase II trial moving onto a Phase III trial. This was demonstrated by changing the power and the effect size, it simulated the simplest case which can be extended to other common scenarios in trial design. These cases are described as follows:

1. **Case 1a** - Phase II to Phase III trials, identical target effect sizes for T1 and T2. This was completed in the initial simulations.

2. **Case 1b** - Phase II to Phase III trials, observed effect sizes in T1 used in design of T2.

   - For the Phase II to Phase III scenarios, the ultimate aim is to solve these two cases mathematically and find the truncation point for commonly used powers (80% and 90%) and different target effect sizes. Once this truncation point is found, an adjustment can de developed to reduce the effect of the truncation, or the regression to the mean, resulting in a less biased effect size to use in other sample size calculations and trial designs. This will be developed in chapter 7.

3. **Case 2** - Pilot Studies.

   - Whilst the results from a pilot study don't necessarily impact the target sample size since the studies are used to assess feasibility and provide estimates of parameters such as the population standard deviation, the results still need to be deemed "encouraging". Therefore, instead of using a P-value in the simulations, the 95% confidence interval for the observed effect size can be used. This, along with a positive effect size, will form the basis for progression to main trial. The main trial would not occur if there were a negative observed effect. Another method could be to use a one-sided 80% or 90% confidence limit to ensure the sample size is large enough to include the postulated effect size for the main trial.

     The reason pilot studies need to be included in the simulations is that even though it is not recommended to use pilot studies to elicit the MCID or target difference, based on the results from chapter 4, it does occur. Pilot studies are also included in meta-analyses so need to be considered in their own right.

4. Alongside these scenarios, it would be useful to simulate the case where the observed difference is used in the sample size calculation for the second trial. This will result in a distribution of sample sizes as opposed to mean differences. The distribution of sample sizes will show that they are either over- or under-estimated and under-powered. This emphasises why the adjustment is important in the simulated context as well as re-iterating the issues discovered in the HTA review in chapter 4.

6.4 Methods

If two 2-armed studies were randomly simulated many times using the Normal distribution, one could investigate the means of each arm in each study from two perspectives: if the first study formed a Normal distribution as in the unbiased case, or if the first study formed a truncated Normal distribution, as is expected if only the significant results are taken forward (the biased case). The biased case is to be considered for the remainder of this chapter. The simulations are performed in *R* (Version 3.3.3).

6.4.1 Phase II to Phase III - Part 2

From section 6.2, the initial simulations demonstrated the bias when T1 and T2 were designed and executed using the same target effect size, $d = 10$. This highlights the bias in the most simple case.

The second part of the scenario for Phase II to Phase III is based on the first trial (T1) being designed on a set target difference, whilst the second trial (T2) is designed using the observed difference from each simulated T1.

The method used in this set of simulations is similar to that used in section 6.2, however the sample size for T2 is recalculated using the observed effect size for each T1 simulation. This results in $10,000$ separate sample sizes for T2. As with the initial simulations, the non-significant trials at T1 stage are removed and the means compared for T1 and T2. These simulations were completed for various powers and effect sizes.

For example, T1 is designed using a mean difference of 10, but had an observed difference of 11.5 when the significant trials are selected for progression. T2 is then designed using 11.5 as the target effect size. This scenario development mirrors the findings of the HTA review in chapter 4.

6.4.2 Pilot Studies

Another context which needs considering is the context of pilot study to main trial. This is commonly used in publicly funded trials (Thabane et al., 2010b), though was found to be used sparingly in chapter 4. However, this method is used in tandem with other methods, so should be considered in its own right.

The simulations to be performed under the context of pilot studies to main trial are similar to those performed previously. However, contrary to the method used up to this point, the pilot study does not have a formal power calculation to establish the appropriate number of participants needed. This is because pilot studies are not formally testing a hypothesis, they are demonstrating how the main trial will work, on a small scale (Whitehead et al., 2014). There are various theories on determining the number of participants for a pilot study, with some being rules of thumb methods (Browne, 1995; Julious, 2005; Teare et al., 2014). There are other methods which have more mathematical foundations, based on imprecision in the estimates of variance in the pilot study (Kieser and Hauschke, 2005), however, since the focus of this research is the moving from one trial to the next, as opposed to the intricacies of pilot study design, the rule of thumb methods will be used.

The rules of thumb methods have a set sample size for the pilot study, and based on that the parameters needed to design the main trial can be estimated. This results in a greater sample size overall irrespective of the size of the main trial (Whitehead et al., 2016). If the pilot study sample size is dependent on the main trial suggested sample size, as discussed by Whitehead *et al.*, then the sample size for the pilot study can vary slightly depending on the given parameters of the main trial, the significance (α) and the power ($1 - \beta$). The variables which are not fixed are the power (set between 75% and 95%) and the standardised effect size, which is to be set to match those used in previous scenarios for consistency (0.2, 0.3, 0.4, 0.5, 0.6, 0.8). Due to there being various methods to determine the sample size for a pilot study and a limited amount of time available for this research, it was decided to focus on the method proposed by Whitehead et al. (Whitehead et al., 2016). This decision was based on there being a lower limit to the size of a pilot study, yet the pilot study size altered depending on the target effect size of the main trial (and the planned sample size of the main trial).

There is no set criteria for a pilot study to be deemed "encouraging", though there is some advice in the literature. For all the simulations, the pilot study would not lead to a main trial if the effect was not positive, so all effect sizes must be positive to be included for continuation to a main trial.

The simulation plan will be similar to that used in the Phase II to Phase III simulations. The first set will be based around the same mean difference for both the pilot and the main. The second set will be based around the observed mean from the pilot study being used in the main trial. It is not advised, but to provide a baseline for comparison of the simulation results, the point estimate from the pilot study shall be taken only if the point estimate is positive.

Lee *et al.* define an encouraging pilot trial to be when the 85% confidence interval for the pilot study effect size contains the target effect size or MCID planned for the main trial (Lee et al., 2014b). This will be used as one of the possible progression conditions, along with some commonly used conditions. These conditions were gathered after personal discussions with trialists throughout the course of the

PhD. Therefore, for these simulations there will be 5 different progression criteria separately. These are as follows:

1. **Basic Case** - If $d_{pilot} > 0$,
2. **Lee condition** - if $d_{pilot} > 0$ and d_{target} is in the 85% confidence interval,
3. **Conventional condition** - if $d_{pilot} > 0$ and d_{target} is in the 95% confidence interval,
4. **Conservative Condition** - if $d_{pilot} > 0.5 \times d_{target}$
5. **Strict Condition** - if $d_{pilot} > 0.5 \times d_{target}$ and d_{target} is in the 95% confidence interval.

Methods of adjustment have been considered in chapter 7.

6.5 Results

This section presents the results of the simulations. It is separated in to two sections, the first being the results when the observed effect size is used to design the second trial for Phase II to Phase III or trials in sequence. The second section is for the scenario of pilot study to main trial.

6.5.1 Phase II to Phase III results

Consider Table 6.3, it can be seen that for trials with larger power, the difference observed in the average sample size for T2 increases. This could be due to the sample sizes producing a skewed distribution, which becomes slightly more skewed for higher powers, thereby raising the average for higher powered trials. These plots can be found in Appendix D The sample sizes for the second trial (T2) when varying the power are greater than the sample size for T1. However, the opposite is true when the effect size increases, which follows what was demonstrated in chapter 2, that as the effect size increases the sample size decreases.

The ratios of the mean differences appear to follow a similar pattern to those observed in Table 6.1 in section 6.2. The ratio of mean differences tends towards 1 as the power increases, indicating that the truncation point of the distribution is becoming less extreme and therefore has less of an impact on the results.

Table 6.4 also follows a similar pattern to Table 6.2 in section 6.2, with an approximately constant ratio of mean differences of $1.12 - 1.14$. In this case of varying effect size, the average sample size for T2 is less than that used in the first trial, T1.

			Biased Case		
Power	**Sample Size (N)**	**Average SS for T2 (N)**	**Mean Difference (SE)**		**Ratio of Mean Differences**
			Trial 1	**Trial 2**	
80%	393	364	11.30 (2.75)	10.01 (4.14)	1.129
85%	450	456	10.85 (2.68)	10.05 (3.70)	1.080
90%	526	556	10.64 (2.59)	10.05 (3.39)	1.059
95%	651	736	10.30 (2.48)	9.99 (2.94)	1.031
99%	920	1074	10.06 (2.24)	10.00 (2.41)	1.006

Table 6.3: The results from simulations with **constant effect size (0.2)** and variable power for simulations where T2 target effect size is based on T1 average.

				Biased Case		
Effect Size	**Sample Size (N)**	**Average SS for T2 (N)**	**'True' Mean Difference**	**Mean Difference (SE)**		**Ratio of Mean Differences**
				Trial 1	**Trial 2**	
0.2	393	364	10	11.30 (2.75)	10.01 (4.14)	1.129
0.3	175	164	15	16.85 (4.09)	15.01 (6.25)	1.123
0.4	99	92	20	22.49 (5.48)	19.86 (8.19)	1.132
0.5	64	60	25	28.01 (6.82)	24.65 (10.19)	1.136
0.6	45	42	30	33.40 (8.10)	29.82 (12.52)	1.120
0.8	26	24	40	44.48 (10.67)	39.86 (16.24)	1.116

Table 6.4: The results from **constant power (80%)** and varying effect size for simulations where T2 target effect size is based on T1 average.

This is to be expected as the target difference used for designing T2 is larger than that in T1, resulting in an increased sample size.

6.5.2 Pilot Studies results

The following results are based upon 10,000 simulations for the pilot study to main trial scenario. There are 5 progression conditions which are investigated separately, as mentioned in section 6.4.2. These will be briefly described in each section. The lowest number of participants in the pilot study (per arm) was 10 (Whitehead et al., 2016).

6.5.2.1 Basic Case

The basic case is that a pilot study will progress to a main trial if $d_{pilot} > 0$. This is the simplest scenario since if there was a negative observed effect in the pilot study then the likelihood of it progressing to a main trial is extremely small. There would be little justification to perform a trial for placebo versus new active treatment if the placebo showed evidence in the pilot phase for being superior.

Table 6.5 shows the results for 10,000 simulations for the basic case where the only condition is that the pilot study has shown a positive effect of any size. It can be seen that the level of bias decreases as the effect size increases. This was confirmed with a truncated Normal distribution where the truncation point is forced to be at

Basic	**80% Powered Main Trial**				
Standardised Effect Size	**Pilot Study SS**	**Main Trial SS**	**Pilot Trial Mean (SE)**	**Main Trial Mean (SE)**	**Ratio of Means**
0.2	20	412	16.96 (11.43)	10.03 (3.48)	1.691
0.3	14	188	22.06 (14.53)	15.04 (5.11)	1.467
0.4	11	108	26.81 (16.89)	19.96 (6.73)	1.343
0.5	10	70	30.33 (18.18)	24.97 (8.49)	1.215
0.6	10	49	34.05 (19.10)	30.00 (10.31)	1.135
0.8	10	28	41.61 (20.49)	39.91 (13.29)	1.043

90% Powered Main Trial					
Standardised Effect Size	**Pilot Study SS**	**Main Trial SS**	**Pilot Trial Mean (SE)**	**Main Trial Mean (SE)**	**Ratio of Means**
0.2	28	552	15.08 (9.95)	10.01 (3.01)	1.506
0.3	19	252	20.20 (12.84)	15.08 (4.46)	1.340
0.4	15	145	24.95 (14.84)	19.98 (5.90)	1.249
0.5	12	95	29.46 (17.19)	25.04 (7.27)	1.177
0.6	11	68	33.83 (18.83)	30.13 (8.62)	1.123
0.8	10	39	41.72 (20.66)	40.05 (11.38)	1.042

Table 6.5: The results of $10,000$ simulations when moving from pilot study to main trial, dependant on the observed effect in the pilot study (d_{pilot}) being greater than 0.

0 due to the condition of progression. As the effect size increased, the distribution moved further from 0, resulting in less bias caused by the introduction of $d_{pilot} > 0$.

6.5.2.2 Lee Condition

The condition which shall be referred to as the Lee condition, as described by Lee *et al.*, states that other confidence interval widths should be used when assessing the potential success of a trial based on a pilot (Lee et al., 2014b). As discussed in section 6.4.2, this shall be based on an 85% confidence interval.

The results of this condition of progression can be seen in Table 6.6. It can be observed that the Lee condition results in a lower ratio of means compared to the basic case as it is more strict than the basic case.

6.5.2.3 Conventional Condition

The results for these simulations are presented in Table 6.7. The conditions for progression to main trial are that the target difference of main trial, d_{target}, is within the 95% confidence interval of the pilot trial and that $d_{pilot} > 0$. Again, these results appear to be slightly more strict than the previous two conditions. All the ratio of means for each condition will be compared in section 6.5.3.

Lee	80% Powered Main Trial				
Standardised Effect Size	Pilot Study SS	Main Trial SS	Pilot Trial Mean (SE)	Main Trial Mean (SE)	Ratio of Means
0.2	20	412	16.10 (10.16)	10.03 (3.46)	1.605
0.3	14	188	20.87 (12.84)	14.98 (5.18)	1.393
0.4	11	108	25.11 (15.13)	19.93 (6.85)	1.260
0.5	10	70	29.37 (16.71)	24.95 (8.49)	1.177
0.6	10	49	32.93 (17.75)	29.99 (10.04)	1.098
0.8	10	28	40.63 (19.25)	39.95 (13.43)	1.017
90% Powered Main Trial					
Standardised Effect Size	Pilot Study SS	Main Trial SS	Pilot Trial Mean (SE)	Main Trial Mean (SE)	Ratio of Means
0.2	28	552	14.27 (8.86)	9.93 (3.01)	1.437
0.3	19	252	19.09 (11.51)	15.00 (4.45)	1.273
0.4	15	145	23.68 (13.57)	19.97 (5.88)	1.186
0.5	12	95	28.19 (15.73)	24.97 (7.28)	1.129
0.6	11	68	32.32 (16.88)	29.97 (8.53)	1.078
0.8	10	39	40.98 (19.15)	40.15 (11.34)	1.021

Table 6.6: The results of $10,000$ simulations when moving from pilot study to main trial, dependant on the d_{pilot} being greater than 0 and d_{target} being within the 85% confidence interval of the pilot study results.

Conventional	80% Powered Main Trial				
Standardised Effect Size	Pilot Study SS	Main Trial SS	Pilot Trial Mean (SE)	Main Trial Mean (SE)	Ratio of Means
0.2	20	412	15.85 (10.06)	10.00 (3.49)	1.585
0.3	14	188	20.82 (13.00)	15.02 (5.18)	1.386
0.4	11	108	25.26 (15.02)	20.00 (6.85)	1.263
0.5	10	70	29.16 (16.80)	25.03 (8.51)	1.165
0.6	10	49	33.07 (17.74)	29.87 (10.16)	1.107
0.8	10	28	40.95 (19.25)	40.10 (13.43)	1.021
90% Powered Main Trial					
Standardised Effect Size	Pilot Study SS	Main Trial SS	Pilot Trial Mean (SE)	Main Trial Mean (SE)	Ratio of Means
0.2	28	552	14.54 (8.94)	10.03 (3.03)	1.450
0.3	19	252	19.09 (11.39)	14.94 (4.41)	1.278
0.4	15	145	23.55 (13.67)	19.94 (5.92)	1.181
0.5	12	95	28.17 (15.54)	24.94 (7.26)	1.300
0.6	11	68	32.51 (17.12)	30.05 (8.49)	1.082
0.8	10	39	40.71 (18.95)	40.05 (11.36)	1.016

Table 6.7: The results of $10,000$ simulations when moving from pilot study to main trial, dependant on a positive observed effect in the pilot study and the 95% confidence interval containing the target effect.

Conservative	80% Powered Main Trial				
Standardised Effect Size	Pilot Study SS	Main Trial SS	Pilot Trial Mean (SE)	Main Trial Mean (SE)	Ratio of Means
0.2	20	412	19.55 (10.26)	9.93 (3.51)	1.969
0.3	14	188	25.55 (12.92)	15.03 (5.16)	1.700
0.4	11	108	31.04 (14.68)	19.96 (6.88)	1.555
0.5	10	70	35.42 (15.89)	25.02 (8.41)	1.416
0.6	10	49	39.46 (16.26)	29.99 (10.02)	1.316
0.8	10	28	46.89 (17.03)	40.13 (13.27)	1.168

90% Powered Main Trial					
Standardised Effect Size	Pilot Study SS	Main Trial SS	Pilot Trial Mean (SE)	Main Trial Mean (SE)	Ratio of Means
0.2	28	552	17.58 (8.99)	9.93 (3.01)	1.770
0.3	19	252	23.85 (11.43)	15.04 (4.43)	1.586
0.4	15	145	29.10 (12.84)	20.01 (5.86)	1.448
0.5	12	95	33.96 (14.66)	24.84 (7.31)	1.367
0.6	11	68	38.68 (15.74)	30.05 (8.49)	1.287
0.8	10	39	46.81 (17.08)	39.96 (11.50)	1.171

Table 6.8: The results of $10,000$ simulations when moving from pilot study to main trial, dependant on the observed effect in the pilot study being greater than $0.5d_{target}$. This is more conservative.

6.5.2.4 Conservative Condition

The conservative condition dictates that the observed effect size in the pilot study is greater than $0.5 \times d_{target}$; the results of which are shown in Table 6.8. These results are far larger than those presented in the previous tables, yet all the results have the same pattern of becoming more similar as the standardised effect size increases.

6.5.2.5 Strict Condition

The strict condition is $d_{pilot} > 0.5 \times d_{target}$ and d_{target}being contained within the 95% confidence interval of the pilot study. These results are presented in Table 6.9. These results are the most inflated, indicating that the strict condition pushes the truncation point further towards the true mean, inflating the observed mean.

Strict	80% Powered Main Trial				
Standardised Effect Size	Pilot Study SS	Main Trial SS	Pilot Trial Mean (SE)	Main Trial Mean (SE)	Ratio of Means
0.2	20	412	18.57 (9.01)	10.01 (3.53)	1.855
0.3	14	188	24.53 (11.35)	14.95 (5.11)	1.641
0.4	11	108	29.68 (13.08)	20.00 (6.76)	1.484
0.5	10	70	34.30 (13.96)	25.00 (8.48)	1.372
0.6	10	49	38.02 (14.96)	30.03 (9.93)	1.266
0.8	10	28	46.08 (15.80)	40.07 (13.28)	1.150

90% Powered Main Trial					
Standardised Effect Size	Pilot Study SS	Main Trial SS	Pilot Trial Mean (SE)	Main Trial Mean (SE)	Ratio of Means
0.2	28	552	16.62 (7.78)	10.05 (3.01)	1.654
0.3	19	252	22.71 (9.83)	14.99 (4.49)	1.515
0.4	15	145	27.80 (11.37)	19.90 (5.83)	1.397
0.5	12	95	32.87 (12.97)	25.05 (7.34)	1.312
0.6	11	68	37.93 (14.31)	30.10 (8.58)	1.260
0.8	10	39	46.21 (15.77)	39.98 (11.26)	1.156

Table 6.9: The results of $10,000$ simulations when moving from pilot study to main trial, dependant on the observed effect in the pilot study being greater than $0.5d_{target}$ and the 95% confidence interval containing the target effect.

Ratio of Means	Constant Effect Size (0.2)	
Power	Initial	Trials in Sequence
80%	1.121	1.29
85%	1.101	1.080
90%	1.068	1.059
95%	1.031	1.031
99%	1.005	1.006

80% Powered Main Trial		
Effect Size	Initial	Trials in Sequence
0.2	1.121	1.129
0.3	1.120	1.123
0.4	1.125	1.132
0.5	1.117	1.136
0.6	1.128	1.120
0.8	1.112	1.116

Table 6.10: The ratio of means (level of bias) for the **trials in sequence** simulation scenarios.

6.5.3 Ratio of Means

The simulations performed in this chapter can be compared using the ratio of means presented in each table. These have been collated and can be seen in Tables 6.10 and 6.11. When displayed in this way, the differences between the levels of bias can be seen more clearly.

With trials in sequence (Table 6.10) both sets of simulations show that the level of bias decreases with increasing power, yet remain relatively stable when the effect size changes. Recall that the initial trials are those which are defined in section 6.2, whilst the trials in sequence are presented in section 6.5.1

The pilot study to main study simulations show the varying levels of bias for the different progression conditions, as shown in Table 6.11. The two conditions with the least bias appear to be the condition advised by Lee *et al.*, and the conventional condition. This can be explained by the inclusion of a confidence interval in the condition. The first three conditions begin with $d_{pilot} > 0$, however the Lee and conventional conditions also state that d_{pilot} must be observed in a pre-specified confidence interval. This has the potential to reduce the upper tail of the distribution, thereby countering the bias imposed with the lower truncation and reducing the inflation of the observed mean.

The conservative condition provides the highest level of bias for varying power and effect size. This can be attributed to the condition which it imposes, that $d_{pilot} > 0.5 \times d_{target}$. Since $d_{target} > 0$, then 0.5 times this value will force the truncation point above 0 and create a greater level of bias and more inflated observed mean.

Ratio of Means	**80% Powered Main Trial**				
Effect Size	**Basic**	**Lee**	**Conventional**	**Conservative**	**Strict**
0.2	1.691	1.605	1.585	1.969	1.855
0.3	1.467	1.393	1.386	1.700	1.641
0.4	1.343	1.260	1.263	1.555	1.484
0.5	1.215	1.177	1.165	1.416	1.372
0.6	1.135	1.098	1.107	1.316	1.266
0.8	1.043	1.017	1.021	1.168	1.150

90% Powered Main Trial					
Effect Size	**Basic**	**Lee**	**Conventional**	**Conservative**	**Strict**
0.2	1.506	1.437	1.450	1.770	1.654
0.3	1.340	1.273	1.278	1.586	1.515
0.4	1.249	1.186	1.181	1.448	1.397
0.5	1.177	1.129	1.130	1.367	1.312
0.6	1.123	1.078	1.082	1.287	1.260
0.8	1.042	1.021	1.016	1.171	1.156

Table 6.11: The ratio of means (measure of bias) for the **pilot to main trial** simulation scenarios.

6.5.4 Standardisation of Observed Effects

The results in this chapter have demonstrated the bias which occurs when moving from one trial to the next based on various target effect sizes or powers. It is easier to compare the effect sizes if they are all on the same scale, so 40 is not being compared with 10. One way to do this is to use the ratio of means, as presented in the results tables and previous section. Another way to compare the observed effect sizes is to standardise them, similar to the technique used in chapter 4. This is done by dividing the observed mean difference by the standard deviation for each set of simulations.

Table 6.12 shows the standardised observed effects for the Phase II to Phase III results. Similarly to the results seen with the ratio of means, the standardised observed effect sizes show that the results for T1 are higher than T2, though they become more similar as the power increases. The standardised observed effect is fractionally higher for trials in sequence context than the initial simulations context for T1, but lower for T2. There is still evidence that as power increases the bias or difference between the means reduces, whilst the changing effect size with constant power does not deviate much.

Table 6.13 shows the standardised observed effect sizes for the various progression conditions for pilot study to main trial context. It can be observed that the lowest standardised observed effects are for the basic progression conditions, when $d_{pilot} > 0$. Since this is the minimum condition which would be deemed as "encouraging", this result shall be taken forward to the adjustment stage in chapter 7.

As the progression conditions intensify (from left to right in Table 6.13) it can be seen that the discrepency between the pilot and main trial results increase. This is due to the extra progression conditions imposing a more severe truncation point in the distributions for the pilot studies.

Constant Effect Size (0.2)				
	Initial Sims		**Trials in Sequence**	
Power	**T1**	**T2**	**T1**	**T2**
80%	0.225	0.201	0.226	0.200
85%	0.219	0.199	0.217	0.201
90%	0.213	0.199	0.213	0.201
95%	0.206	0.200	0.206	0.200
99%	0.201	0.200	0.201	0.200

Constant Power (80%)				
	Initial Sims		**Trials in Sequence**	
Effect Size	**T1**	**T2**	**T1**	**T2**
0.2	0.225	0.201	0.226	0.200
0.3	0.337	0.301	0.337	0.300
0.4	0.450	0.400	0.450	0.397
0.5	0.558	0.499	0.560	0.493
0.6	0.672	0.595	0.668	0.596
0.8	0.892	0.802	0.890	0.797

Table 6.12: The standardised observed effect sizes for **trials in sequence** simulations.

80% Powered Main Trial										
	Basic		**Lee**		**Conventional**		**Conservative**		**Strict**	
Effect Size	**Pilot**	**Main**	**Pilot**	**Main**	**Pilot**	**Main**	**Pilot**	**Main**	**Pilot**	**Main**
0.2	0.339	0.201	0.322	0.201	0.317	0.200	0.391	0.199	0.371	0.200
0.3	0.441	0.301	0.417	0.300	0.416	0.300	0.511	0.301	0.491	0.299
0.4	0.536	0.399	0.502	0.399	0.505	0.400	0.621	0.399	0.594	0.400
0.5	0.607	0.499	0.587	0.499	0.583	0.501	0.708	0.500	0.686	0.500
0.6	0.681	0.600	0.659	0.600	0.661	0.597	0.789	0.600	0.760	0.601
0.8	0.832	0.798	0.813	0.799	0.819	0.802	0.938	0.803	0.922	0.801

90% Powered Main Trial										
	Basic		**Lee**		**Conventional**		**Conservative**		**Strict**	
Effect Size	**Pilot**	**Main**	**Pilot**	**Main**	**Pilot**	**Main**	**Pilot**	**Main**	**Pilot**	**Main**
0.2	0.302	0.201	0.285	0.199	0.291	0.201	0.352	0.199	0.332	0.201
0.3	0.404	0.302	0.382	0.300	0.382	0.299	0.477	0.301	0.454	0.300
0.4	0.499	0.400	0.474	0.399	0.471	0.399	0.582	0.402	0.556	0.398
0.5	0.589	0.501	0.564	0.499	0.563	0.499	0.679	0.497	0.657	0.501
0.6	0.677	0.603	0.646	0.599	0.650	0.601	0.774	0.601	0.759	0.602
0.8	0.834	0.801	0.820	0.803	0.814	0.801	0.936	0.799	0.924	0.800

Table 6.13: The standardised observed effect sizes for **pilot study to main trial** simulations.

6.6 Discussion

The results presented in this chapter demonstrate that there is a regression to the mean effect when moving from one trial to the next. This is due to only significant trials or encouraging results moving forward to the next trial.

In its simplest form, the case of moving from one trial to the next (i.e. Phase II to Phase III) invokes a bias of around 12%, so the mean observed in trial 1 is inflated by that amount compared to the "true" mean. As the power increases, the bias decreases. This is highlighted by the movement of the truncation point of the distribution of T1 further from the mean, thereby reducing the average treatment effect for T1. For a 99% powered trial there is almost zero bias introduced.

A trial which targets 90% power is more likely to meet its target recruitment than a 80% powered trial (Sully et al., 2013). Studies with 80% power are less likely to recruit, and based on the results seen in this chapter they also have the most biased results. If a study is designed to have 80% power but fails to recruit the target sample size, it is likely that the resulting power of that study to be less than 80%. From personal discussion with trialists, investigators tend to use 80% power because they are aware that there will be problems in recruiting the sample size if the study were powered to 90%. The use of 80% is often an indication that recruitment will be difficult. This is then compounded by 80% powered trials experiencing problems with recruitment, thus becoming underpowered. Due to trials failing to meet 80% power, there is concern that any point estimates from previous work used in current sample size calculations are largely biased. Therefore, a power-based adjustment method could be useful for those trialists using previous research to design studies.

The changing of standardised target effect size had no effect on the bias, it remained constant dependent solely on the power of the studies. These simulations were completed under the alternative hypothesis, that there is a true treatment difference. When the treatment effect from T1 is used to calculate the sample size required for T2, the bias increases fractionally to between $12 - 14\%$.

For the pilot to main trial simulations, these demonstrated what is already deemed common knowledge, that the point estimate of a pilot study should not be used as the target effect size for designing a main trial. It can be clearly seen from each of the pilot study cases that the pilot study mean is far greater than that observed in the main trial for smaller to moderate effect sizes. If this overly ambitious mean were to be used to design the main trial, whilst the required sample size would be reduced the chance of seeing an effect size that large is incredibly small.

One difference between the results for the sequential trials and the pilot study simulations is that a change in effect size appears to have an impact on the level of bias for the pilot study simulations. As the effect size increases, the level of bias decreases.

This contrasts with the results observed for the sequential trials, which showed only slightly shifting in the bias when effect size changed. This could be attributed to the sample sizes for the pilot and main trials becoming more similar as the standardised effect size increases. It would also occur due to the "true" mean which the trial is being simulated around getting further away from zero. This would impact all the simulations where the condition is based on a positive observed treatment effect.

The level of bias observed in the pilot to main trial cases is greater than that observed in the Phase II to Phase III cases. This can be seen clearly in Tables 6.10 and 6.11.

A limitation of this research is that the pilot study to main trial sample sizes have been based on only one method of sample size estimation for pilot studies. This was primarily due to the time each simulation took. The implications of this include the possibility that the results for the bias are not generalisable to the other methods of pilot study sample size estimation. Therefore, these results should be interpreted with caution.

The pilot study simulations show a variety of progression conditions, with all other than the basic condition detailing a form of 'sufficient encouragement'. The basic case shall be taken forward to chapter 7 for an adjustment method to be developed and tested. This is due to the differing nature of each individual progression condition, yet all pilot studies must show a positive effect if they are to be moved on to a main trial.

6.7 Conclusions

This chapter has demonstrated that there is a regression to the mean effect which occurs when moving from one trial to the next. The level of bias varies depending on the power of the trials, as well as the design of the pilot study.

From the results, it is clear than an adjustment would be beneficial to allow trialists' to use previous data in a non-biased way to design future trials. This adjustment would be dependent on the power of the initial trial in the context of trials in sequence, as well as the design of the trials. This means whether it is previous research moving to a second trial, or a pilot study to main trial. Since pilot studies are not powered, the adjustment would be dependent on the continuation criteria, such as the observed effect size being positive.

In chapter 7, an adjustment which has been developed from the work presented in this chapter is further developed and compared with current adjustment methods discussed in chapter 5. These adjustment methods are rules-of-thumb, which whilst they are easy to recall and apply, could run the risk of being overly sensitive for higher powers. The requirement of a new method stems from the varying levels of bias depending on trial design as outlined in this chapter.

7. Development of Adjustment Methods

7.1 Introduction

A review of the literature in chapter 5 indicated that when using results from one trial to design the next, a phenomenon known as regression to the mean could be occuring. There are some general adjustments which have been developed (Kirby et al., 2012; Wang et al., 2006), however these are rule-of-thumb methods so could be overly sensitive to certain parameters such as sample size of trial. Regression to the mean was demonstrated in chapter 6, where a series of simulations showed the effects of regression to the mean in two common scenarios, trials in sequence and pilot study to main trial.

In this chapter, pilot studies will be considered under the context that they are deemed ‘successful’ if they observe $d_{obs} > 0$ only.

7.1.1 Chapter Aims

This chapter aims to show that the distributions observed in chapter 6 are consistent with a truncated Normal distribution. Conditional on this result, this chapter will mathematically confirm the results observed in chapter 6 and develop an adjustment by which the observed effects can be refined. The developed adjustment will be compared with two other adjustments identified from the review in chapter 5 and all three applied to the observed effect sizes in the first trial to assess whether the bias is reduced.

7.2 Normal Distributions

This section will focus on the standard Normal distribution and then progress to the truncated Normal distribution.

7.2.1 The Normal Distribution

The Normal distribution for random variable X is denoted $X \sim N(\mu, \sigma^2)$ for mean μ and variance σ^2. The standard Normal distribution occurs when $\mu = 0$ and $\sigma^2 = 1$, such that $X \sim N(0, 1)$.

The probability density function (pdf) for a standard Normal distribution is denoted ϕ, given by

$$\phi_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right). \tag{7.1}$$

for $-\infty < x < \infty$.

The cumulation distribution function (cdf) for the standard Normal distribution, denoted $\Phi(x)$, is given by

$$\Phi(x) = P(X \leq x) = \int_{-\infty}^{x} \phi(w) dw \tag{7.2}$$

where w is a dummy variable. The cdf gives the probability that $X \leq x$ for some value x in the domain.

The distributions observed in practice, however, are not usually standard Normal distributions. The standard Normal distribution can be transformed to represent known mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$. The probability density function (pdf), $\phi(x)$, is given by

$$\phi_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{7.3}$$

This is based on the domain $X \in (-\infty, \infty)$.

The expectation for a Normal distribution is given by

$$E[X] = \mu \tag{7.4}$$

and the variance is given by

$$Var[X] = \sigma^2. \tag{7.5}$$

These results will be used in the next section to demonstrate mathematically the cause of regression to the mean.

7.2.2 Truncated Normal Distribution

The truncation Normal distribution is the general Normal distribution bounded by a random variable from either above, below or both. This could occur if there was a floor- or ceiling-effect with the data, for example if blood-pressure values had to be above or below a pre-specified threshold to enter into a trial, there could be a truncation at that threshold.

7.2.2.1 Two-sided Truncated Normal

Suppose $X \sim N(\mu, \sigma^2)$ and let Y be a truncated Normal, denoted $TN(\mu, \sigma^2, a, b)$ where (a, b) are restrictions on the domain of X $(-\infty \leq a < b \leq \infty)$ (Johnson and Thomopoulos, 2002). These results are the two-sided results, such that there is a truncation point on either side of the distribution.

The probability density function of a truncated Normal distribution is given by

$$f\big(y|(a,b)\big) = \frac{\frac{1}{\sigma}\phi\big(\frac{y-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)} \tag{7.6}$$

for $a \leq y \leq b$ and $f(y) = 0$ otherwise.

The probability of X lying within the interval of (a, b) is given by

$$\Phi\bigg(\frac{b-\mu}{\sigma}\bigg) - \Phi\bigg(\frac{a-\mu}{\sigma}\bigg) \tag{7.7}$$

where $\Phi(..)$ is the cumulative distribution function described in Equation 7.2.

The expectation for the truncated Normal distribution is given by (Kortum, 2002; Olive, 2015),

$$E(Y) = \mu + \left[\frac{\phi\big(\frac{a-\mu}{\sigma}\big) - \phi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)}\right]\sigma \tag{7.8}$$

and the variance is given by

$$VAR(Y) = \sigma^2\left[1 + \frac{\big(\frac{a-\mu}{\sigma}\big)\phi\big(\frac{a-\mu}{\sigma}\big) - \big(\frac{b-\mu}{\sigma}\big)\phi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)}\right] - \sigma^2\left[\frac{\phi\big(\frac{a-\mu}{\sigma}\big) - \phi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)}\right]^2.$$

7.2.2.2 One-sided Truncated Normal

The results for a left-truncated Normal distribution are as follows, such that $b \to \infty$.

The probability density function of a truncated Normal distribution is given by

$$f\big(y|(a,b)\big) = \frac{\frac{1}{\sigma}\phi\big(\frac{y-\mu}{\sigma}\big)}{1 - \Phi\big(\frac{a-\mu}{\sigma}\big)} \tag{7.9}$$

for $a \leq y$ and $f(y) = 0$ otherwise.

The probability of X lying within the interval of (a, ∞) is given by

$$1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{7.10}$$

where $\Phi(..)$ is the cumulative distribution function described in Equation 7.2.

The expectation for the truncated Normal distribution is given by (Kortum, 2002; Olive, 2015), if we let $E(Y) = \mu*$ where $\mu*$ is the expectation of the truncated Normal distribution,

$$E(Y) = \mu* = \mu + \left[\frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right]\sigma \tag{7.11}$$

and the variance is given by

$$VAR(Y) = \sigma^2\left[1 + \frac{\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right] - \sigma^2\left[\frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right]^2.$$

This section has presented the pdf, cdf, mean and variance for a truncated Normal distribution. However, for this thesis the focus is primarily on the mean, since this can lead to the estimation of the target effect size. Thus, from this point onwards, the focus shall be on the equation for the mean.

As observed in chapter 6, the distribution which occurs when only the significant trials are taken forward is a one-sided truncated distribution. Due to this, the focus of the chapter will remain on the one-sided distribution.

7.2.3 Truncation Point

Based on the results observed in chapter 6, the truncation point of each distribution appears to be related to the power of the trial. This is also demonstrated in the previous section. Recall that power is 1 - the probability of making a Type II error, and in the simulations the power of each trial is known.

From the expectation equation (7.11), if we define $\mu*$ to be the expectation of the truncated Normal distribution, and let $b \rightarrow \infty$, this equation becomes

$$\mu* = \mu + \sigma\frac{\phi(A)}{1 - \Phi(A)} \tag{7.12}$$

where $A = \frac{a-\mu}{\sigma}$, μ is the expection or mean of the underlying Normal distribution (the untruncated Normal distribution) and σ is the population standard deviation.

It can be observed that $\mu* > \mu$ since $\sigma\frac{\phi(A)}{1-\Phi(A)} > 0$, so when the distribution is left-truncated, the mean is higher than the standard Normal expectation. This confirms the observations in chapter 6 where the mean in the initial trial (whether it was a pilot trial or phase II trial) is higher than the mean of the second trial. If one is able to find the truncation point, then it is possible to calculate the mean for the distribution if it were not truncated.

As shown in chapter 6, our data for the initial trial is left-truncated, i.e. it has a truncation point on the left side of the distribution and none on the right. This means that equation 7.7 for the probability of X lying in the area greater than a becomes

$$P[X > a] = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{7.13}$$

since $b \to \infty$. This equation looks similar to the expression for the power of a trial $(1-\beta)$. Recall β is the probability of making a Type II error, whilst the power of a trial is the probability of observing a difference if there truly is a difference to be observed (i.e. if the alternative hypothesis is true).

As we have deduced that the truncation point and the power are connected based on the results observed in chapter 6, one should be able to calculated the truncation point based on the power of a trial.

7.2.3.1 Finding the Truncation Point

Let us consider a Truncated Normal distribution with mean, $\mu = 10$, variance $\sigma^2 = 50^2$ and truncation points a and ∞. This forms a left-truncated Normal distribution, as shown in Figure 7.1.

A method to establish the truncation point depends on knowing the proportion of the distribution which has been 'cut off'. If we consider the simulations performed in chapter 6, for 80% power and a 5% significance level, it corresponds that the proportion of results which had a P-value< 0.05 is close to the power of the trial. Therefore, one could estimate the truncation point by taking the inverse of the cdf at the proportion β.

For a Normal distribution, this is given by $\Phi^{-1}(p)$, where p is the proportion of the distribution to the left of the cut-off, or truncation. Since in chapter 6 there were $10,000$ simulations to begin with, after the selection process subject to the condition that $P < 0.05$, we can calculate the proportion of trials still included as follows

$$p = \frac{\text{Number of Trials from T1 with } p < 0.05}{\text{Total Number of Trials (Simulations)}}. \tag{7.14}$$

This value is approximately equal to the power of the trials, since the trials are simulated under the alternative hypothesis.

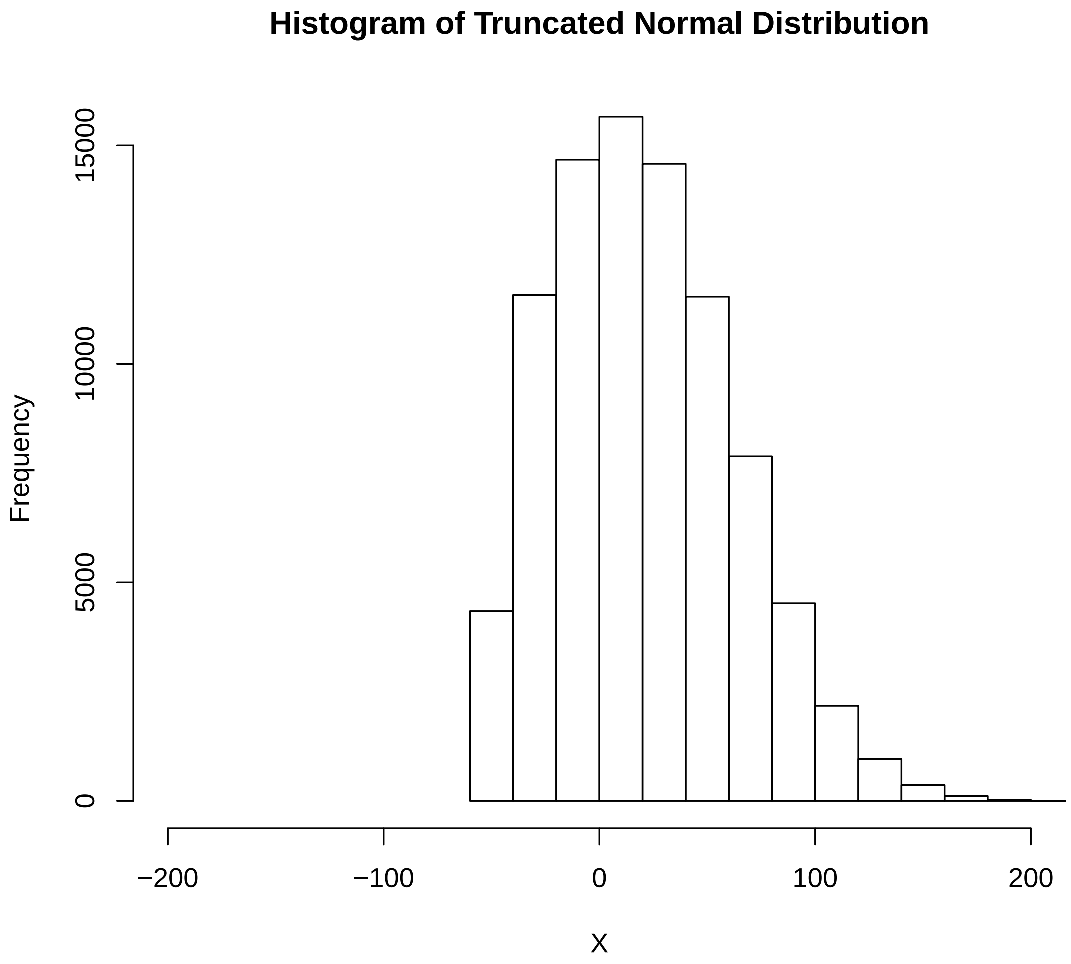


Figure 7.1: Histrogram showing a truncated Normal distribution.

This approach depends on the results of the trial, however, the truncation point can be calculated *a priori*. In chapter 6, the data which arises from the t-tests form a t-distribution. The results for a t-distribution of t-statistics, given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s/\sqrt{n}} \tag{7.15}$$

We can define $d = \bar{x}_1 - \bar{x}_2$. Under the null hypothesis, $\mu_1 - \mu_2 = 0$ so the equation becomes

$$t = \frac{d}{s/\sqrt{n}}. \tag{7.16}$$

If the number in each group can be assumed to be equal, the degrees of freedom are $2n - 2$. Therefore, the truncation point can be given by the proportion of trials excluded due to having $P \geq 0.05$ corresponding to the value $t_{2n-2,1-\alpha/2}$. Therefore, the truncation point for small samples ($n < 30$) could be calculated by taking the inverse cdf of a t-distribution with mean d, standard deviation s and $2n - 2$ degrees of freedom.

For large sample sizes ($n \geq 30$), the t-distribution approximates to a standard Normal distribution, therefore the adjustments posed in this chapter which are based on the Normal distribution hold.

Back in chapter 2, it was shown that the power of a trial can be calculated using a non-central t-distribution as

$$1 - \beta = 1 - T^{-1}\left(t_{1-\frac{\alpha}{2}, n_A(r+1)-2}, n_A(r+1) - 2, \sqrt{\frac{rn_Ad_S^2}{(r+1)\sigma^2}}\right), \tag{7.17}$$

where $T^{-1}(\dots)$ is the cumulative density function of a non-central t-distribution, with non-centrality parameter $\sqrt{\frac{rn_A}{(r+1)}}$ (Senn, 1993). In this chapter, the focus is on two-arm trials with $r = 1$ and the non-centrality parameter becomes

$$\sqrt{\frac{n}{2}\frac{d_S^2}{\sigma^2}}. \tag{7.18}$$

Chapter 2 also stated the justification for the t-distribution being non-central since the power is estimated under the alternative hypothesis, therefore the distribution would be non-central.

The simulations presented in chapter 6 were also performed under the alternative hypothesis, hence a non-centrality parameter should be introduced as described above. It can be observed that

$$\sqrt{\frac{d_S^2}{\sigma^2}} \tag{7.19}$$

is the standardised effect size, which can be denoted ES. Therefore, the non-centrality parameter becomes

$$ES \times \sqrt{\frac{n}{2}}. \tag{7.20}$$

The distribution of the effect sizes multiplied by $\sqrt{n/2}$ gives a Normal distribution $N(ES\sqrt{n/2}, 1)$.

Let $E(Y)$ be deonted $\mu*$, which is the mean of the truncated Normal distribution. Since the truncation point, a, can be calculated using $t_{2n-2,1-\alpha/2}$, and the truncated mean $\mu*$ is known, using equation 7.12 and re-arranging in terms of the true mean μ, gives

$$\mu = \mu * -\sigma\frac{\phi(A)}{1 - \Phi(A)} \tag{7.21}$$

where $A = \frac{a-\mu}{\sigma}$

7.2.3.2 Power-based Truncation Point

The results from section 7.2.3.1 can be used as a method of investigating the truncation point which is based on the concept of the minimum detectable difference used in chapter 4. The minimum detectable difference (MDD) is the smallest difference that can be statistically detected in a particular study (Cook et al., 2014) and is related to the truncation point. There are a number of published methods to calculate the minimum detectable difference, a few of which are outlined in this section.

Valk and colleagues (Valk et al., 2000), along with Pijls and colleagues (Pijls et al., 1999), described the "smallest detectable difference" for two measurements (test-retest) for a single group as

$$\sqrt{(Z_\alpha + Z_\beta) \times \frac{\sigma^2}{n}} \tag{7.22}$$

where α is the significance level, $1 - \beta$ is the statistical power, σ^2 is the variance of the within-person differences and n is the number of observations.

Another method, used for two independent groups (equal size and variance), is given by (Bridges, 1997; Hanson et al., 2003b)

$$\sqrt{2}\big(t_{\alpha,\nu} + t_{\beta,\nu}\big)\frac{SD}{\sqrt{n}} \tag{7.23}$$

where $t_{\alpha,\nu}$ (and $t_{\beta,\nu}$ respectively) are the $100(1-\alpha)$ percentile of the t-distribution with ν degrees of freedom, n is the number of observations per group and α and β are the significance level and Type II error.

One intuitive way to calculate it is when the power is set to 50%, this gives the minimum value that the 95% confidence interval around the point estimate will exclude the null value. Using the simplified equation from chapter 2,

$$n = \frac{2\sigma^2\big(Z_{1-\beta} + Z_{1-\alpha/2}\big)^2}{d^2} \tag{7.24}$$

with the power is set to 50%, the term $Z_{1-\beta}$ becomes equal to 0. If this is rearranged to be in terms of d, it becomes

$$d_{det} = \sqrt{\frac{2\sigma^2 Z^2_{1-\alpha/2}}{n}} \tag{7.25}$$

where d_{det} is the detectable difference, α is the significance level, n is the sample size per arm and σ^2 is the variance.

Power	x
80	0.700
81	0.691
82	0.682
83	0.673
84	0.663
85	0.654
86	0.645
87	0.635
88	0.625
89	0.615
90	0.605
95	0.544
99	0.457

Table 7.1: The adjustment values for the detectable difference. **Note: x is the value by which the target difference d should be multiplied.**

If the P-value achieved in the trial is equal to 0.05, and the sample size is achieved, then the ratio of the detectable difference to the target difference can be calculated by

$$\frac{d_{det}}{d}$$

for various powers. This will provide an adjustment value for the target difference, d, showing the minimum detectable difference. The results can be seen in Table 7.1.

The values in Table 7.1 are the values of the detectable difference. If a trial has a P-value of 0.05, for 80% power then the value of the detectable difference would be $0.7d$, where d is the target difference. These values are linked to the truncation point described earlier in the chapter, since the truncation point is the value at which the P-values become significant and the detectable difference is the proportion of d which will observed a P-value of 0.05.

7.2.4 Comparison of Simulation Results and Truncation Results

It is now possible to compare the results observed in the simulations in chapter 6 with those calculated in Section 7.2.3, both in terms of the truncation point and the ratio of bias between the observed (truncated) and true mean.

The results of the truncation point calculations is separated into two sections. The first intermediary section is presented in Appendix E. The final results are presented in Table 7.2 and Table 7.3. Each of the variables shall be further explained to ensure clarity. Since the values we are interested in are ratios, they are scale-independent.

These calculations are based on a significance level $\alpha = 0.05$. The results presented in this section are the results for the standardised effect size, where the effect size in the tables in the Appendix are multiplied by $\sqrt{2/n}$. These tables demonstrate that the truncation point based on the t-distribution, a, matches the truncation point as calcuated using Table 7.1. The results in the Appendix are for $\mu = ES\sqrt{n/2}$ where ES is the effect size and $\sigma = 1$.

Each variable in the results tables (Tables 7.2 and 7.3) are described below. The values for μ are those of the true effect sizes and $\mu*$ show the biased mean estimate on the standardised scale.

- Sample size per arm (n) is set to the same values calculated in chapter 6
- μ is the mean difference, based on the non-central t-distribution. Thus $\mu = ES$ where ES is the effect size.
- a_{det} is calculated by $x.\mu$, where x is the associated value from Table 7.1
- a is the mathematical truncation point calculated as $a = t_{2n-2,1-\alpha/2}$
- $\mu*$ has been calculated using Equation 7.12.

The ratios for the chapter 6 bias are brought forward from Tables 6.3 and 6.4 and calculated by taking the inverse of the ratio. The ratio of $\mu/\mu*$ is calculated from the data in Table 7.2 and Table 7.3. It can be observed that $a_{det} \approx a$, indicating that the truncation point as calculated mathematically (a) is equivalent to the truncation point calculated by the detectable difference (a_{det}).

The difference between $\mu*$ and μ shows the amount of bias which occurs from the truncation of the underlying distribution, with $\mu*$ being the mean of the truncated distribution (or the observed mean) and μ being the mean of the untruncated distribution (the 'true' mean').

The observed ratio from chapter 6 is approximately the same as the ratio of the mathematical means ($\mu/\mu*$).

These results confirm the observations from chapter 6, and provide a mathematically sound solution for the truncation which occurs for trials in sequence.

Trials in Sequence							
Effect Size = 0.2							
	Sample	**Truncation**		**Mean Difference**		**Ratio**	
Power	**Size (n)**	a_{det}	a	μ	$\mu*$	**Ch 6 Bias**	$\mu/\mu*$
80	393	1.962	1.963	0.200	0.225	0.885	0.889
85	450	1.962	1.963	0.200	0.218	0.926	0.916
90	526	1.962	1.962	0.200	0.212	0.945	0.943
95	651	1.963	1.962	0.200	0.206	0.970	0.971
99	950	1.960	1.961	0.200	0.201	0.994	0.994

Table 7.2: A comparison of mathematically calculated truncation points and ratios of mean differences with simulated values for various powers, having multiplied the Effect Sizes (ES) by $\sqrt{2/n}$.

Trials in Sequence							
Power = 80%							
	Sample	**Truncation**		**Mean Difference**		**Ratio**	
Effect	**Size (n)**	a_{det}	a	μ	$\mu*$	**Ch 6 Bias**	$\mu/\mu*$
0.2	393	1.962	1.963	0.200	0.225	0.886	0.889
0.3	175	1.962	1.963	0.300	0.338	0.891	0.889
0.4	99	1.962	1.963	0.400	0.450	0.884	0.889
0.5	64	1.962	1.963	0.500	0.561	0.880	0.891
0.6	45	1.962	1.963	0.600	0.672	0.892	0.892
0.8	26	1.962	1.963	0.800	0.893	0.896	0.896

Table 7.3: A comparison of mathematically calculated truncation points and ratios of mean differences with simulated values for various effect sizes, having multiplied the Effect Sizes (ES) by $\sqrt{2/n}$.

7.2.5 Pilot Study to Main Trial

For the pilot study to main trial results, the adjustment can be solved mathematically by setting the truncation point, a, equal to zero. This is based on the premise that a pilot study will continue to a main trial if the observed effect is positive, which is the least strict condition of continuation discussed in chapter 6 (the basic case). The results are presented in Tables 7.4 and 7.5.

It can be observed that the ratio of mean differences observed from the simulations in chapter 6 are similar to the ratio of mathematical mean differences developed in this chapter. This confirms the results observed in chapter 6 and the mathematical bias are the same.

Pilot to Main Trial Main Trial Power = 80%						
	Sample	**Truncation**	**Mean Difference**		**Ratio**	
Effect	**Size (n)**	a	μ	$\mu*$	**Ch 6 Bias**	$\mu/\mu*$
0.2	20	0	0.200	0.340	0.591	0.588
0.3	14	0	0.300	0.440	0.682	0.682
0.4	11	0	0.400	0.533	0.745	0.751
0.5	10	0	0.500	0.610	0.823	0.820
0.6	10	0	0.600	0.680	0.881	0.883
0.8	10	0	0.800	0.837	0.959	0.955

Table 7.4: A comparison of mathematically calculated truncation points and ratios of mean differences with simulated values for various effect sizes for **pilot study to main trial**.

Pilot to Main Trial Main Trial Power = 90%						
	Sample	**Truncation**	**Mean Difference**		**Ratio**	
Effect	**Size (n)**	a	μ	$\mu*$	**Ch 6 Bias**	$\mu/\mu*$
0.2	28	0	0.200	0.304	0.664	0.657
0.3	19	0	0.300	0.403	0.747	0.745
0.4	15	0	0.400	0.493	0.801	0.812
0.5	12	0	0.500	0.587	0.850	0.853
0.6	11	0	0.600	0.669	0.891	0.897
0.8	10	0	0.800	0.837	0.960	0.955

Table 7.5: A comparison of mathematically calculated truncation points and ratios of mean differences with simulated values for various effect sizes for **pilot study to main trial**.

These tables show that the mean which occurs under the truncated Normal distribution ($\mu*$) is more inflated than both the 'true' mean (μ) and the corresponding observed mean from Tables 7.2 and 7.3.

In contrast with the trials in sequence results, where the ratios remained approximately constant as the effect size varied, the ratios presented in Tables 7.4 and

7.5 increase as the effect size increases. This is due to the ratio of the pilot study sample size and main trial sample size. As the effect size increases, the main trial sample size decreases dramatically but the pilot study sample size is capped at a minimum of 10 participants per group. Therefore as the effect size increases, the ratio of $Pilot_{SampleSize}/Main_{SampleSize}$ increases.

These adjustments appear to be more extreme than those in Table 7.6, which is due to the inflated mean observed in the pilot studies. This inflated mean is due to the very small sample size used in pilot studies.

As the effect size increases the strength of the adjustment required decreases, with the adjustment for an effect size of 0.8 being equal irrespective of whether the power of the main trial is 80% or 90%.

As previously discussed in chapter 6, the pilot study sample sizes are based on recommendations by Whitehead et al. (Whitehead et al., 2016), however this section has demonstrated that the results of the mathematical solution for this context actually depend on the conditions set for moving forward with the main trial. For example, these results are all based on a truncation point occurring at 0 since if $d < 0$ then the intervention is assumed to be not promising. Therefore, whilst it is true that there was limited generalisability of the simulation results, the mathematical results presented and used for the remainder of the thesis are independent of this and depend only on the condition of a promising pilot study. This will ensure that the adjustment is not restricted to one particular method for sample size calculation relating to a pilot study-main trial context.

7.3 Development of the Adjustment Method

This section will detail the possible adjustment method which has been developed based on the results from chapter 6 and the results seen thus far in this chapter. For simplicity, the adjustment method developed in this chapter will be referred to as the Rothwell adjustment.

7.3.1 Rothwell Adjustment

An adjustment to be applied to the results of trial 1 (T1) can be taken from Table 7.2 and Table 7.3. This adjustment, which is the ratio of $\mu * /\mu$ in the table, is calculated by taking the inverse of the ratio of means, thereby providing the level of bias by which to adjust the results from trial 1.

Trials in Sequence	
Effect Size = 0.2	
Power	**Rothwell Adjustment (x)**
80	0.89
85	0.92
90	0.94
95	0.97
99	0.99
Power = 80%	
Effect	**Rothwell Adjustment (x)**
0.2	0.89
0.3	0.89
0.4	0.89
0.5	0.89
0.6	0.89
0.8	0.90

Table 7.6: The Rothwell adjustments for **trials in sequence** for various powers and effect sizes. **Note: x is the value by which the observed difference $\mathbf{d_{T1}}$ should be multiplied.**

For the trials in sequence, it was shown in chapter 6 that a change in standardised effect size had little impact on the bias, it was more affected by the change in power. Since the simulation results from chapter 6 are validated using the mathematical results shown in Table 7.2, the associated Rothwell adjustments shall be taken from the mathematical solution and are shown in Table 7.6. Therefore, the results presented in Table 7.6 are for a constant effect size of 0.2. Results are given to 2 decimal places where appropriate for ease of use. All the tables containing adjustment values in this chapter are showing the value by which the observed effect size or difference in Trial one should be multiplied. This gives a more unbiased estimate of the 'true' effect size.

Pilot Study to Main Trial 80% Powered Main Trial	
Effect Size	**x** (Rothwell)
0.2	0.59
0.3	0.68
0.4	0.74
0.5	0.82
0.6	0.88
0.8	0.96

90% Powered Main Trial	
Effect Size	**x** (Rothwell)
0.2	0.66
0.3	0.75
0.4	0.80
0.5	0.85
0.6	0.89
0.8	0.96

Table 7.7: The adjustment values for **pilot study to main trial** designs. **Note: x is the value by which the observed effect $\mathbf{d_{pilot}}$ should be multiplied.**

The results from the simulations and the mathematical solution shown earlier in this chapter demostrate that irespective of the value of μ, where μ is the 'true' mean, as long as the power of trial 1 is known then the adjustment can be applied to the observed effect size from Trial 1. The value of the adjustment depends solely on the power of trial 1, thus as long as the power of the trial is known, and the observed effect size $\mu*$, then the result can be adjusted as appropriate for the values shown in Tables 7.6 and 7.7. From this point forwards, the adjustment values will be as in Table 7.6 for trials in sequence and Table 7.7 for pilot to main trial.

As discussed in the previous section (section 7.2.5), the adjustment for trials in sequence is dependent on the power of the first trial. The adjustment for the pilot study to main trial context is dependent on the continuation criteria imposed on the pilot study, for example this could be that the observed effect size was positive ($d_{pilot} > 0$) or that the 90% confidence interval contains the MCID.

7.4 Comparison of Other Methods

This section compares the adjustment developed in this chapter with adjustments described in chapter 5. For the context of trials in sequence, the adjustments can be compared with those proposed by Wang and Kirby. For the pilot to main trial context, since there is no discussion in the Wang and Kirby papers about the application of their respective adjustments to pilot studies, only the Rothwell adjustment will be applied to the simulation data and the results discussed.

Recall, the first of the previous adjustments is a subtractive method proposed by Wang *et al.* (Wang et al., 2006),

$$\hat{\Delta} - 1 \times s.e(\hat{\Delta}) \tag{7.26}$$

where Δ is the observed difference and $s.e(\hat{\Delta})$ is the standard error of the observed difference from the first trial .

The second method is a multiplicative method which was proposed by Kirby *et al.* (Kirby et al., 2012), it provides a more general rule-of-thumb,

$$\hat{\Delta} \times 0.9. \tag{7.27}$$

7.5 Adjustments for Trials in Sequence

Thus far in the chapter, the various adjustments have been discussed in detail. In order to fully compare these adjustment and assess their individual merits, each adjustment is applied to the simulated data and the results discussed.

7.5.1 Methods

The simulation results in chapter 6 have been used, with the observed mean differences in trial 1 having each adjustment method applied. The values which form the various adjustment methods are presented in Table 7.8. It can be seen that the Kirby adjustment is constant, the Wang adjustment is sensitive to high effect sizes, and the Rothwell adjustment varies slightly but remains constant for increasing effect size. There are two point at which the adjustments can be applied, the first being to the overall average results across the $10,000$ simulations, the second being during each individual simulation. Both applications are presented in this section, with the methods for both being near idenitical except for the point of adjustment.

Trials In Sequence			
Varying Power, Constant Effect Size			
Power	**x** (Rothwell)	**x** (Kirby)	**y** (Wang)
80	0.89	0.90	0.139
85	0.92	0.90	0.126
90	0.94	0.90	0.113
95	0.97	0.90	0.097
99	0.99	0.90	0.074
Varying Effect Size, Constant Power			
Effect	**x** (Rothwell)	**x** (Kirby)	**y** (Wang)
0.2	0.89	0.90	0.139
0.3	0.89	0.90	0.309
0.4	0.89	0.90	0.551
0.5	0.89	0.90	0.853
0.6	0.89	0.90	1.207
0.8	0.90	0.90	2.093

Table 7.8: All adjustment methods results for trials in sequence by power. **Note: x is the value by which the observed difference $\mathbf{d_{T1}}$ should be multiplied. y is the value which the observed difference from T1 should be subtracted by.**

7.5.2 Results

The results for the implementation of the adjustments for bias are given in this section for trials in sequence. As there were two possible points at which to introduce the adjustment, both have been presented.

7.5.2.1 Application to Average Trial 1 Results

These adjustments have been applied to the average results from T1 across the $10,000$ simulations, presented in Table 7.9. It can be seen that the Rothwell adjustment produces consistent results as the power varies, with the adjusted T1 value being close to the 'true' effect. The Kirby adjustment functions well for 80% power and varying effect size with constant power of 80%, however, as the power increases the adjustment becomes more severe. From the opposite side, the Wang adjustment appears to be conservative for varying effect size and only begins to function well for very high powers.

7.5.2.2 Application to Individual Trial 1 Results

The results for the trials in sequence are shown in Table 7.10. The effect of these adjustments is that the truncated distributions observed in chapter 6 are shifted by a fixed amount, causing the mean to more closely follow that for the second trial. It can be observed in the table the effect of implementing the adjustments to

Trials In Sequence - Adjusted Average Mean Differences			
Effect Size= 0.2	Adjusted T1 Mean Difference		
Power	Rothwell	Kirby	Wang
80	10.06	10.17	11.16
85	9.98	9.77	10.72
90	10.00	9.58	10.53
95	9.99	9.27	10.20
99	9.96	9.05	9.99

Power= 80%	Adjusted T1 Mean Difference		
Effect	Rothwell	Kirby	Wang
0.2	10.06	10.17	11.16
0.3	15.00	15.17	16.54
0.4	20.01	20.24	21.94
0.5	24.93	25.21	27.16
0.6	29.73	30.06	32.19
0.8	40.03	40.03	42.39

Table 7.9: The adjusted average means from T1 using each method, with the ‘true’ difference shown.

individual trial data. For the individually adjusted data, the Kirby adjustment does not appear to perform as well as it did for the average mean adjustment, whilst the Rothwell adjustment again proves to be relatively stable when compared with the ‘true’ effect. The Wang adjustment works well for high powers, but does not reduce the bias enough for lower powers.

These results can be seen in Figures 7.2 and 7.3. Figure 7.2 shows that the Rothwell adjustment follows the ‘true’ mean difference more closely than the other adjustments, with the Kirby adjustment moving away from the ‘true’ difference as the power increases. Opposing this, the Wang adjustment and the unadjusted difference get closer to the ‘true’ mean difference as the power increases.

Figure 7.3 shows that as the effect size increases with constant power, the unadjusted mean difference and the Wang adjusted values move away from the ‘true’ value, whereas the Kirby and Rothwell methods remain constantly close to the ‘true’ difference.

Trials In Sequence - Individually Adjusted Mean Differences			
Effect Size= 0.2	Adjusted T1 Mean Difference		
Power	Rothwell	Kirby	Wang
80	10.06	9.70	10.72
85	10.16	9.83	10.43
90	9.99	9.54	10.13
95	10.01	9.26	9.90
99	9.96	9.04	9.72

Power= 80%	Adjusted T1 Mean Difference		
Effect	Rothwell	Kirby	Wang
0.2	10.06	9.70	10.72
0.3	15.03	15.19	15.58
0.4	19.96	20.20	20.25
0.5	25.35	25.25	24.55
0.6	30.02	30.26	28.60
0.8	39.97	40.02	35.86

Table 7.10: The average of individually adjusted means in T1 using each method, with the ‘true’ difference shown.

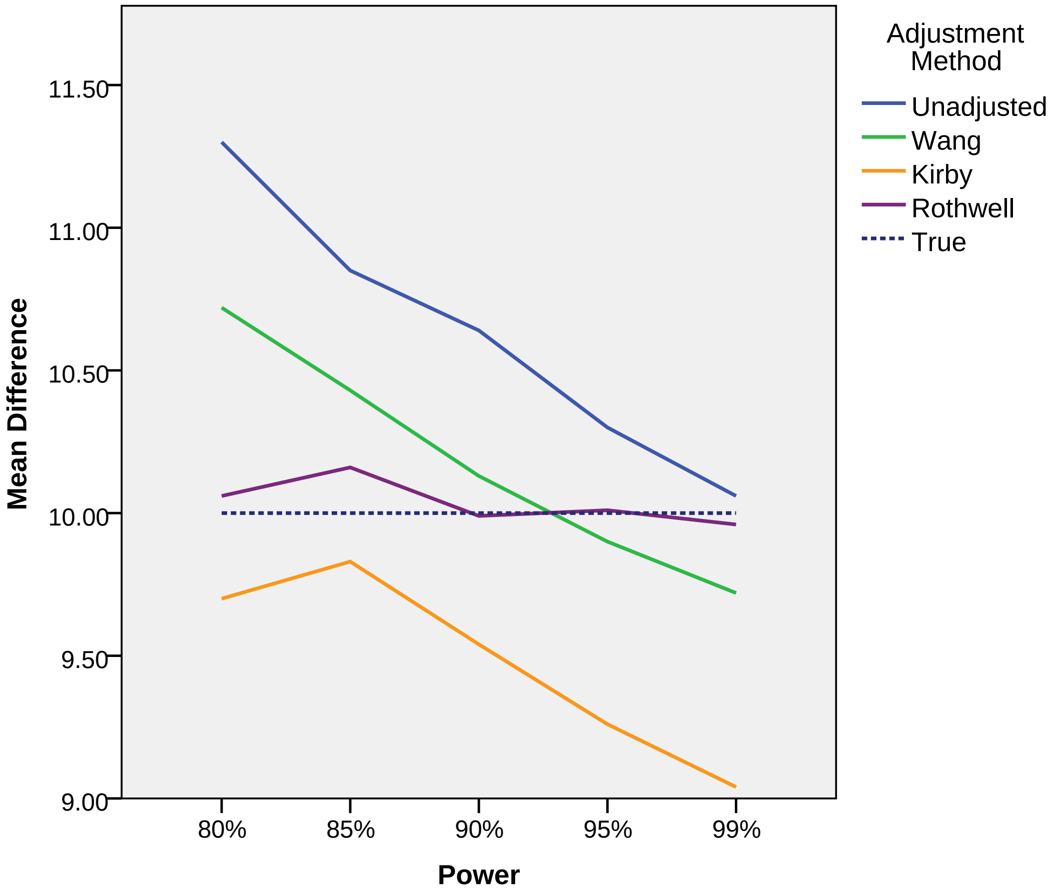


Figure 7.2: A line plot showing the various adjustment methods for **trials in sequence** compared with the unadjusted values and the 'true' value for different powers.

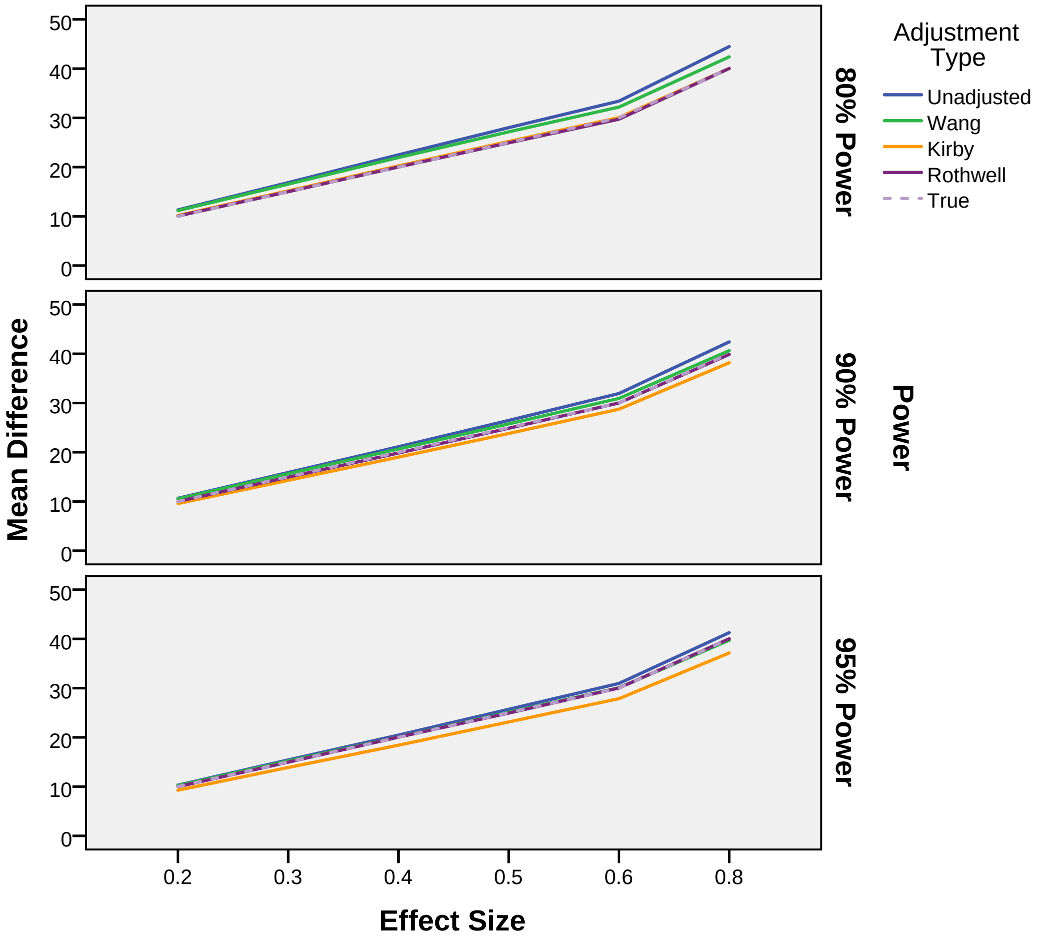


Figure 7.3: A line plot showing the various adjustment methods for **trials in sequence** compared with the unadjusted values and the 'true' value for different effect sizes.

7.6 Adjustment for Pilot Study to Main Trial

As previously discussed in this chapter and chapter 6, it is not advised to use the point estimate from a pilot study due to the sample size being very small, it was of interest to see the contrast between the adjustment presented for this context and the one for trials in sequence. The Rothwell adjustment was the only adjustment to be considered for this context due to there being no discussion in the papers by Kirby and Wang of using their adjustments for pilot study data. The values of the adjustments for pilot study results can be seen in Table 7.7.

7.6.1 Methods

The Rothwell adjustment was applied to the data from chapter 6 under the context of a pilot study to main trial. Once more, this adjustment could be applied at two different points of the simulations, with each being presented in turn.

7.6.2 Results

The results of adjusting for a regression to the mean effect for a pilot study to main trial context have been split in two sections by at which point the adjustment was implemented. These results should be taken with caution, as mentioned earlier.

7.6.2.1 Average Initial Trial Results

The results when the adjustment displayed in Table 7.7 is applied to the observed mean of a pilot study can be seen in Table 7.11. The adjustment does appear to work well in this scenario, providing a closer estimate of the 'true' mean difference than that observed in the pilot study.

7.6.2.2 Individual Initial Trial Results

The adjustments were also applied to each simulation, to assess whether on a trial by trial basis they reduce the truncation effect of the results. Figure 7.4 shows the adjusted results for various effect sizes and power for the pilot study to main trial scenario. The adjustment implemented is the one detailed in Table 7.7. It can be seen in the graph that the adjusted values from the pilot studies are inflated when compared with the 'true' mean. The Rothwell adjustment for pilot studies shows that the observed effect size can be modified to follow more closely the 'true' effect size. This could be a useful adjustment for point estimates in pilot studies to allow them to be used more effectively in designing the main trial.

Pilot Study to Main Trial		
80% Powered Main Trial		
Effect Size	**'True' Mean**	**Adj. Pilot Mean (Rothwell)**
0.2	10	10.01
0.3	15	15.00
0.4	20	19.84
0.5	25	24.87
0.6	30	29.96
0.8	40	39.95
90% Powered Main Trial		
Effect Size	**'True' Mean**	**Adj. Pilot Mean (Rothwell)**
0.2	10	9.95
0.3	15	15.15
0.4	20	19.96
0.5	25	25.04
0.6	30	30.11
0.8	40	40.05

Table 7.11: The adjusted average means from the pilot study, by power and effect size.

Pilot Study to Main Trial		
80% Powered Main Trial		
Effect Size	**'True' Mean**	**Adj. Pilot Mean (Rothwell)**
0.2	10	10.10
0.3	15	14.97
0.4	20	19.58
0.5	25	24.90
0.6	30	29.93
0.8	40	40.16
90% Powered Main Trial		
Effect Size	**'True' Mean**	**Adj. Pilot Mean (Rothwell)**
0.2	10	10.16
0.3	15	14.93
0.4	20	19.66
0.5	25	24.90
0.6	30	30.00
0.8	40	40.52

Table 7.12: The average adjusted individual means from the pilot study.

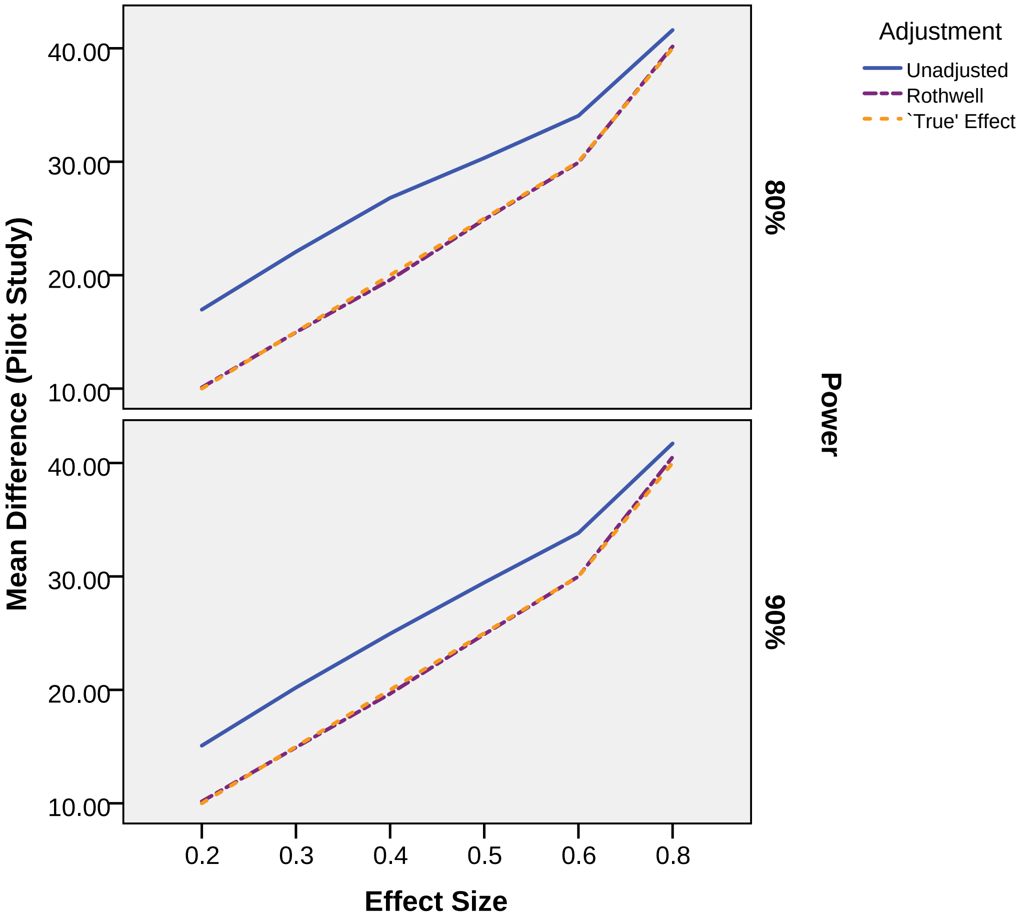


Figure 7.4: A line plot showing the various adjustment methods for **pilot study to main trial** compared with the unadjusted values and the 'true' value for different effect sizes.

7.7 Effect on Sample Size

As discussed in early chapters, the sample size calculation is very sensitive to the target difference, *d*. If the adjustments discussed in this chapter are applied to the observed mean in the initial trial, then the sample size for T2 is calculated, they will be very different to those sample size calculations which would be based on the unadjusted mean difference observed in T1.

To assess this difference, Table 7.13 shows the unadjusted sample sizes for T2, along with the sample sizes based on each adjustment method. It can be observed that all the adjustments result in a higher sample size for T2 than that of T1. This is not unexpected, since all the adjustments reduce the observed mean difference, and as described in chapter 2 as the target difference decreases the sample size increases. As the power increases, the adjustment by Kirby results in a much greater sample size than that used in T1.

Trials In Sequence					
Varying Power, Effect Size= 0.2.			**Adjusted Sample Size (T2)**		
Power	**SS T1**	**Unadj. SS T2**	Rothwell	Kirby	Wang
80	393	364	460	490	406
85	450	456	519	554	491
90	526	556	633	691	610
95	651	651	782	913	800
99	920	1074	1096	1333	1158
Varying Effect Size, Power= 80%.			**Adjusted Sample Size (T2)**		
Effect	**SS T1**	**Unadj. SS T2**	Rothwell	Kirby	Wang
0.2	393	364	460	490	406
0.3	175	164	206	201	191
0.4	99	92	116	114	114
0.5	64	60	72	73	77
0.6	45	42	52	51	57
0.8	26	24	30	30	37

Table 7.13: The average sample size per arm for T2 based on the adjusted mean differences from T1.

7.8 Worked Examples

This section will briefly provide two worked examples to demonstrate how the Rothwell adjustment is to be implemented.

7.8.1 Trials in Sequence Context

In this example, let us consider a trial of a new treatment to reduce systolic blood pressure compared with an existing drug. The effect size is the reduction in systolic blood pressure. The first trial (the Phase II trial) has been performed and produced an average observed effect size of 12 mmHg which we wish to use to guide the planning for the Phase III trial. In order to properly adjust this effect size to reduce the effect of regression to the mean, one must first consider the power planned of the Phase II trial. Suppose this is 90%. Therefore, the adjustment will be 0.94, which will give an adjusted target effect size of 11.28. This value could be used as possiblye plausible effect size for the sample size calculation for the Phase III trial.

7.8.2 Pilot Study Context

Let us consider a trial investigating the effect of group therapy on depression. The outcome measure is the self-reported PHQ-9 questionnaire. A reduction of 10 points is deemed clinically important. The pilot trial observed an effect size of 16 point reduction after 3 months of therapy. Then designing the pilot trial, the study design of the main trial was considered using the approach of Whitehead (Whitehead et al., 2016). The main trial is planned to have a power of 80%, and using a standardised expected effect size of 0.2 (assuming a standard deviation of 50). Due to this a sample size of 20 was used for the pilot trial. It was decided that the main trial would start if $d_{pilot} > 0$. For this decision rule, an adjustment of 0.59 would be used. Therefore, using the effect observed in the pilot trial, a plausible estimate of the treatment difference would be 16×0.59 which is 9.44. This is also around the value that is of clinical importance. This gives reasonable confidence and justification to use 10 as a clinically important effect in the sample size calculation for the main trial. The power planned for the main trial, per the Whitehead method, has an effect on the sample size used for the pilot study. If the power is planned for 80%, this would result in a smaller sample size used in the pilot study compared to a main trial of 90% power, therefore increasing the standard deviation.

7.9 Discussion

This chapter has explained what a truncated Normal distribution is, as well as discussed some possibly solutions to find the point of truncation in the distribution. The truncation point has been shown mathematically to match up with the detectable difference calculation, and the ratio of bias observed in chapter 6 matches the ratio of the truncated and true means. All this shows that the simulations performed in chapter 6 hold true and the results have been mathematically confirmed.

The adjustments for the early data found in chapter 5 were compared to one developed in this chapter based on the simulations in chapter 6. It has been shown that the adjustments by Kirby and Rothwell work well on simulated data for a variety of powers and effect sizes. However, the Kirby adjustment seems to be too conservative for higher powers compared to the Rothwell adjustment proposed in this dissertation. The Rothwell adjustment is relatively stable across the various powers and effect sizes, with the Kirby adjustment performing similarly well for various effect sizes and the Wang adjustment being too affected by the sample size for higher powers, resulting in the adjusted mean differences being too conservative for lower powers.

With regards to the sample size for T2 based on the results from T1, there needs to be some consideration about the trade-off between the number required per arm and the chance of 'success' of the trial. Based on the results in Table 7.13, it can be seen that the Kirby adjustment results in needing generally more participants than the Rothwell adjustment, however, the proportion of trials at T2 which show $P < 0.05$ is fractionally greater for the Kirby adjusted samples than the other samples. This could be chance, or due to the increased number of participants. Sometimes it is not feasible to increase the number of participants by that great an amount, therefore this could become a problem with regards to implementing in practice.

In the next chapter, the Rothwell adjustment shall be assessed on a large real data set to test its robustness on the variability of real-life data as opposed to the simulated data.

7.9.1 Limitations

The limitations of this section fall largely with the time constraint of writing this thesis and the computer power required for each set of simulations when assessing how well each adjustment performs. However, since the results have also been proven mathematically this mitigates this factor considerably. For further work, it would be important to test the adjustments on more simulations and a wider variety of powers and effect sizes to provide a comprehensive table of adjustments which trialists can refer to. The table would contain each combination of power and effect size, with the associated adjustment value. This could be useful for trialists when designing studies based on previous work to ensure that the regression to the mean effect was reduced.

Another important limitation is this work has assumed other trials have the same endpoint. This is not always the case, in fact it is common for the initial trial to be performed on a surrogate end point.

The chapter has focused on the mean of the one-sided truncated Normal distribution. The results could be extended to both the two-sided distribution and the variance, however this is outside the scope of this thesis. It could be considered for further work.

7.10 Conclusion

This chapter has shown that the work completed in a simulated context in chapter 6 can be demonstrated mathematically, and the results for both are similar. There does exist a bias when the results from the first trial come from a truncated Normal distribution, which leads to the adjustments further developed in this chapter being required.

The adjustment developed in this chapter has been applied to simulated data, which emulates a 'perfect' trial scenario. The Rothwell adjustment performed well on the simulated data sets, and was more consistent at higher powers than the Kirby or Wang adjustments. However, there is a practical consideration with this adjustment, as the likelihood of trials having greater than 90% power is small.

The Rothwell adjustment would be useful for trialists when designing trials based on previous research. Once the power and approxiamte target effect size are known, the appropriate adjustment could be applied to the previous data to reduce the possible effect of regression to the mean. This could, in the long term, reduce the number of trials which do not achieve statistical significance if there is truly a difference between treatments. It is always possible that the previous work were anomalies, uncharacteristically high differences even if there is no difference in treatments. It can not be said that any adjustment will avoid this problem, that is the nature of clinical trials. However, all these adjustments have the potential to provide a less inflated estimate of the treatment effect and improve the statistical accuracy of the sample size calculation for future trials, therefore reducing long-term cost and unnecessary further trials.

8. Application to Real Data

8.1 Introduction

The adjustment developed in chapter 7 has been shown to work well for simulated data. This simulated data are data which occurs under ideal conditions, yet data collected in real-life trials is rarely from ideal conditions. In order to fully assess the adjustments from the previous chapter, hereafter called the Rothwell adjustment, its impact on real data should be investigated.

8.1.1 Chapter Aims

This chapter builds on the results from chapter 7 and applies the Rothwell adjustment to a real data set. The structure of the data set is described, along with the methods implemented to put the data in the same contexts as those used in chapters 6 and 7. These are the trials in sequence scenario and the pilot study to main trial context.

8.2 Data

This section briefly describes the data set being used in this chapter. The full data set is described in a dissertation by Ho Ching-Ping (Ching-Ping, 2016), therefore a summary is presented here to aid understanding.

8.2.1 Description of Data Set

The data set being used is from Unilever, a company which develops a number of products including antiperspirants. This data comes from a series of antiperspirant trials at Unilever's Research and Development department in Leeds. The data has been provided to the University for a Masters dissertation and used in this thesis to demonstrate the implementation of the Rothwell adjustment on real data.

There are two different objectives for the trials, one being a test product versus a placebo and the second being a head to head comparison with the test product being compared with a second antiperspirant currently on the market. There are 262 trials in the data set, of which 94 are placebo-controlled and 168 are head-to-head comparisons. All the trials are crossover trials, with each product being investigated randomly assigned to a side of the participants' body. All the trials are designed with 80% power and 5% significance level.

In order for a product to be classed as an antiperspirant, it must show some capacity to reduce underarm sweat production. This is tested using a Hot Room, which puts participants in a room with increased temperature and humidity.

The primary outcome of these trials is sweat reduction percentage, which is skewed data. Originally, during the analysis of these studies, the data were log-transformed. However, for the purpose of this chapter the data are kept on the original scale. This is partly to enable clear interpretation of results and ensure the focus of the chapter is the testing of the Rothwell adjustment, not the results of the trials themselves. In order for a product to be deemed "effective", it must demonstrate the log mean sweat reduction being at least 15%, therefore the overall target effect size is -0.15 for a product to be deemed successful.

The results from the trials are expressed as a percentage of the sweat weight reduction of one treatment relative to another. This is calculated as

$$SWR = mean(I) - mean(C)$$

where SWR is sweat weight reduction percentage, I is the investigational product and C is the control product.

8.2.2 Structure of Data

In the full data set there were 263 trials, however, one was removed due to all sweat records being missing, so there were 262 studies in the initial data set. The total number of participants was 8979. Table 8.1 shows the summary statistics for the sample sizes in the whole data set.

Sample Size			
Mean	**Median**	**Minimum**	**Maximum**
34	34	26	43

Table 8.1: Summary statistics for sample sizes in full data set of 262 studies.

In the next section, a subset of the full data set is described. This subset was the data used in this chapter.

8.3 Methods

The data which is used in this analysis is a subset of the data set described thus far. The method used to determine the subset is described in this section, as well as further details of the selection of trials and implementation of the adjustments developed in chapter 7 and shown in Table 8.2.

Effect Size = 0.2	
Power	**Rothwell Adjustment (x)**
80	0.89
85	0.92
90	0.94
95	0.97
99	0.99

Table 8.2: The Rothwell adjustments for trials in sequence for various powers as shown in chapter 7. **Note: x is the value by which the observed difference $\mathbf{d_{T1}}$ should be multiplied.**

8.3.1 Selection of Subset

As described in earlier sections, the full data set contained 262 trials. Each trial had two paired arms, with each pair being the left arm or the right arm. The pairs were randomly allocated either a placebo control or active control, and the active treatment. There were 94 trials which were placebo-controlled and the remaining 168 were head to head comparisons.

Trial subjects were removed if there were no sweat records available for either arm (56 subjects), and if the control axilla sweat weight was less than 100mg, though this did not occur in this data set.

To enable direct comparison to the context of trials in sequence, a subset of this large data set was extracted to consist of only trials where comparisons had been made on the same products. For example, if there were two different trials both comparing a placebo and a particular treatment, these were included.

Doing this allowed the trials to be treated as if they were in sequence, mimicking the context used in previous chapters. Of the 262 trials in the full data set, 135 trials compared the same products, 40 were placebo-controlled and the remaining 95 were head-to-head comparisons.

There were 19 occurences of trials which compared the same products more than twice, 15 of which had three comparisons of the same products and 4 which had four comparisons.

Set	Trial ID	Product 1	Product 2	T1 or T2?	Included?
1	A	Placebo	R1	1	Yes
1	B	Placebo	R1	1	No
1	C	Placebo	R1	2	Yes
2	D	S1	R1	2	Yes
2	E	S1	R1	1	Yes
2	F	S1	R1	1	Yes
2	G	S1	R1	2	Yes

Table 8.3: An illustrative example of the trial matching. For the Placebo vs. R1 trial, T1 was randomly selected from trials A and B.

8.3.1.1 Trial Matching

Since the aim was to use the trials as if they were sequential, one trial had to be set as Trial 1 whilst the other was allocated as Trial 2. For the occurrences where there were 3 trials comparing the same products, the trials were randomly allocated either the number 1 or 2. The trials which compared the same products and were allocated the same trial order number were then randomly selected such that there were only two trials comparing each set of products.

In the cases where there were 4 trials comparing the same products, these were also randomly allocated either 1 or 2 but all were included because each trial could be matched to a second trial.

Table 8.3 provides an illustrative example of this allocation process, so each Trial ID is an individual comparison with each product displayed. The trials were separated into Sets, which were defined as trials comparing the same products. Thus, set 1 contains trials A, B and C all of which compared a placebo with an active product (R1). Set 2 consisted of trials D-G which compared active control S1 to active product R1. Each trial was randomly allocated to T1 or T2, with the condition that at least one trial in each set is allocated T1 and at least one to T2. The final stage of selection for the subsetted data was to randomly select pairs of T1 and T2 from each set. If there were an odd number of trials in a specific set, one trial was randomly excluded such that there were matched trials in T1 and T2.

A total of 60 trial pairs were included in the subset for analysis, 17 placebo-controlled pairs of trials and 43 head to head pairs of trials. There were 15 trials which were not selected due to the random trial number selection process described earlier. There were 8090 observations in total, 4045 for each axilla. These are from 120 distinct trials, 34 placebo-controlled trials and 86 head to head comparisons.

8.3.1.2 Structure of Subset

The subsetted data set was used as the primary data set for the rest of this chapter. The summary statistics for the work described in this chapter are in Tables 8.4 and 8.5. These trials are to be analysed in the context of trials in sequence (T1 to T2). Since the sample sizes are already relatively small, it was decided to not use this data for the pilot to main trial context.

	Sample Size			
	Mean	**Median**	**Minimum**	**Maximum**
Placebo-Controlled	33	33	27	41
Head-to-Head	33	34	26	41

Table 8.4: Summary statistics for sample sizes in the subsetted data.

	Placebo-Controlled	**Head-to-Head**	**Total**
Frequency	34	86	120

Table 8.5: The frequency of placebo and head-to-head trials in the subsetted data.

8.3.2 Analysis

The purpose of this chapter was to assess how robust the Rothwell adjustment is when applied to real data. This was evaluated by applying the Rothwell adjustment to the observed mean difference of statistically significant trials. Up to this point, it has been discussed how the data was formatted to emulate trials in sequence. The data set used from this point forward was the subsetted data of the paired trials, as described in section 8.3.1.1. These trials were completed within-person, so each person had one product on the left axilla and another on the right. The statistical tests used to compare the two treatments was a paired *t*-test, since each person acted as their own control and the trials were all crossover trials. The trials were powered at 80% with a 5% significance level.

The paired *t*-tests were performed on the data, with the non-signficant Trial 1 studies being excluded and not proceeding to Trial 2. This mirrors the pathway used in the simulations in chapter 6.

The target mean difference is calculated by rearranging the sample size equation for a crossover trial, as shown in chapter 2. This equation can be rearranged to get

$$\delta = \sqrt{\frac{2 \times \left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2}{n}}, \tag{8.1}$$

where δ is the standardised target difference and n is the sample size.

The average standardised target effect size is 0.68 for both the placebo and active controlled trials, which is a large effect size as defined by Cohen (Cohen, 1988). Even though the target effect size is large, the adjustments have been demonstrated to not be affected by the size of the target effect size, only the power. Therefore, the adjustment values in Table 8.2 are suitable. Chapter 6 demonstrated that the different effect sizes made little impact on the level of adjustment required. Since these trials are crossover trials, whilst the target effect size is large it is known that the standard deviation will be lower than those seen in parallel group trials.

The second stage of this analysis was to repeat the t-tests on the pairs of trials and then apply the Rothwell adjustment to the observed mean difference from each significant T1. This provides the adjusted mean differences for significant T1 and these can be used to estimate the required T2 sample size, allowing direct comparison to the true T2 sample size. It will also allow comparison of the adjusted T1 mean difference with the observed T2 mean difference, to assess whether the Rothwell adjustment performs as well on real data compared to simulated data. The simulated data showed that the Rothwell adjustment provided a mean difference closer to that of the 'true' mean difference, as opposed to the inflated mean difference observed in trial 1.

These results are presented separately for the placebo controlled trials and the head to head comparisons, since it would be expected that the observed difference is greater for the placebo controlled trials. The results are then assessed by the ratio $\frac{Active}{Control}$. If this ratio is greater than 1, the control treatment reduces the sweat rate more than the active treatment. This ratio is scale-independent and thus allows direct comparison with the results from chapters 6 and 7.

8.4 Results

The results for the placebo controlled trials and head to head comparisons are presented in this section. Within each of these categories, the unadjusted results are presented as well as the results after application of the Rothwell adjustment. These trials are only being used to investigate the trials in sequence context, not the pilot study to main trial context.

Consider the full data set, before any subsetting. If paired t-tests are performed on each trial, 70 out of 120 trials would be statistically significant, such that $P < 0.05$.

The data described in section 8.3.1 is the subsetted data to be used in this analysis, and the results will be presented separately for placebo comparisons and head-to-head comparisons.

	Average Mean Diff			All Results		Sig Only	
Comparison	T1	T1*	T2	Ratio	Inverse	Ratio*	Inverse*
Placebo Controlled	0.899	0.899	0.871	1.03	0.97	1.03	0.97
Head-to-Head	0.118	0.510	0.279	0.68	1.48	1.83	0.55

Table 8.6: The average mean differences for T1 and T2 and the statistically significant (*) T1 trials, and the ratio of both T1 values over T2.

8.4.1 Placebo Controlled Trials

For the placebo controlled trials, it would be expected that the difference between the active treatment and control treatment would be large. All 34 placebo controlled trials are statistically significant when tested individually. The average mean differences are calculated using the absolute values. The overall average mean difference for these trials is 0.885, whilst the average mean difference for the trials allocated T1 is 0.899 and trials allocated T2 is 0.871, as shown in Table 8.6. This table also shows the ratio of T1 over T2, for both the placebo controlled trials and the head-to-head comparisons. The placebo controlled results show a higher ratio of difference between T1 and T2 average observed mean differences, which could be expected due to the nature of placebo-controlled trials.

The results of each trial set are shown in Table 8.7. In this case no adjustment has been implemented on the results of the trials. All the trials are statistically significant, the trials where the mean difference is positive show that the active treatment is less effective at reducing sweat than the control.

The second set of results (Table 8.8) show the observed treatment effect and the adjusted treatment effect in T1, as well as the implications this has on the sample size for T2. Recall that the Rothwell adjustment, as shown in Table 8.2, in this case is to multiply the observed difference by 0.89 due to the trials being 80% powered. Since all the trials in the placebo controlled subset are statistically significant, the adjustment is applied to all T1 trials.

As the studies are powered to 80% and there are 17 trials, even if d was true one would expect 20% of the trials to not be statistically significant. Since the studies are powered at 80%, it would be expected for there to be approximately 4 studies which did not reach statistical significance due to the Type II error. This does not occur with the placebo-controlled trials, all studies were statistically significant. This results in the implementation of the Rothwell adjustment potentially causing an over-adjustment, it is inducing an adjustment based on 80% power when in actual fact the studies are behaving as though they are powered at over 90%, where it would be expected to observe approximately 2 non-significant studies. The standardised target effect size, as calculated from the sample size calculation, is around 0.68 yet the average standardised observed effect size is 6.11 for T1 and 4.95 for T2.

Product	**Trial 1**		**Trial 2**	
Comparison	**Mean Diff (SD)**	**P-value**	**Mean Diff (SD)**	**P-value**
1	0.91 (0.14)	< 0.001	0.94 (0.20)	< 0.001
2	0.91 (0.25)	< 0.001	0.87 (0.18)	< 0.001
3	0.86 (0.13)	< 0.001	0.88 (0.14)	< 0.001
4	0.97 (0.24)	< 0.001	1.38 (0.20)	< 0.001
5	0.84 (0.15)	< 0.001	0.82 (0.20)	< 0.001
6	1.34 (0.25)	< 0.001	0.82 (0.19)	< 0.001
7	0.81 (0.12)	< 0.001	0.82 (0.13)	< 0.001
8	1.08 (0.14)	< 0.001	0.84 (0.17)	< 0.001
9	1.02 (0.16)	< 0.001	0.64 (0.20)	< 0.001
10	0.94 (0.16)	< 0.001	0.96 (0.28)	< 0.001
11	-0.89 (0.10)	< 0.001	-0.81 (0.12)	< 0.001
12	-0.76 (0.13)	< 0.001	-0.84 (0.37)	< 0.001
13	-0.83 (0.09)	< 0.001	-0.85 (0.10)	< 0.001
14	-0.78 (0.20)	< 0.001	-0.99 (0.25)	< 0.001
15	-0.89 (0.27)	< 0.001	-0.80 (0.29)	< 0.001
16	-0.74 (0.10)	< 0.001	-0.70 (0.14)	< 0.001
17	-0.71 (0.10)	< 0.001	-0.85 (0.14)	< 0.001

Table 8.7: The results from the initial paired t-tests for placebo controlled trials in sequence. **Performed under the assumption of no bias.**

Therefore the adjustment is potentially likely to be too conservative for these trials since they are observing a difference much higher than that which they are powered for.

Table 8.9 shows the average sample size used in T1 and T2, compared with the average sample size for T2 if the adjusted mean differences had been used to calculate it. The adjusted sample size for T2 is lower than that which was used (20 vs. 34). This table also shows the average ratio of T1 mean over T2 mean separately for the unadjusted T1 mean and the adjusted T1 mean. As discussed earlier in this section, the adjustment is shown to be too conservative for this data as the trials are powered to a target standardised effect size of 0.68 but are observing standardised effect sizes greater than this. It can be observed that the average ratio when using the unadjusted T1 is higher than the ratio which occurs when the Rothwell adjustment has been used on the T1 mean differences.

Product	**Trial 1 - Mean Diff**		**Trial 2**	**Ratio**	
Comparison	**Unadj**	**Adj**	**Mean Diff**	**Unadj T1/T2**	**Adj T1/T2**
1	0.91	0.81	0.94	0.97	0.86
2	0.91	0.81	0.87	1.04	0.93
3	0.86	0.77	0.88	0.98	0.88
4	0.97	0.86	1.38	0.71	0.63
5	0.84	0.75	0.82	1.02	0.91
6	1.34	1.19	0.82	1.62	1.44
7	0.81	0.72	0.82	1.00	0.89
8	1.08	0.96	0.84	1.29	1.15
9	1.02	0.91	0.64	1.59	1.41
10	0.94	0.83	0.96	0.97	0.87
11	−0.89	−0.79	−0.81	1.09	0.97
12	−0.76	−0.68	−0.84	0.91	0.81
13	−0.83	−0.74	−0.85	0.98	0.87
14	−0.78	−0.69	−0.99	0.79	0.70
15	−0.89	−0.79	−0.80	1.12	0.99
16	−0.74	−0.66	−0.70	1.06	0.94
17	−0.71	−0.63	−0.85	0.83	0.74

Table 8.8: The results from the paired *t*-tests for placebo controlled trials in sequence, **having adjusted for bias in Trial 1**.

	Average Sample Size			**Average Ratio T1/T2**	
Comparison	T1	T2 - Adj**	T2 - Unadj	Unadj	Adj
Placebo-controlled	33	20	34	1.06	0.94

Table 8.9: The average sample size per arm for T2 (using the unadjusted and adjusted mean difference from T1) and average ratio of T1 over T2. ****SS calculated using significant T1s only**.

8.4.2 Head to Head Comparison Trials

There are 36 out of 86 trials which are statistically significant (42%). The average difference in sweat production for the trials allocated Trial 1 is 0.19, which is considerably smaller than the observed average mean for the placebo controlled trials. This average indicates that the active products do not reduce sweat production considerably when compared with active controls. The average observed effect for the significant T1 trials is 0.510, whereas the average observed effect size for all the T2 trials was 0.279, which again shows the bias which occurs when moving from T1 to T2.

The average standardised observed effect size for all T1 trials is 0.92 and for the T2 trials is 1.76. For the statistically significant trials in T1 the average standardised observed effect size is 2.42 and for the T2 trials which followed these significant results is 1.12, again demonstrating a regression to the mean effect.

As shown in Table 8.6, it can be observed that the average for the statistically significant results in T1 is higher than the average for T2. The ratio shows that the statistically significant results of T1 give an inflated mean difference of around 83%, which is considerably higher than the bias demonstrated in chapter 6 and the mathematical solution of chapter 7. The inverse of this ratio is 0.55, therefore the average difference in the results from trial 1 and trial 2 is more than the expected bias of 0.89.

Table 8.10 shows the results for the 43 pairs of trials. If T1 gives a non-significant result, then that pair is stopped. Out of the 43 trials performed as T1, there were 15 (35%) which were statistically significant, therefore these products would be progressed to the next stage, T2. Of these 15 trials at the second stage, only 7 (47%) were statistically significant, which corresponds to 16.2% of the total number of trials performing a head to head comparison.

It can be observed that the adjusted T1 values are identical to 2 decimal places to the unadjusted T1 values, when close to 0. However, these values would typically not result in needing the adjustment as they would not have a significant P-value, therefore the adjustment would not be implemented. It is also noticable in Table 8.10 that though the comparisons are of the same products for each row, the mean differences are very different for T1 and T2. Some of the trial pairs switch from showing the active product is superior to showing the placebo is superior, in particular trial pairs 7 and 24 which show a swapping of result and both results reaching statistical significance.

Table 8.11 compares the unadjusted mean differences for T1 with the Rothwell-adjusted mean differences from T1 and the observed differences for the progressed T2 results. Trial 2 only occurs if Trial 1 was statistically significant, as shown in

Product Comparison	**Trial 1**		**Trial 2**	
	Mean Diff (SD)	**P-value**	**Mean Diff (SD)**	**P-value**
1	0.08 (0.19)	0.025	0.03 (0.13)	0.200
2	0.07 (0.25)	0.129		
3	−0.001 (0.03)	0.780		
4	−0.94 (0.17)	< 0.001	0.02 (0.44)	0.804
5	0.02 (0.25)	0.629		
6	−0.05 (0.24)	0.241		
7	1.00 (0.28)	< 0.001	−0.10 (0.18)	0.001
8	−0.03 (0.21)	0.496		
9	−0.11 (0.26)	0.023	0.07 (0.35)	0.265
10	0.01 (0.13)	0.527		
11	−0.02 (0.20)	0.626		
12	1.78 (0.26)	< 0.001	−0.05 (0.18)	0.103
13	0.95 (0.14)	< 0.001	−0.06 (0.17)	0.057
14	−0.08 (0.33)	0.129		
15	−0.11 (0.21)	0.007	0.91 (0.20)	< 0.001
16	0.01 (0.20)	0.825		
17	0.65 (0.57)	< 0.001	0.20 (0.15)	< 0.001
18	−0.001 (0.21)	0.968		
19	0.07 (0.22)	0.051		
20	0.01 (0.17)	0.757		
21	−0.01 (0.17)	0.841		
22	−0.01 (0.14)	0.751		
23	0.04 (0.14)	0.179		
24	−0.86 (0.24)	< 0.001	0.96 (0.39)	< 0.001
25	0.11 (0.23)	0.009	−0.17 (0.25)	< 0.001
26	−0.22 (0.28)	< 0.001	0.97 (0.19)	< 0.001
27	0.82 (0.18)	< 0.001	−0.07 (0.24)	0.150
28	0.01 (0.02)	0.003	−0.003 (0.03)	0.364
29	−0.002 (0.03)	0.694		
30	−0.01 (0.02)	0.012	0.01 (0.02)	0.008
31	< 0.001 (0.03)	0.848		
32	−0.001 (0.03)	0.846		
33	< 0.001 (0.02)	0.812		
34	0.004 (0.03)	0.308		
35	−0.002 (0.03)	0.700		
36	< 0.001 (0.03)	0.882		
37	−0.002 (0.01)	0.311		
38	0.01 (0.03)	0.013	0.001 (0.02)	0.689
39	0.01 (0.02)	0.129		
40	0.001 (0.02)	0.760		
41	−0.004 (0.02)	0.227		
42	0.003 (0.02)	0.405		
43	0.003 (0.02)	0.399		

Table 8.10: The results from the paired t-tests for active-controlled trials in sequence.

Table 8.10. Some of the results appear to go from being statistically significant in favour of the active treatment to statistically significant in favour of the control treatment (See Comparison 24 in Table 8.11). Recall that if the mean difference is negative, that indicates that the active treatment caused less sweat production than the control treatment.

Table 8.12 shows the sample sizes used for T1 and T2, compared with the calculated sample size for T2 based on the significant results from T1. It also presents the average ratio of T1 mean difference over the T2 mean for each the unadjusted and adjusted T1 mean differences. This only occurs for the significant T1 results. The results show that due to the variance in the magnitude of the observed differences for T1, the new adjusted average sample size for T2 is 495, which is much larger than the average of 34 participants for T2. However, this ranges from 13 per arm to over 2500 per arm.

The products that the trials were matched upon were identical, and all coding was thoroughly assessed for errors throughout this research. However, the protocols for each trial was not available, therefore whilst every effort was made to match the trials appropriately, it could be that there are other factors which makes the matched trials not identical. For example, the two trials could be testing the same product but under different conditions. This is further discussed in section 8.5.

Product	Trial 1 - Mean Diff		Trial 2	Ratio	
Comparison	Unadj	Adj	Mean Diff	Unadj	Adj
1	0.08	0.07	0.03	2.99	2.66
2	0.07	0.06			
3	−0.001	−0.001			
4	−0.94	−0.84	0.02	−43.84	−39.02
5	0.02	0.02			
6	−0.05	−0.04			
7	1.00	0.89	−0.10	−9.56	−8.51
8	−0.03	−0.02			
9	−0.11	−0.09	0.07	−1.52	−1.36
10	0.01	0.01			
11	−0.02	−0.01			
12	1.78	1.58	−0.05	−34.57	−30.77
13	0.95	0.85	−0.06	−16.04	−14.27
14	−0.08	−0.07			
15	−0.11	−0.10	0.91	−0.12	−0.11
16	0.01	0.01			
17	0.65	0.58	0.20	3.17	2.82
18	−0.001	−0.001			
19	0.07	0.06			
20	0.01	0.01			
21	−0.01	−0.01			
22	−0.01	−0.01			
23	0.04	0.03			
24	−0.86	−0.76	0.96	−0.89	−0.79
25	0.11	0.09	−0.17	−0.62	−0.55
26	−0.22	−0.20	0.97	−0.23	−0.20
27	0.82	0.73	−0.07	−12.15	−10.82
28	0.01	0.01	−0.003	−3.59	−3.19
29	−0.002	−0.002			
30	−0.01	−0.008	0.01	−0.85	−0.76
31	< 0.001	< 0.001			
32	−0.001	< −0.001			
33	< 0.001	< 0.001			
34	0.004	0.004			
35	−0.002	−0.002			
36	< 0.001	< 0.001			
37	−0.002	−0.002			
38	0.01	0.01	0.001	11.17	9.94
39	0.01	0.005			
40	0.001	0.001			
41	−0.004	−0.004			
42	0.003	0.003			
43	0.003	0.003			

Table 8.11: The results from the paired *t*-tests for active-controlled trials in sequence, **having adjusted for bias in Trial 1**.

	Average Sample Size			**Average Ratio T1/T2**	
Comparison	T1	T2 - Adj**	T2 - Unadj	Unadj	Adj
Head-to-Head	33	495	34	-7.11	-6.33

Table 8.12: The average sample size per arm for T2 (using the unadjusted and adjusted mean difference from T1) and average ratio of T1 over T2. ****SS calculated using significant T1s only**.

8.5 Discussion

This data set has been used to illustrate the practical application of the adjustment. Whilst the adjustment functioned well for the simulated data in chapter 7, it was more difficult to manage when implemented on the real data set used in this chapter. There are a number of observations which will be discussed here.

For the placebo-controlled trials, all the trials were statistically significant when analyzed individually, which was not unexpected. This is due to the target effect size for a placebo controlled trial to be larger than that expected in a head-to-head comparison. Simply, placebo-controlled trials are aiming to show efficacy, whereas a head-to-head comparison aims to conventionally build upon efficacy and show superiority compared to another similar product.

If the T1 mean difference is adjusted using the Rothwell adjustment and this adjusted mean difference is used in the sample size calculation for T2, it has been shown that the sample sizes required for T2 are lower than those used. The limitations of this application are discussed in the next section.

The adjustment was only implemented on the real data under the context of trials in sequence. It would be worth doing further work in the context of pilot study to main trial, where the adjustment is tested on a real data set to assess its accuracy. The studies were powered at 80% based on an average standardised target effect size of 0.68.

The average absolute observed effect size for T1 in the placebo-controlled trials was 0.899, compared to 0.871 in the T2 trials. Since all the T1 trials reached statistical significance and the observed standardised effect size of 6.11 is much greater than the target standardised effect size of 0.68, this could indicate that the placebo-controlled trials were over-powered. For the head-to-head comparisons, the target was also 0.68, but the observed for the statistically significant T1 trials was 0.510. This was not unexpected since the difference for active-controlled trials is usually smaller than placebo-controlled trials. The observed effect size for the T2 trials which would have occurred (those which had a successful T1 trial) was 0.242, considerably lower than the T1 result. Again this shows that regression to the mean is occuring. However, the problems with the sign-swapping in the head-to-head results leads me to conclude that this data was perhaps not the best example to test the Rothwell adjustment.

It can be concluded that whilst this exercise was worth doing to demonstrate the adjustment, it did not prove to be as helpful as initially expected. The observation that in some of the head-to-head comparisons the sign of the observed mean difference swaps from negative to positive (or positive to negative) yet both results are reaching statistical significance indicates that these trials are possibly not as similar as originally thought. This is discussed further in the limitations section.

Another possible explanation for the sign-swapping observed in the head to head comparisons is that this is real life data and thus these things can just occur due to real life variation. The sample sizes of these trials are relatively small, with the average achieved sample size being around 33 participants. Each participant acts as their own control due to the crossover design (as discussed in chapter 2), resulting in a smaller variance and therefore a smaller sample size.

The data used in this chapter are based on trials with 80% power. The trials were powered on a target standardised effect size of 0.68, yet for the placebo-controlled trials the average observed standardised effect was 6.11. None of the placebo-controlled trials had $P \geq 0.05$. In fact, the probability of all 17 placebo-controlled trials being statistically significant is $0.80^{17} = 0.023$, therefore the probability of seeing at least one trial with $P \geq 0.05$ is 0.977. It would be expected under 80% power that 3 trials would be not significant. This gives evidence to support the conjecture that the trials have been overpowered.

8.5.1 Limitations

There are a number of limitations which must be considered when discussing the work in this chapter. Each will be discussed, along with potential implications and considerations.

The data used in this chapter is based on a series of crossover trials performed on a variety of products. Whilst the Rothwell adjustment should, in theory, be transferable to various designs of trials, it was designed based upon parallel group trials. As discussed in chapter 2, the theory of sample size calculations for parallel group trials is similar to that for crossover trials, and can be extended to non-inferiority and equivalence designs. However, since this adjustment has been developed under the parallel group setting, it can not automatically be assumed to function as well under a different design. It would be expected to be robust as the adjustment has been developed based on a truncated Normal distribution, which would still be the distribution for the other trial designs, but further research is needed to explore this further.

Another potential limitation of this real data application is the fact that the trials were not designed as trials in sequence, they are all individual trials. Alongside this fact, they are all trials on a variety of different products. So whilst the trials put in sequence are testing the same products, grouping these trials together and summarising the data should be done with caution as there are some contradictory results. Some trials are showing that the investigational product is superior and some showing that the placebo or control is overwhelmingly superior. This is most noticable for the placebo-controlled trials, though is also applicable to the head to head comparisons.

The individual protocols for the trials could not be accessed, therefore the analysis in this chapter has been completed under the assumption that these trials could be paired up as described in section 8.3.1. Given this, there is no way of determining if the studies which were comparing the same products were performed exactly the same way. They could be on different populations, different genders or under different test conditions. The method of inducing sweat production could vary, for example one study could be performed on stationary participants and another study comparing the same products could be conducted on participants whilst walking. Other ways that the studies could differ even when testing the same products include differing levels of fitness of the participants or different ages of participants. The data provided were the sweat production rates and subject, there is no other demographic or study information. It was decided to not use this data for the pilot to main trial context due to not having enough information about the similarity of the paired trials.

The adjustment was designed under the alternate hypothesis, therefore it is assumed for an 80% powered trial there will be 20% of trials which will not reach statistical significance. However, in the placebo-controlled trials with real-data all the trials were statistically significant. This results in the Rothwell adjustment causing an over-adjustment, because T1 are designed with 80% power and achieve a much higher power than that. As the adjustments depend on the power of the first trial, realistically the appropriate adjustment required for the placebo-controlled T1 trials would be close to 1.00 as opposed to the 0.89 adjustment which was used. The use of 0.89 actually introduces an adjustment when one of this magnitude is not necessarily required. It would be worth considering whether the adjustment should be associated not only with the planned power of T1, but also the achieved power of T1.

Since all the trials are independent, the comparison of the sample sizes for the second trial based on the results from the first should be taken with caution. There is no way to know which of these trials occurred first in real life, it was randomly allocated in this chapter to demonstrate the application of the Rothwell adjustment, and highlight its advantages for trials in sequence.

8.6 Conclusion

This chapter has centred around the application of the Rothwell adjustment to a large set of real data, consisting of multiple trials. The aim was to determine whether it is robust to the nuances of real data. Whilst the Rothwell adjustment performed well against the Kirby and Wang adjustments on simulated data in chapter 7, further work would be useful to apply it to more real data sets to fully assess its validity.

The Rothwell adjustment has been shown to be useful thus far, as it has validated the results of the simulations and been proven mathematically. It would be a useful tool for trialists to consider when designing trials based on previous research. Once the power and approximate target effect size are known, the appropriate adjustment can be applied to the previous data to reduce the possible effect of regression to the mean. In the long term this has the potential to reduce the number of trials which do not achieve statistical significance if there is truly a difference between treatments. It is always possible that the previous work were anomalies, uncharacteristically high differences even if there is no difference in treatments. It can not be said that any adjustment will avoid this problem, that is the nature of clinical trials. However, what all these adjustments can do is provide a less inflated estimate of the treatment effect and improve the statistical accuracy of the sample size calculation for future trials.

The work in chapter 6 and earlier in this chapter were based on simulated data, which is an ideal condition. When the data is from real life, more variation occurs. The adjustment performed well in terms of reducing the observed effect size from T1 and therefore reducing the ratio of bias in the placebo trials, however, for the head-to-head trials the observed effect sizes which reached statistical significance were sometimes small. The data sometimes did not behave as one would expect, such as getting a statistically significant positive result at T1 then getting a statistically significant negative result at T2.

The adjustment did not perform as well with the real data as it had done previously with the simulated data. Whilst this was expected due to the higher amount of variation in real life data, it highlights the need for further work to assess the Rothwell adjustment when applied to real data. The adjustment did not appear to function as well as anticipated, as it did not have an effect for many of the observed mean differences in the head-to-head comparisons due to some having very small observed effect sizes. However, these trials reached statistical significance because the target effect size for these trials was large (0.68) yet the standardised observed effect sizes were larger. Recall that only the statistically significant trials in T1 would be progressed to a confirmatory trial (T2), hence for all the statistically significant T1 trials the adjustment did reduce the observed mean difference, therefore it did have an effect and perform as desired.

9. Discussion and Recommendations

9.1 Introduction

There is a lot of uncertainty about the choice of the target difference in clinical trial design, and a growing trend of seemingly positive primary trials failing to reach statistical significance at the secondary stage. This impacts not only the chances of a drug getting onto the market but also the finances of trial units and pharmaceutical companies.

This thesis aimed to investigate the commonly used methods of choosing the target difference in clinical trial design, as well as establish a range of plausible target effect sizes to assist funding bodies in reviewing grant applications. It also investigated the effect of regression to the mean on trials in sequence, and develop an adjustment to be applied to the observed treatment effect in the first trial or collection of previous research to better estimate the target effect size to be used for the second trial.

This chapter aims to collate the research completed in the process of writing this thesis, summarising the results and discussing the potential implications and practicalities of this work. The format of the chapter will be to recap the research questions, then to discuss the associated research aimed at answering those questions. Finally, an overall summary of the thesis is presented along with limitations of the research as a whole and suggestions for further work.

9.1.1 Research Aims

To recap the research questions presented in this thesis, each question will be discussed separately. The questions were as follows:

1. How are effect sizes quantified in the design of clinical trials?
2. Are the reported target effect sizes similar to *a priori* effect sizes? Is this effect size clinically important?
3. What range of observed effect sizes are being seen in different clinical areas or populations?

4. Are there more optimal methods for quantifying the effect size?
5. Are there more optimal methods to adjust for the bias of moving from one trial to the next?

9.2 Discussion of Research

This section will discuss each of the research questions in detail, as well as link the questions to the appropriate chapters in the thesis and, by extension, the associated work within those chapters. Each of the questions is discussed in terms of the implications of the research conducted, as well as the limitations, which are discussed throughout the chapter. The research conducted solved all the research questions posed at the start of the thesis, to the best of my knowledge.

9.2.1 Quantifying Effect Sizes in the Design of Clinical Trials

The first research question aimed to establish how effect sizes are quantified in clinical trials. This was investigated through the review in chapter 4, which showed that a variety of methods were used to elicit the target effect, however the most common method was the review of evidence, either in isolation or in conjunction with other methods. According to Cook *et al.* (Cook et al., 2014), the review of evidence method is used to "specify an important and/or a realistic difference", as mentioned in chapter 3.

The discussion in chapter 4 about whether there are more optimal methods for quantifying effect sizes is based around the findings of the HTA review and the work completed with the DELTA2 group (Cook et al., 2018).

The choice of the target difference depends on the research aim and the question. Are we looking for an important effect, or a realistic effect, or both? Each of these questions result in a different 'optimal' method for target difference elicitation, the context of the trial and its aims are of vital importance.

Chapter 3 provided a list by Cook *et al.* (Cook et al., 2014), which identified different methods of elicitation under the categories of important difference, realistic difference and an important and/or realistic difference.

A realistic effect would be estimated using previous work; if an effect size, x, has been seen in one trial, it would be reasonable to expect that a similar effect would be observed again, so long as the trial was performed on a comparable population and intervention. The quantification of an important difference depends on the context in which the question is being asked. A clinician may have a very different perspective of an important difference compared to a patient.

As far as I am aware, I am the first person to perform this review and gather data on the target and observed effect sizes. This work has shown that the most commonly used and reported method of target difference elicitation is the review of evidence base, which is the use of previous research to inform estimates for the current trial. The use of multiple methods to establish a more accurate estimate is advised, and the case studies in chapter 4 show examples of good practice. The median target effect size was shown to be 0.30 from the review in chapter 4. This result may be useful for future trial design and funding bodies, as it could be used as a bench-mark for the target effect size.

9.2.2 Comparison of Target Effects with Observed Effects by Clinical Area

Having illustrated the potential implications of over-estimating the target effect size, it prompted the question of how many trials are actually achieving their target or estimated effect size used in the original sample size calculation.

Effect sizes are known to vary across research areas, with areas such as nutrition and genetics consistently reporting extremely small effect sizes (Siontis and Ioannidis, 2011). If it is in fact true that trialists are over-estimating the target effect size in their sample size calculations, it would be useful to know, given a particular disease area or research area, a range of plausible effect sizes. This could be used to assist trialists when designing trials for which there is no previous research or data.

The review detailed in chapter 4 indicated that for publicaly-funded parallel group superiority trials, the median target standardised effect size is 0.30 and the median observed standardised effect size is 0.11. These values would provide trialists and funding bodies with a baseline by which to compare proposed target effect sizes in grant applications. Though there are some large effects being observed, there are also a large proportion which are very small. On average, the target effect size was greater than the observed effect size. As discussed in chapter 4, this is not unexpected.

These results have not been found prior to this, as far as I am aware, and they could prove useful in the design of trials in the future.

9.2.3 Optimal Methods for Quantifying the Target Effect Size

The other important finding from chapter 4 was that the most common method of elicitation of the target effect size is the use of previous research. This was either as the primary method to estimate the target difference or in tandem with other methods. There are some inherent issues with using this approach, the first of which being based around publication bias. If trialists' are using information from previous trials to estimate a target effect size for their trial, the original trial could potentially be a 'random high', therefore an over-estimation of target effect size would occur and the trial could be under-powered to detect a difference even if one truly exists. The second issue with this method is that a regression to the mean effect could be introduced. This is further discussed in chapter 5, and forms the justification for an adjustment to be developed to reduce the observed effect size to account for this.

The case studies presented in chapter 4 provide examples of good practice. The importance of transparency when discussing the estimate for the target effect size is highlighted. It would be more useful for reviewers and funding bodies to see fully where the estimate comes from, in order to establish whether it is realistic. If an adjustment has been provided to down-weight the estimate, or to account for other uncertainties about the parameter estimates in the sample size calculation, this need to be clearly explained. The use of multiple methods of elicitation is also discussed as being optimal, since the more information that the trialists' base the estimate of target effect size on, the more accurate the estimate should be.

The benefit of using the HTA trial reports for the review was that part of the requirements of funding from the HTA is that all trials are reported, irrespective of the statistical significance of the trial. This enhances the results of the review as whilst only one journal was used, which could be seen as a limitation, there is not publication bias in the sense of the journal only publishing statistically significant trials. Also, these trials are considered some of the best in the UK, with trials being published in high-impact journals such as The Lancet, New England Journal of Medicine and the British Medical Journal as well as being published in the HTA journal.

9.2.4 Optimal Methods of Adjusting for Bias

There were four chapters which investigated the question of optimal methods for adjusting the bias which occurs for trials in sequence. Each one built up to the development of the Rothwell adjustment and then assesses the robustness of this adjustment. This section will discuss each chapter and how it lead to the development of the Rothwell adjustment that is proposed, as well as discussing briefly the limitations of the work presented and any potential implications of this research.

9.2.4.1 Regression to the Mean

Chapter 5 presented a systematic review with regression to the mean as the key component. It also detailed the other currently available methods to adjust for this effect for trials in sequence, the first method by Wang *et al.* which was based on the standard error of the first trial (Wang et al., 2006), and the second method by Kirby *et al.* which was a rule-of-thumb adjustment to down-weight the observed effect size by 10% (Kirby et al., 2012). These methods were tested on simulated data and appeared to perform well, though as discussed in the chapter they were generic methods and there had been little discussion in the respective papers on the generalisability of these methods and application to other trial designs.

9.2.4.2 Simulations of Trials in Sequence

The phenomenon of regression to the mean was shown to occur for trials in sequence in chapter 6, demonstrating that for trials with 80% power and any effect size, the approximate over-estimation of the effect size is $12-13\%$ compared to that observed in the second trial. The amount of over-estimation varies depending on the power of the studies, and is not affected by the effect size used in the design of the study. The bias is determined by taking the ratio of the trial 1 result over the trial 2 result, which is scale-independent. Also, since the ratio is the number of interest, it is not influenced by the numerical value of $\mu*$, where $\mu*$ is the observed effect size.

The distribution of trial 1, when only the statistically significant results are progressed to trial 2, is a left-truncated Normal distribution. This follows through logic as well, since if trials are stopped if $P < 0.05$ then for trials which observe an effect smaller than the critical value at which $P < 0.05$, they will not continue so there is a cut-off point on the left of the distribution. The knowledge that this distribution follows that of a left-truncated Normal is used to confirm and validate the development of the Rothwell adjustment, discussed later in the chapter.

9.2.4.3 Development of Adjustment

The research presented in chapter 7 demonstrates the methods which are currently in the literature and compared them with a new method developed by JCR from the simulation results in chapter 6. This adjustment was then proven mathematically in chapter 7 for both trials in sequence and pilot study to main trial contexts, demonstrating that the results from the simulations held for any value of μ. The adjustment varied depending on the power of the initial trial, to the best of my knowledge there is no other research which presents this. This adjustment was different to all the adjustments found in chapter 5 and provides a more stable adjustment (as shown in chapter 6) to be applied to the results of previous trials.

This result means that if a trialist knows the planned power of the first study, they can apply the appropriate adjustment to the observed effect from the first trial to get a more accurate estimate for the effect size. This, in turn, provides a more accurate estimate for the sample size calculation. The application of the adjustment will inevitably increase the sample size required for the second trial, however this needs to be balanced with the more accurate estimate for the target difference and the cost of the trial. There will be cases where the cost of the adjusted trial may be too large, but this needs to be considered with improvements in transparency of the estimated outcome.

The systematic review in chapter 5 demonstrated that this is an important area of research. There were documented investigations of the phenomenon of regression to the mean, and some adjustments provided by Wang *et al.* and Kirby *et al.*, however the adjustment developed in this thesis is a considered refinement.

The adjustment by Wang was shown in chapter 7 to be relatively unstable, functioning well for high powered trials but dependent on the standard error of the initial trial, therefore extremely varied in performance. The Kirby adjustment was a flat adjustment, which initially appeared to perform well, however was shown to over-adjust the initial trial result when the power was not 80%. For low-powered trials, it would follow that the Kirby adjustment would not adjust the results enough, since the Kirby adjustment was constant at 0.90 and the Rothwell adjustment was 0.89 for trials with 80% power but increased (decreased) as the power increased (decreased).

The proposed Rothwell adjustment is based on a truncated Normal distribution and has been validated with simulations. This is the first method, to the best of my knowledge, which takes the power of the initial trial into consideration, as well as the first method to be implemented in the pilot study context.

The adjustments which have been proposed are dependent on the truncation point of the distribution of the first trial. The truncation point is affected by the power of T1 for trials in sequence, and for pilot study to main trial scenarios it is affected by the progression criteria of the pilot study.

9.2.4.4 Application of Rothwell Adjustment

The adjustment was developed in chapters 6 and 7, and tested on the simulated data in chapter 7. For the simulated data, the Rothwell adjustment consistently reduced the bias and resulted in a target effect size for Trial 2 which was closer to the target and 'true' effect size. It proved to be more stable than the Kirby and Wang adjustments.

The next step in the development process was to test the adjustment on a real data set. The problems which usually occur with real data are that it rarely behaves as you hope it will, there is also an issue when there is a large quantity of missing data. Luckily, with the real dataset available there was not much missing data. The adjustment was used on a subset of the data which were matched to emulate trials in sequence. The real data were from a series of crossover trials with a continuous endpoint. Though the adjustment was developed based on a parallel-group design, it can be used for a crossover design since, as shown in chapter 2, the crossover design is simply an extension of the parallel-group design. The adjustment reduced the observed effect size towards the target effect size for placebo-controlled trials in chapter 8.

It would be useful to test the adjustment on a wider variety of real data to assess its robustness to the nuances of real life data. Further work would be recommended in this area.

9.3 Implications of Research

The research conducted and presented in this thesis demonstrated the occurance of regression to the mean for trials in sequence. This will have an impact on all trials which are sequential, even if they are not planned as such.

From the context discussed in this thesis of trials in sequence and pilot study to main trial, the context of trials in sequence can be broadened to include interim analyses, systematic review results being used to design a new trial, meta-analyses results or the meta-analysis itself. A limitation of the research is that an assumption of identical end-points has been made. It could occur that the end-point used in the first trial or pilot study is a surrogate, with the primary end-point for the second trial being a secondary end-point in the initial trial. Further work could extend to apply adjustments for secondary endpoints.

In terms of the meta-analysis itself, a meta-analysis would not occur unless some of the results of the trials to be included were providing positive indications.

Choosing a smaller target effect size could cause a trial to become infeasible. For example, if a smaller, more realistic, target effect size results in a sample size which is too large and therefore causes the trial to be too costly, one solution for the trialists' is to choose another endpoint which is important. This alternative endpoint could result in a smaller sample size and therefore make the trial affordable. However, consideration must be given to the association between the original and alternate endpoints to ensure the trial aims are still met. A surrogate endpoint could result in a lower required sample size as well. Another possible solution to a trial being financially infeasible is to consider a multi-site trial, or to collaborate with other researchers if there is no appropriate alternative endpoint to increase potential funding.

If trialists are aiming too high with their target effect size, they could have a reduced sample size for the trial but this can result in an under-powered trial. If a trial is underpowered, it is less likely to observe an effect when one truly exists, thus if there is truly a difference between treatments the chance of the trial observing this is reduced. Consideration should be given to the cost of a potential increase in initial target sample size versus the cost of extending the trial or having an underpowered trial. The implications of having an underpowered trial are fairly serious; if the drug truly works and would be beneficial to patients but can not reach statistical significance due to being underpowered, another trial would have to possibly be performed to confirm efficacy or the drug would never reach the market. A second trial would be inevitably more costly than correctly increasing the sample size of the initial trial, and the implications of the drug never reaching the market are that patients do not benefit.

It would be useful to extend the work presented in this chapter to the context where one study is powered on a surrogate endpoint and the second study is powered on a different primary endpoint, which could have been a secondary endpoint in trial 1. If the power for trial 1 is high, the bias should be minimized as shown by the Rothwell adjustment not being as strict as the power increases.

9.4 Overall Summary

This thesis has provided evidence of a regression to the mean effect when using the results of previous research to design future trials. It has used this information to design an adjustment to apply to the results of the previous research to reduce the potential bias resulting from regression to the mean.

Currently, in UK publicly funded trials the median target effect size is 0.30, which corresponds to a small effect under the Cohen categorisation (Cohen, 1988). This provides evidence for trialists and funding bodies that trial teams claiming to target a large effect size (above 0.5 on the standardised scale) could be being overly optimistic as this large an effect is rarely seen. The fact that this research has shown an average effect size of 0.30 provides a good starting point for trialists when discussing the design of a trial.

Advice for funding bodies based on this research would be that trialists' presenting grant applications with target effect sizes (when standardised) greater than 0.3 are being optimistic, as compared to those extracted from chapter 4 the average target standardised effect size is 0.3 and the average observed standardised effect size is 0.11, which is much lower and provides possible evidence as to why 66% of trials are not reaching statistical signficance. There is the consideration that the treatment truly does not work, which needs to be included in any evaluation of statistical insignificance before concluding that the target effect size was overly optimistic.

The Rothwell adjustment has proven to be mathematically sound and has worked well under simulated conditions. Further testing of the adjustment is required when working with real data of different types, of particular interest would be its functionality with results from systematic reviews and meta-analyses, as well as other more complex trial designs.

Systematic reviews were not considered as part of this research, however logically an adjustment should be considered when using the results of those to design a new, confirmatory trial. Whilst one would hope that the effect of regression to the mean be reduced by the systematic review, since these reviews aim to collate all existing evidence to provide an overall estimate of treatment effect, one can not rule out the effect of publication bias on this. If the systematic review is based on only published trial results, there could be publication bias introduced, which compounded with regression to the mean could result in a very high target effect size.

9.4.1 Further Work

The potential for further work stemming from this research is quite large. It is still a current area of interest and this thesis documents only a small portion of the various trial designs and primary outcome measures. This is strictly a frequentist approach, as stated in chapter 1.

This research could be developed to different trial designs, including but not limited to non-inferiority trials, equivalence trials and adaptive designs. It could be particularly useful for adaptive designs, and could be evaluated at an interim analysis point to determine whether the adjustment is providing a more accurate estimate of the effect size. Another aspect which was mentioned earlier is when the first trial is powered on another endpoint and the secondary endpoint in trial 1 is the primary endpoint in trial 2.

Non-inferiority and equivalence trials are not designed in terms of a target difference, they are based on a non-inferiority or equivalence margin, as described in chapter 2. However, the same issue with regression to the mean could occur in this context as well. The non-inferiority/equivalence margin comes from what is deemed clinically meaningful, which is similar to the target difference.

The work presented in this thesis is primarily focused on continuous end-points, so there is an avenue to develop the adjustment for binary end points as well. Survival endpoints are slightly different and their associated trials are complex, therefore it could be that a different adjustment is required for these trials and endpoint.

The Rothwell adjustment has been demonstrated on a real data set, however it would be useful to confirm its efficacy and robustness through further analysis of various trial designs and real data which is from trials in sequence, such as Phase II to Phase III trials. If the adjustment performs well in this context, it could be further developed to the context of meta-analyses and systematic reviews.

Bibliography

Agostino, R. B., Massaro, J. M., and Sullivan, L. M. (2003). Noninferiority trials: design concepts and issues the encounters of academic consultants in statistics. *Statistics in Medicine*, 22(2):169–186.

Allison, D., Elobeid, M. A., Cope, M. B., Brock, D., Faith, M. S., Veur, S. V., Berkowitz, R., Cutter, G., McVie, T., Gadde, K., and Foster, G. (2010). Sample size in obesity trials: Patient perspective versus current practice. *Med. Decis. Mak.*, 30(1):68–75.

Allison, D. B., Loebel, A. D., Lombardo, I., Romano, S. J., and Siu, C. O. (2009). Understanding the relationship between baseline bmi and subsequent weight change in antipsychotic trials: effect modification or regression to the mean? *Psychiatry Res*, 170(2-3):172–6.

Altman, D. G. (1980). Medicine and mathematics - statistics and ethics in medical-research .3. how large a sample. *British Medical Journal*, 281(6251):1336–1338.

Altman, D. G. (1999). *Practical statistics for medical research.* London ; Boca Raton : Chapman and Hall/CRC, 1999, London ; Boca Raton.

Ambrosius, W. T. and Mahnken, J. D. (2010). Power for studies with random group sizes. *Statistics in Medicine*, 29(10):1137–1144.

Andrews, P., Avenell, A., Noble, D., Campbell, M. K., Croal, B., Simpson, W., Vale, L., Battison, C., Jenkinson, D., and Cook, J. (2011). Randomised trial of glutamine, selenium, or both, to supplement parenteral nutrition for critically ill patients. *Bmj*, 342:d1542.

Asmar, R., Safar, M., and Queneau, P. (2001). Evaluation of the placebo effect and reproducibility of blood pressure measurement in hypertension. *Am J Hypertens*, 14(6 Pt 1):546–52.

Bajard, A., Chabaud, S., Perol, D., Boissel, J. P., and Nony, P. (2009). Revisiting the level of evidence in randomized controlled clinical trials: A simulation approach. *Contemp Clin Trials*, 30(5):400–10.

Barbui, C., Violante, A., and Garattini, S. (2000). Does placebo help establish equivalence in trials of new antidepressants? *European Psychiatry*, 15(4):268–273.

Barnett, A. G., van Der Pols, J., and Dobson, A. (2005). Regression to the mean: what it is and how to deal with it. *Int. J. Epidemiol.*, 34(1):215–220.

Barrett, B., Brown, D., Mundt, M., and Brown, R. (2005a). Sufficiently important difference: expanding the framework of clinical significance. *Medical decision making : an international journal of the Society for Medical Decision Making*, 25(3):250.

Barrett, B., Brown, R., Mundt, M., Dye, L., Alt, J., Safdar, N., and Maberry, R. (2005b). Using benefit harm tradeoffs to estimate sufficiently important difference: The case of the common cold. *Medical Decision Making*, 25(1):47–55.

Barrett, B., Harahan, B., Brown, D., Zhang, Z., and Brown, R. (2007). Sufficiently important difference for common cold: severity reduction. *Ann Fam Med*, 5(3):216–23.

Beaton, D. E., Boers, M., and Wells, G. A. (2002). Many faces of the minimal clinically important difference (mcid): a literature review and directions for future research. *Current opinion in rheumatology*, 14(2):109.

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–302.

Bernstein, E., Edwards, E., Dorfman, D., Heeren, T., Bliss, C., and Bernstein, J. (2009). Screening and brief intervention to reduce marijuana use among youth and young adults in a pediatric emergency department. *Acad Emerg Med*, 16(11):1174–85.

Bhorade, A. M., Gordon, M. O., Wilson, B., Weinreb, R. N., and Kass, M. A. (2009). Variability of intraocular pressure measurements in observation participants in the ocular hypertension treatment study. *Ophthalmology*, 116(4):717–24.

Bjorkedal, E. and Flaten, M. (2011). Interaction between expectancies and drug effects: an experimental investigation of placebo analgesia with caffeine as an active placebo. *Psychopharmacology*, 215(3):537–548.

Bland, M. (2000). *An Introduction to medical statistics*. Oxford : Oxford University Press, 2000, Oxford, 3rd ed. edition.

Bradford-Hill, A. (1990). Memories of the british streptomycin trial in tuberculosis: the first randomized clinical trial. *Controlled clinical trials*, 11(2):77–79.

Brand, C., Snaddon, J., Bailey, M., and Cicuttini, F. (2001). Vitamin e is ineffective for symptomatic relief of knee osteoarthritis: a six month double blind, randomised, placebo controlled study. *Ann Rheum Dis*, 60(10):946–9.

Bridges, T. (1997). The influence of worm age, duration of exposure and endpoint selection on bioassay sensitivity for neanthes arenaceodentata (annelida: Polychaeta). *Environmental Toxicology and Chemistry*, 16(8):1650–1658.

Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in medicine*, 14(17):1933–1940.

Brush, G. G. (1988). *How to Choose the Proper Sample Size*, volume 12 of *ASQC Statistical How-To Series*. ASQC Quality Press, Milwaukee, WI.

Burge, P. S., Calverley, P. M., Jones, P. W., Spencer, S., and Anderson, J. A. (2003). Prednisolone response in patients with chronic obstructive pulmonary disease: results from the isolde study. *Thorax*, 58(8):654–8.

Burneo, J. G., Montori, V. M., and Faught, E. (2002). Magnitude of the placebo effect in randomized trials of antiepileptic agents. *Epilepsy Behav*, 3(6):532–534.

Burnham, D. B., Miller, D., Karlstadt, R., Friedman, C. J., and Palmer, R. H. (1994). Famotidine increases plasma alcohol concentration in healthy subjects. *Aliment Pharmacol Ther*, 8(1):55–61.

Campbell, D. T. and Kenny, D. A. (1999). *A primer on regression artifacts*.

Campbell, M. J. (1990). *Medical statistics : a commonsense approach*. Chichester : Wiley, 1990, Chichester.

Campbell, M. J. (2013). Doing clinical trials large enough to achieve adequate reductions in uncertainties about treatment effects. *Journal of the Royal Society of Medicine*, 106(2):6871.

Campbell, M. J., Julious, S. A., and Altman, D. G. (1995). Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ: British Medical Journal*, 311(7013):1145–1148.

Cannito, M. P., Suiter, D. M., Beverly, D., Chorna, L., Wolf, T., and Pfeiffer, R. M. (2012). Sentence intelligibility before and after voice treatment in speakers with idiopathic parkinson's disease. *J Voice*, 26(2):214–9.

Chan, I. (2003). Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. *Statistical Methods In Medical Research*, 12(1):37–58.

Chapurlat, R. D., Blackwell, T., Bauer, D. C., and Cummings, S. R. (2001). Changes in biochemical markers of bone turnover in women treated with raloxifene: influence of regression to the mean. *Osteoporos Int*, 12(12):1006–14.

Chen, J. J., Tsong, Y., and Kang, S.-H. (2000). Tests for equivalence or noninferiority between two proportions*. *Drug Information Journal*, 34(2):569–578.

Ching-Ping, H. (2016). Investigating the effect of novel adaptive designs in human studies of deodorants and their impact on the recruited number of subjects. Master's thesis.

Chow, S.-C., Shao, J., and Wang, H. (2002). A note on sample size calculation for mean comparisons based on noncentral t -statistics. *Journal of Biopharmaceutical Statistics*, 12(4):441–456.

Chuang-Stein, C. and Kirby, S. (2014). The shrinking or disappearing observed treatment effect. *Pharm Stat*, 13(5):277–80.

ChuangStein, C. and Kirby, S. (2014). The shrinking or disappearing observed treatment effect. *Pharmaceutical Statistics*, 13(5):277–280.

Clarke, C. E., Patel, S., Ives, N., Rick, C. E., Woolley, R., Wheatley, K., Walker, M. F., Zhu, S., Kandiyali, R., Yao, G., and Sackley, C. M. (2016). Clinical effectiveness and cost-effectiveness of physiotherapy and occupational therapy versus no therapy in mild to moderate parkinson's disease: a large pragmatic randomised controlled trial (pd rehab). *Health Technol Assess*, 20(63):1–96.

Cocks, K., King, M., Velikova, G., St-James, M. M., Fayers, P., and Brown, J. (2011). Evidence-based guidelines for determination of sample size and interpretation of the european organisation for the research and treatment of cancer quality of life questionnaire core 30. *J. Clin. Oncol.*, 29(1):89–96.

Cocks, K., King, M. T., Velikova, G., Fayers, P. M., and Brown, J. M. (2008). Quality, interpretation and presentation of european organisation for research and treatment of cancer quality of life questionnaire core 30 data in randomised controlled trials. *European Journal of Cancer*, 44(13):1793–1798.

Cocks, K. and Torgerson, D. J. (2013). Sample size calculations for pilot randomized trials: a confidence interval approach. *Journal of clinical epidemiology*, 66(2):197–201.

Cohen, J. (1973). Statistical power analysis and research results. *American Educational Research Journal*, 10(3):225–229.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ : Lawrence Erlbaum Associates, 1988, Hillsdale, NJ], 2nd ed. edition.

Conboy, L. A., Wasserman, R. H., Jacobson, E. E., Davis, R. B., Legedza, A. T., Park, M., Rivers, A. L., Morey, E. B., Nam, B. H., Lasagna, L., Kirsch, I., Lembo, A. J., Kaptchuk, T. J., and Kerr, C. E. (2006). Investigating placebo effects in irritable bowel syndrome: a novel research design. *Contemp Clin Trials*, 27(2):123–34.

Connett, J. E., Smith, J. A., and McHugh, R. B. (1987). Sample size and power for pair-matched case-control studies. *Stat Med*, 6(1):53–9.

CONSORT (2010). Consort statement checklist. `http://www.consort-statement.org/checklists/view/32-consort/66-title`. Accessed: 2016-10-10.

Cook, J., Julious, S., Hampson, L., Hewitt, C., Berlin, J., Ashby, D., Sones, W., Emsley, R., Fergusson, D., Walters, S., Wilson, E., Maclennan, G., Stallard, N., Rothwell, J., Bland, M., Smith, R., Brown, L., Ramsay, C., Cook, A., Armstrong, D., Altman, D., and Vale, L. (2018). Choosing the target difference ("effect size") for a randomised controlled trial - delta2 guidance.

Cook, J., Julious, S., Sones, W., Rothwell, J., Ramsay, C., Hampson, L., Emsley, R., Walters, S., Hewitt, C., Bland, M., et al. (2017). Choosing the target difference (effect size) for a randomised controlled trial-delta 2 guidance protocol. *Trials*, 18(1):271.

Cook, J. A., Hislop, J., Adewuyi, T. E., Harrild, K., Altman, D. G., Ramsay, C. R., Fraser, C., Buckley, B., Fayers, P., Harvey, I., Briggs, A. H., Norrie, J. D., Fergusson, D., Ford, I., and Vale, L. D. (2014). Assessing methods to specify the target difference for a randomised controlled trial: Delta (difference elicitation in trials) review. *Health technology assessment (Winchester, England)*, 18(28):v–vi, 1–175.

Cook, J. A., Hislop, J., Altman, D. G., Fayers, P., Briggs, A. H., Ramsay, C. R., Norrie, J. D., Harvey, I. M., Buckley, B., Fergusson, D., Ford, I., and Vale, L. D. (2015). Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. *Trials*, 16.

Cooper, S. P., Hardy, R. J., Labarthe, D. R., Hawkins, C. M., Smith, E. O., Blaufox, M. D., Cooper, C. J., Entwisle, G., and Maxwell, M. H. (1988). The relation between degree of blood pressure reduction and mortality among hypertensives in the hypertension detection and follow-up program. *Am J Epidemiol*, 127(2):387–403.

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., and Schuler, T. C. (2007). Understanding the minimum clinically important difference: a review of concepts and methods. *The spine journal : official journal of the North American Spine Society*, 7(5):541.

Council, N. R. (2010). The prevention and treatment of missing data in clinical trials. https://www.ncbi.nlm.nih.gov/books/NBK209904/. Accessed: 2017-08-24.

Counsell, N., Biri, D., Fraczek, J., and Hackshaw, A. (2017). Publishing interim results of randomised clinical trials in peer-reviewed journals. *Clinical Trials*, 14(1):67–77.

Cranney, A., Welch, V., Wells, G., Adachi, J., Shea, B., Simon, L., and Tugwell, P. (2001). Discrimination of changes in osteoporosis outcomes. *Journal of Rheumatology*, 28(2):413–421.

Crawford, M. R., Bartlett, D. J., Coughlin, S. R., Phillips, C. L., Neill, A. M., Espie, C. A., Dungan, G. C., n., Wilding, J. P., Calverley, P. M., Grunstein, R. R., and Marshall, N. S. (2012). The effect of continuous positive airway pressure usage on sleepiness in obstructive sleep apnoea: real effects or expectation of benefit? *Thorax*, 67(10):920–4.

Crosby, R. D., Kolotkin, R. L., and Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of clinical epidemiology*, 56(5):395.

Cummings, J. L., Tractenberg, R. E., Gamst, A., Teri, L., Masterman, D., and Thal, L. J. (2004). Regression to the mean: implications for clinical trials of psychotropic agents in dementia. *Curr Alzheimer Res*, 1(4):323–8.

Cummings, S. R., Palermo, L., Browner, W., Marcus, R., Wallace, R., Pearson, J., Blackwell, T., Eckert, S., and Black, D. (2000). Monitoring osteoporosis therapy with bone densitometry: misleading changes and regression to the mean. fracture intervention trial research group. *Jama*, 283(10):1318–21.

Davis, C. E. (1976). The effect of regression to the mean in epidemiologic and clinical studies. *American journal of epidemiology*, 104(5):493.

Denke, M. A. and Frantz, I. D. (1993). Response to a cholesterol- lowering diet: Efficacy is greater in hypercholesterolemic subjects even after adjustment for regression to the mean. *The American Journal of Medicine*, 94(6):626–631.

DeRogatis, L. R., Graziottin, A., Bitzer, J., Schmitt, S., Koochaki, P. E., and Rodenberg, C. (2009). Clinically relevant changes in sexual desire, satisfying sexual activity and personal distress as measured by the profile of female sexual function, sexual activity log, and personal distress scale in postmenopausal women with hypoactive sexual desire disorder. *J Sex Med*, 6(1):175–83.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., and Smith, H., J. (1987). Publication bias and clinical trials. *Control Clin Trials*, 8(4):343–53.

Dunnett, C. W. and Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics*, 33(4):593.

Durham, T. A. (2008). *Introduction to statistics in pharmaceutical clinical trials*. London ; Chicago : Pharmaceutical Press, 2008, London ; Chicago.

Ederer, F. (1972). Serum cholesterol changes: Effects of diet and regression toward the mean. *Journal of Chronic Diseases*, 25(5):277–289.

Elton, P. J., Ryman, A., Hammer, M., and Page, F. (1994). Randomised controlled trial in northern england of the effect of a person knowing their own serum cholesterol concentration. *Journal of Epidemiology and Community Health (1979-)*, 48(1):22–25.

Emerson, S. and Fleming, T. (1990). Parameter-estimation following group sequential hypothesis-testing. *Biometrika*, 77(4):875–892.

Enck, P. and Klosterhalfen, S. (2005). The placebo response in functional bowel disorders: perspectives and putative mechanisms. *Neurogastroenterology & Motility*, 17(3):325–331.

Everitt, B. S. e. and Palmer, C. e. (2010). *Encyclopaedic Companion to Medical Statistics*. Wiley.

Farlow, M. R., Small, G. W., Quarg, P., and Krause, A. (2005). Efficacy of rivastigmine in alzheimer's disease patients with rapid disease progression: results of a meta-analysis. *Dement Geriatr Cogn Disord*, 20(2-3):192–7.

Fayers, P. and Hays, R. D. (2014). Don't middle your mids: regression to the mean shrinks estimates of minimally important differences. *Qual. Life Res.*, 23(1):1–4.

Fayers, P. M. and Machin, D. (1995). Sample-size - how many patients are necessary. *British Journal of Cancer*, 72(1):1–9.

Fethney, J. (2010). Statistical and clinical significance, and how to use confidence intervals to help interpret both. *Australian Critical Care*, 23(2):93–97.

Finney, J. W. (2008). Regression to the mean in substance use disorder treatment research. *Addiction*, 103(1):42–52.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York ; Chichester : Wiley, c1986, New York ; Chichester.

Fleiss, J. L. and Levin, B. (1988). Sample size determination in studies with matched pairs. *Journal of Clinical Epidemiology*, 41(8):727–730.

Fleming, T. R. (1982). One-sample multiple testing procedure for phase ii clinical trials. *Biometrics*, 38(1):143.

Flight, L. and Julious, S. A. (2016). Practical guide to sample size calculations: superiority trials. *Pharmaceutical Statistics*, 15(1):75–79.

for the Evaluation of Medicinal Products, E. A. (2000). Points to consider on switching between supriority and non-inferiority.

Forbes, G. and Holt, H. (1948). Streptomycin treatment of pulmonary tuberculosis.

Friedman, L. M. (1985). *Fundamentals of clinical trials.* Littleton, Mass. : PSG Pub. Co., 1985, Littleton, Mass., 2nd ed. edition.

Friedman, L. M. (2010). *Fundamentals of clinical trials.* New York : Springer, c2010, New York.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Gardner, M. J. and Heady, J. A. (1973). Some effects of within- person variability in epidemiological studies. *Journal of Chronic Diseases*, 26(12):781–795.

GL Burrell, F Thoemmes, D. M. (2010). Visual displays of regression toward the mean using sas sgplot.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995–1004.

Gray, A. J., Goodacre, S., Newby, D. E., Masson, M. A., Sampson, F., Dixon, S., Crane, S., Elliott, M., and Nicholl, J. (2009). A multicentre randomised controlled trial of the use of continuous positive airway pressure and non-invasive positive pressure ventilation in the early treatment of patients presenting to the emergency department with severe acute cardiogenic pulmonary oedema: the 3cpo trial. *Health Technol Assess*, 13(33):1–106.

Greineder, D. K., Loane, K. C., and Parks, P. (1999). A randomized controlled trial of a pediatric asthma outreach program. *J Allergy Clin Immunol*, 103(3 Pt 1):436–40.

Guenther, W. (1981). Sample size formulas for normal theory t tests. *The American Statistician*, 35(4):243–244.

Guideline, I. H. T. (1995). Structure and content of clinical study reports e3. *Recommended for Adoption at Step*, 4.

Guyatt, G., Walter, S., and Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, 40(2):171–178.

Hanson, M., Sanderson, H., and Solomon, K. (2003a). Variation, replication, and power analysis of myriophyllum spp. microcosm toxicity data. *Environmental toxicology and chemistry*, 22(6):1318–1329.

Hanson, M., Sanderson, H., and Solomon, K. (2003b). Variation, replication, and power analysis of myriophyllum spp. microcosm toxicity data. *Environmental toxicology and chemistry*, 22(6):1318.

Harris, J., Brand, J., Cote, M., Faucett, S., and Dhawan, A. (2017). Research pearls: The significance of statistics and perils of pooling. part 1: Clinical versus statistical significance. *Arthroscopy: The Journal of Arthroscopic & Related Surgery.*

Hauschke, D., Steinijans, V., Diletti, E., and Burke, M. (1992). Sample size determination for bioequivalence assessment using a multiplicative model. *Journal of Pharmacokinetics and Biopharmaceutics*, 20(5):557–561.

Hays, R. and Woolley, J. (2000). The concept of clinically meaningful difference in health-related quality-of-life research. *Pharmacoeconomics*, 18(5):419–423.

Heather, N. (2014). Interpreting null findings from trials of alcohol brief interventions. *Front Psychiatry*, 5:85.

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, 2(3):167–171.

Heimendinger, J. and Laird, N. (1983). Growth changes: Measuring the effect of an intervention. *Evaluation Review*, 7(1):80–95.

Hill, C. J., Bloom, H., Black, A., and Lipsey, M. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3):172–177.

Holland, J. (2013). Fixing a broken drug development process. *Journal of Commercial Biotechnology*, 19(1):5–6.

Hulley, S. B. (2013). *Designing clinical research.* Philadelphia, Pennsylvania : Lippincott Williams and Wilkins, 2013, 4th edition / stephen b. hulley, steven r. cummings, warren s. browner, deborah g. grady and thomas b. newman. edition.

ICH-E9 (2014). Addendum to statistical principles for clinical trials on choosing appropriate estimands and defining sensitivity analyses in clinical trials. Accessed: 2017-08-24.

Ioannidis, J. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648.

Ioannidis, J. P., Munaf, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, 18(5):235241.

Irwig, L., Glasziou, P., and Wilson, A. (1990). Estimating an individuals true cholesterol level and response to intervention. *Medical Decision Making*, 10(4):327–327.

Jacobson, N. S., Follette, W. C., and Revenstorf, D. (1984). Psychotherapy outcome research - methods for reporting variability and evaluating clinical-significance. *Behavior Therapy*, 15(4):336–352.

Jaeschke, R., Singer, J., and Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, 10(4):407–415.

James, K. E. (1973). Regression toward the mean in uncontrolled clinical studies. *Biometrics*, 29(1):121.

Jeffcoate, W. J., Price, P. E., Phillips, C. J., Game, F. L., Mudge, E., Davies, S., Amery, C. M., Edmonds, M. E., Gibby, O. M., Johnson, A. B., Jones, G. R., Masson, E., Patmore, J. E., Price, D., Rayman, G., and Harding, K. G. (2009). Randomised controlled trial of the use of three dressing preparations in the management of chronic ulceration of the foot in diabetes. *Health Technol Assess*, 13(54):1–86, iii–iv.

Johnson, A. and Thomopoulos, N. (2002). Use of the left-truncated normal distribution for improving achieved service levels. In *Proceedings of the 2002 Annual Meeting of the Decision Sciences Institute*, pages 2033–2041.

Johnson, K., McMorris, B., Raynor, L., and Monsen, K. (2013). What big size you have! using effect sizes to determine the impact of public health nursing interventions. *Appl. Clin. Inform.*, 4(3):434–444.

Johnston, M., Hays, R. D., and Hui, K. (2009). Evidence-based effect size estimation: An illustration using the case of acupuncture for cancer-related fatigue. *Bmc Complementary And Alternative Medicine*, 9.

Jones, E., Jarvis, P., Lewis, J., and Ebbutt, A. (1996). Trials to assess equivalence: The importance of rigorous methods. *British Medical Journal*, 313(7048):36–39.

Julious, S. (2010a). *An introduction to statistics in early phase trials [electronic resource]*. Wiley-Blackwell, Oxford.

Julious, S. and Zariffa, N. (2002). The abc of pharmaceutical trial design: some basic principles. *Pharmaceutical Statistics*, 1(1):45–53.

Julious, S. A. (2004). Sample sizes for clinical trials with normal data. *Statistics in Medicine*, 23(12):1921–1986.

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics*, 4(4):287–291.

Julious, S. A. (2010b). *Sample sizes for clinical trials [electronic resource]*. Boca Raton : CRC Press/Taylor & Francis, c2010, Boca Raton.

Julious, S. A. and Campbell, M. J. (2012). Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. *Statistics in Medicine*, 31(24):2904–2936.

Julious, S. A., Campbell, M. J., and Altman, D. G. (1999). Estimating sample sizes for continuous, binary, and ordinal outcomes in paired comparisons: Practical hints. *Journal of Biopharmaceutical Statistics*, 9(2):241–251.

Julious, S. A., Campbell, M. J., and Walters, S. J. (2007). Predicting where future means will lie based on the results of the current trial. *Contemporary Clinical Trials*, 28(4):352–357.

Julious, S. A. and Walters, S. J. (2014). Estimating effect sizes for health-related quality of life outcomes. *Stat. Methods Med. Res.*, 23(5):430–439.

Kaul, S. and Diamond, G. A. (2006). Good enough: A primer on the analysis and interpretation of noninferiority trials. *Annals of Internal Medicine*, 145(1):62–69.

Kazis, E. L., Anderson, J. J., and Meenan, F. R. (1989). Effect sizes for interpreting changes in health status. *Medical Care*, 27(3 Suppl):S178–S189.

Khanna, D., Tseng, C. H., Furst, D. E., Clements, P. J., Elashoff, R., Roth, M., Elashoff, D., and Tashkin, D. P. (2009). Minimally important differences in the mahler's transition dyspnoea index in a large randomized controlled trial–results from the scleroderma lung study. *Rheumatology (Oxford)*, 48(12):1537–40.

Kieser, M. and Hauschke, D. (2005). Assessment of clinical relevance by considering point estimates and associated confidence intervals. *Pharm. Stat.*, 4(2):101–107.

Kirby, S., Burke, J., Chuang-Stein, C., and Sin, C. (2012). Discounting phase 2 results when planning phase 3 clinical trials. *Pharmaceutical Statistics*, 11(5):373–385.

Klosterhalfen, S. and Enck, P. (2006). Psychobiology of the placebo response. *Auton Neurosci*, 125(1-2):94–9.

Kortum, S. (2002). Lecture 4: Selection. `http://web.ist.utl.pt/~ist11038/compute/qc/,truncG/lecture4k.pdf`.

Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., and Yesavage, J. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives Of General Psychiatry*, 63(5):484–489.

Krause, A. and Pinheiro, J. (2007). Modeling and simulation to adjust p values in presence of a regression to the mean effect. *Am. Stat.*, 61(4):302–307.

Krum, H. and Tonkin, A. (2003). *Why do phase III trials of promising heart failure drugs often fail? The contribution of "regression to the truth"*, volume 9, pages 364–7. United States.

Lamb, S. E., Lall, R., Hansen, Z., Castelnuovo, E., Withers, E. J., Nichols, V., Griffiths, F., Potter, R., Szczepura, A., and Underwood, M. (2010). A multi-centred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. the back skills training (best) trial. *Health Technol Assess*, 14(41):1–253, iii–iv.

Lee, E. C., Whitehead, A. L., Jacques, R. M., and Julious, S. A. (2014a). The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC medical research methodology*, 14(1):41.

Lee, E. C., Whitehead, A. L., Jacques, R. M., and Julious, S. A. (2014b). The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC medical research methodology*, 14:41.

Lenth, R. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193.

Lewis, J. (1999). Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in medicine*, 18(15):1903–1942.

Lin, H. M. and Hughes, M. D. (1995). Use of historical marker data for assessing treatment effects in phase i/ii trials when subject selection is determined by baseline marker level. *Biometrics*, 51(3):1053–63.

Lin, H. M. and Hughes, M. D. (1996). Analysis of uncontrolled treatment changes in hiv clinical trails. *Stat Med*, 15(19):2053–67.

Lipman, E. L., Boyle, M. H., Cunningham, C., Kenny, M., Sniderman, C., Duku, E., Mills, B., Evans, P., and Waymouth, M. (2006). Testing effectiveness of a community-based aggression management program for children 7 to 11 years old and their families. *J Am Acad Child Adolesc Psychiatry*, 45(9):1085–93.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, 16(4):421–437.

Lubsen, J., Voko, Z., Poole-Wilson, P. A., Kirwan, B. A., and de Brouwer, S. (2007). Blood pressure reduction in stable angina by nifedipine was related to stroke and heart failure reduction but not to coronary interventions. *J Clin Epidemiol*, 60(7):720–6.

Maltby, N., Mayers, M. F., Allen, G. J., and Tolin, D. F. (2005). Anxiety sensitivity: stability in prospective research. *J Anxiety Disord*, 19(6):708–16.

Manley, S. E., Stratton, I. M., Cull, C. A., Frighi, V., Eeley, E. A., Matthews, D. R., Holman, R. R., Turner, R. C., and Neil, H. A. (2000). Effects of three months' diet after diagnosis of type 2 diabetes on plasma lipids and lipoproteins (ukpds 45). uk prospective diabetes study group. *Diabet Med*, 17(7):518–23.

Martinez-Yelamos, S., Martinez-Yelamos, A., Martin Ozaeta, G., Casado, V., Carmona, O., and Arbizu, T. (2006). Regression to the mean in multiple sclerosis. *Mult Scler*, 12(6):826–9.

Mbizvo, G. K., Nolan, S. J., Nurmikko, T. J., and Goebel, A. (2015). Placebo responses in long-standing complex regional pain syndrome: a systematic review and meta-analysis. *J Pain*, 16(2):99–115.

McAlister, F. A., Connor, A. M., Wells, G., Grover, S. A., and Laupacis, A. (2000). When should hypertension be treated? the different perspectives of canadian family physicians and patients. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 163(4):403.

McCall, W. V., D'Agostino, R., J., Rosenquist, P. B., Kimball, J., Boggs, N., Lasater, B., and Blocker, J. (2011). Dissection of the factors driving the placebo effect in hypnotic treatment of depressed insomniacs. *Sleep Med*, 12(6):557–64.

McCambridge, J., Kypri, K., and McElduff, P. (2014). Regression to the mean and alcohol consumption: A cohort study exploring implications for the interpretation of change in control groups in brief intervention trials. *Drug And Alcohol Dependence*, 135:156–159.

McDonald, J. P. (2003). A review of surgical treatment for obstructive sleep apnoea/hypopnoea syndrome. *The Surgeon*, 1(5):259–264.

McGrath, R. and Meyer, G. (2006). When effect sizes disagree: The case of r and d. *Psychol. Methods*, 11(4):386–401.

Meinert, C. L. (1986). *Clinical trials [electronic resource] : design, conduct and analysis.* New York ; Oxford : Oxford University Press, 1986, New York ; Oxford.

Moebus, S., Lehmann, N., Bodeker, W., and Jockel, K. (2006). An analysis of sickness absence in chronically ill patients receiving complementary and alternative medicine: A longterm prospective intermittent study. *Bmc Public Health*, 6.

Morton, V. and Torgerson, D. (2005). Regression to the mean: treatment effect without the intervention. *Journal Of Evaluation In Clinical Practice*, 11(1):59–65.

Network, U. K. C. R. (2015). `http://public.ukcrn.org.uk/search/`.

Neyman, J. and Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 20A(1/2):175–240.

Neyman, J. and Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part ii. *Biometrika*, 20A(3/4):263–294.

Neyman, J. and Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337.

Neyman, J. and Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Math. Proc. Camb. Phil. Soc.*, 29(4):492–510.

Novack, G. D. and Crockett, R. S. (2009). Regression to the mean. *Ocular Surface*, 7(3):163–165.

of Health Research, N. I. (2010a). Health technology assessment.

of Health Research, N. I. (2010b). Health technology assessmet journals library.

of Health Research, N. I. (2010c). Nihr- our purpose.

Okin, P. M., Devereux, R. B., Jern, S., Julius, S., Kjeldsen, S. E., and Dahlof, B. (2001). Relation of echocardiographic left ventricular mass and hypertrophy to persistent electrocardiographic left ventricular hypertrophy in hypertensive patients: the life study. *Am J Hypertens*, 14(8 Pt 1):775–82.

Olive, D. (2015). Chapter 4 - truncated distributions. `http://lagrange.math.siu.edu/Olive/`.

Osterbrand, M., Fahlen, M., Odn, A., and Eliasson, B. (2007). A method to predict the metabolic effects of changes in insulin treatment in subgroups of a large population based patient cohort. *Eur J Epidemiol*, 22(3):151–157.

Owen, D. B. (1965). A special case of a bivariate non-central t-distribution. *Biometrika*, 52(3/4):437–446.

Oxberry, S. G., Bland, J. M., Clark, A. L., Cleland, J. G. F., and Johnson, M. J. (2012). Minimally clinically important difference in chronic breathlessness: Every little helps. *American Heart Journal*, 164(2):229–235.

Palmieri, V., Dahlof, B., DeQuattro, V., Sharpe, N., Bella, J. N., de Simone, G., Paranicas, M., Fishman, D., and Devereux, R. B. (1999). Reliability of echocardiographic assessment of left ventricular structure and function: the preserve study. prospective randomized study evaluating regression of ventricular enlargement. *J Am Coll Cardiol*, 34(5):1625–32.

Pereira, T., Horwitz, R., and Ioannidis, J. (2012). Empirical evaluation of very large treatment effects of medical interventions. *JAMA*, 308(16):1676–1684.

Persu, A., Jin, Y., Elmula, F., Jacobs, L., Renkin, J., and Kjeldsen, S. (2014). Renal denervation after symplicity htn-3: An update. *Current Hypertension Reports*, 16(8).

Pijls, L., de Vries, H., Donker, A., and van Eijk, J. (1999). Reproducibility and biomarker-based validity and responsiveness of a food frequency questionnaire to estimate protein intake. *Am J Epidemiol*, 150(9):987–95.

Pinheiro, J. and Demets, D. (1997). Estimating and reducing bias in group sequential designs with gaussian independent increment structure. *Biometrika*, 84(4):831–845.

Piva, S., Fitzgerald, G., Irrgang, J., Bouzubar, F., and Starz, T. (2004). Get up and go test in patients with knee osteoarthritis 1, 2. *Archives of physical medicine and rehabilitation*, 85(2):284–289.

Qouta, S. R., Palosaari, E., Diab, M., and Punamaki, R. L. (2012). Intervention effectiveness among war-affected children: a cluster randomized controlled trial on improving mental health. *J Trauma Stress*, 25(3):288–98.

Ravelo, A., Watanabe Jonathan, H., Gries Katharine, S., Campbell Jonathan, D., Dmochowski Roger, R., and Sullivan Sean, D. (2009). Treatment success for overactive bladder with urinary urge incontinence refractory to oral antimuscarinics: a review of published evidence. *BMC Urology*, 9(1).

Reeves, B., Pike, K., Rogers, C., Brierley, R., Stokes, E., Wordsworth, S., Nash, R., Miles, A., Mumford, A., Cohen, A., Angelini, G., and Murphy, G. (2016). A multicentre randomised controlled trial of transfusion indication threshold reduction on transfusion rates, morbidity and healthcare resource use following cardiac surgery (titre2). *Health Technology Assessment*, 20(60).

Research, P. and of America, M. (2015). Biopharmaceutical research industry 2015 profile. `http://phrma-docs.phrma.org/sites/default/files/pdf/2014_PhRMA_PROFILE.pdf`.

Richards, D., Hughes-Morley, A., Hayes, R., Araya, R., Barkham, M., Bland, J., Bower, P., Cape, J., Chew-Graham, C., Gask, L., and et al. (2009). Collaborative depression trial (cadet): multi-centre randomised controlled trial of collaborative care for depression - study protocol. *BMC Health Services Research*, 9(1).

Richards, D. A., Bower, P., Chew-Graham, C., Gask, L., Lovell, K., Cape, J., Pilling, S., Araya, R., Kessler, D., Barkham, M., Bland, J. M., Gilbody, S., Green, C., Lewis, G., Manning, C., Kontopantelis, E., Hill, J. J., Hughes-Morley, A., and Russell, A. (2016). Clinical effectiveness and cost-effectiveness of collaborative care for depression in uk primary care (cadet): a cluster randomised controlled trial. *Health Technol Assess*, 20(14):1–192.

Rocconi, L. M. and Ethington, C. A. (2009). Assessing longitudinal change: Adjustment for regression to the mean effects. *Research in Higher Education*, 50(4):368–376.

Rose, G., Heller, R. F., Pedoe, H. T., and Christie, D. G. (1980). Heart disease prevention project: a randomised controlled trial in industry. *Br Med J*, 280(6216):747–51.

Ross, D. C. (1995). Using severity rather than randomization for assignment in clinical trials: Monte carlo tests and practical illustrations of the robbins-zhang method. *J Psychiatr Res*, 29(4):315–32.

Rothmann, M., Li, N., Chen, G., Chi, G. Y. H., Temple, R., and Tsou, H.-H. (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine*, 22(2):239–264.

Rothwell, J., Julious, S., Campbell, M., and Cooper, C. (2018a). Handbook of statistical methods in randomised controlled trials.

Rothwell, J., Julious, S., Campbell, M., and Cooper, C. (2018b). A study of target effect sizes in randomised controlled trials published in the health technology assessment journal.

Royston, P. (1993). Exact conditional and unconditional sample size for pair-matched studies with binary outcome: a practical guide. *Stat Med*, 12(7):699–712.

Ryvicker, M., Feldman, P. H., Rosati, R. J., Sobolewski, S., Maduro, G. A., J., and Schwartz, T. (2011). Improving functional outcomes in home care patients: impact and challenges of disseminating a quality improvement initiative. *J Healthc Qual*, 33(5):28–36.

Sagar, S. M. (2008). Acupuncture as an evidence-based option for symptom control in cancer patients. *Curr Treat Options Oncol*, 9(2-3):117–26.

Salter, G., Roman, M., Bland, M., and Macpherson, H. (2006). Acupuncture for chronic neck pain: a pilot for a randomised controlled trial. *Bmc Musculoskeletal Disorders*, 7.

Samsa, G., Edelman, D., Rothman, M., Williams, G., Lipscomb, J., and Matchar, D. (1999). Determining clinically important differences in health status measures - a general approach with illustration to the health utilities index mark ii. *Pharmacoeconomics*, 15(2):141–155.

Sanders-van Wijk, S., Maeder, M. T., Nietlispach, F., Rickli, H., Estlinbaum, W., Erne, P., Rickenbacher, P., Peter, M., Pfisterer, M. P., and Brunner-La Rocca, H. P. (2014). Long-term results of intensified, n-terminal-pro-b-type natriuretic peptide-guided versus symptom-guided treatment in elderly patients with heart failure: five-year follow-up from time-chf. *Circ Heart Fail*, 7(1):131–9.

Schlesselman, J. J. and Schneiderman, M. A. (1982). Case control studies: Design, conduct, analysis. *Journal of Occupational and Environmental Medicine*, 24(11):879.

Schwartz, D. (1980). *Clinical trials.* London : Academic Press, 1980, London.

Senn, S. (1993). *Cross-over trials in clinical research.* Chichester : Wiley, c1993, Chichester.

Senn, S. and Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharm. Stat.*, 6(3):161–170.

Senn, S. J., Brown, R. A., and James, K. E. (1985). Estimating treatment effects in clinical trials subject to regression to the mean. *Biometrics*, 41(2):555–560.

Shapiro, S. H. and Louis, T. A. (1983). *Clinical trials : issues and approaches.* New York : M. Dekker, c1983, New York.

Shaw, L., Rodgers, H., Price, C., van Wijck, F., Shackley, P., Steen, N., Barnes, M., Ford, G., and Graham, L. (2010). Botuls: a multicentre randomised controlled trial to evaluate the clinical effectiveness and cost-effectiveness of treating upper limb spasticity due to stroke with botulinum toxin type a. *Health Technol Assess*, 14(26):1–113, iii–iv.

Shepard, D. S. and Finison, L. J. (1983). Blood pressure reductions: Correcting for regression to the mean. *Preventive Medicine*, 12(2):304–317.

Siontis, G. C. M. and Ioannidis, J. P. A. (2011). Risk factors and interventions with statistically significant tiny effects. *International Journal of Epidemiology*, 40(5):1292–1307.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Stat. Med.*, 29(21):2169–2179.

Stewart, R. A., Kittelson, J., and Kay, I. P. (2000). Statistical methods to improve the precision of the treadmill exercise test. *J Am Coll Cardiol*, 36(4):1274–9.

Stone, M. A., Inman, R. D., Wright, J. G., and Maetzel, A. (2004). Validation exercise of the ankylosing spondylitis assessment study (asas) group response criteria in ankylosing spondylitis patients treated with biologics. *Arthritis Care & Research*, 51(3):316–320.

Suissa, S. (2008). *Lung function decline in COPD trials: bias from regression to the mean*, volume 32, pages 829–31. Switzerland.

Sully, B. G. O., Julious, S. A., and Nicholl, J. (2013). A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two uk funding agencies. *Trials*, 14:166–166.

SW, R. (1984). Effect of cigarette smoking cessation on risk factors for coronary atherosclerosis. a control clinical trial. *J Trauma Stress*, 53(2):173–84.

Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., and Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*, 15(1):264.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L., Robson, R., Thabane, M., Giangregorio, L., and Goldsmith, C. (2010a). A tutorial on pilot studies: the what, why and how. *Bmc Medical Research Methodology*, 10.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., Robson, R., Thabane, M., Giangregorio, L., Goldsmith, C. H., and et al. (2010b). A tutorial on pilot studies: the what, why and how. *BMC medical research methodology.*

Thomas, J., Lochbaum, M., Landers, D. M., and He, C. (1997). Planning significant and meaningful research in exercise science: Estimating sample size. *Research Quarterly For Exercise And Sport*, 68(1):33–43.

Tickle, M., O'Neill, C., Donaldson, M., Birch, S., Noble, S., Killough, S., Murphy, L., Greer, M., Brodison, J., Verghis, R., and Worthington, H. V. (2016). A randomised controlled trial to measure the effects and costs of a dental caries prevention regime for young children attending primary care dental services: the northern ireland caries prevention in practice (nic-pip) trial. *Health Technol Assess*, 20(71):1–96.

Tilbrook, H. E., Cox, H., Hewitt, C. E., Kang'ombe, A. R., Chuang, L.-H., Jayakody, S., Aplin, J., Semlyen, A., Trewhela, A., Watt, I., and et al. (2011). Yoga for chronic low back pain: A randomized trial. *Annals of Internal Medicine*, 155(9):569578.

Torgerson, D. J., Ryan, M., and Ratcliffe, J. (1995). Economics in sample size determination for clinical trials. *QJM : monthly journal of the Association of Physicians*, 88(7):517.

Tressoldi, P. E., Giofr, D., Sella, F., and Cumming, G. (2012). High impact = high statistical standards? not necessarily so. *SSRN Electronic Journal*.

Twisk, J. and de Vente, W. (2008). The analysis of randomised controlled trial data with more than one follow-up measurement. a comparison between different approaches. *European journal of epidemiology*, 23(10):655.

Twisk, J. and Proper, K. (2004). Evaluation of the results of a randomized controlled trial: how to define changes between baseline and follow-up. *J Clin Epidemiol*, 57(3):223–8.

Valk, G., Grootenhuis, P., van Eijk, J., Bouter, L., and Bertelsmann, F. (2000). Methods for assessing diabetic polyneuropathy: validity and reproducibility of the measurement of sensory symptom severity and nerve function tests. *Diabetes Res Clin Pract*, 47(2):87–95.

van Schayck, C. P., Dompeling, E., van Herwaarden, C. L., Folgering, H., Akkermans, R. P., van Den Broek, P. J., and van Weel, C. (1995). Continuous and on demand use of bronchodilators in patients with non-steroid dependent asthma and chronic bronchitis: four-year follow-up randomized controlled study. *The British journal of general practice : the journal of the Royal College of General Practitioners*, 45(394):239.

van Tulder, M., Malmivaara, A., Hayden, J., and Koes, B. (2007). Statistical significance versus clinical importance - trials on exercise therapy for chronic low back pain as example. *Spine*, 32(16):1785–1790.

Vickers, A. J. and Altman, D. G. (2001). Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ: British Medical Journal*, 323(7321):1123–1124.

Victora, C. G., Morris, S. S., Barros, F. C., Horta, B. L., Weiderpass, E., and Tomasi, E. (1998). Breast-feeding and growth in brazilian infants. *Am J Clin Nutr*, 67(3):452–8.

Voehringer, P. A. and Ghaemi, S. N. (2011). Solving the antidepressant efficacy question: Effect sizes in major depressive disorder. *Clinical Therapeutics*, 33(12):B49–B61.

Walker, S. P., Grantham-McGregor, S. M., Himes, J. H., Powell, C. A., and Chang, S. M. (1996). Early childhood supplementation does not benefit the long-term growth of stunted children in jamaica. *J Nutr*, 126(12):3017–24.

Wang, S., Hung, J., and Neill, R. (2006). Adapting the sample size planning of a phase iii trial based on phase ii data. *Pharmaceutical Statistics*, 5(2):85–97.

White, P., Lewith, G., Prescott, P., and Conway, J. (2004). Acupuncture versus placebo for the treatment of chronic mechanical neck pain: a randomized, controlled trial. *Ann Intern Med*, 141(12):911–9.

White, W. B., Mehrotra, D. V., Black, H. R., and Fakouhi, T. D. (1997). Effects of controlled-onset extended-release verapamil on nocturnal blood pressure (dippers versus nondippers). coer-verapamil study group. *Am J Cardiol*, 80(4):469–74.

Whitehead, A., Julious, S., Cooper, C., and Campbell, M. (2016). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research*, 25(3):1057–1073.

Whitehead, A., Sully, B., and Campbell, M. (2014). Pilot and feasibility studies: is there a difference from each other and from a randomised controlled trial? *Contemp Clin Trials*, 38(1):130–3.

Whitney, C. W. and Von Korff, M. (1992). Regression to the mean in treated versus untreated chronic pain. *Pain*, 50(3):281–285.

Wiens, B. L. (2002). Choosing an equivalence limit for noninferiority or equivalence studies. *Control Clin Trials*, 23(1):2–14.

Wolfe, F., Michaud, K., and Dewitt, E. M. (2004). Why results of clinical trials and observational studies of antitumour necrosis factor (anti-tnf) therapy differ: methodological and interpretive issues. *Ann Rheum Dis*, 63 Suppl 2:ii13–ii17.

Woods, S. W., Stolar, M., Sernyak, M. J., and Charney, D. S. (2001). Consistency of atypical antipsychotic superiority to placebo in recent clinical trials. *Biological Psychiatry*, 49(1):64–70.

Wu, C., Chuang, L., Lin, K., Lee, S. D., and Hong, W. (2011). Responsiveness, minimal detectable change, and minimal clinically important difference of the nottingham extended activities of daily living scale in patients with improved performance after stroke rehabilitation. *Arch. Phys. Med. Rehabil.*, 92(8):1281–1287.

Wyrwich, K. W., Nienaber, N. A., Tierney, W. M., and Wolinsky, F. D. (1999). Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care*, 37(5):469–478.

Yeo, A. and Qu, Y. (2009). Evaluation of the statistical power for multiple tests: a case study. *Pharm. Stat.*, 8(1):5–11.

Yoshioka, A. (1998). Use of randomisation in the medical research council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940s. *BMJ*, 317(7167):1220–1223.

Yudkin, P. L. and Stratton, I. M. (1996). How to deal with regression to the mean in intervention studies. *Lancet*, 347(8996):241–3.

Zhang, J., He, K., Tang, S., Sridhara, R., Blumenthal, G., and Cortazar, P. (2012). Overestimation of the effect size in group sequential trials. *Clinical Cancer Research*, 18(18):4872–4876.

Zhang, Y., Zhang, S., Thabane, L., Furukawa, T. A., Johnston, B. C., and Guyatt, G. H. (2015). Although not consistently superior, the absolute approach to framing the minimally important difference has advantages over the relative approach. *Journal of Clinical Epidemiology*, 68(8):888–894.

A. Chapter 2

The work in this appendix refers to chapter 2. Since the focus of the thesis has been on parallel group superiority trials, it was decided that whilst these sections are relevant to the wider context, they are not directly relevant to the path of this research.

A.1 Non-Inferiority and Equivalence Trials

A.1.1 Non-Inferiority Trials

Non-inferiority trials are undertaken when there is an active control, such as a current treatment on the market. The aim for this trial is to show that the trial or new treatment is as good as the current treatment. This could be sufficient when the side effects are less with the new treatment, or the cost is less. An example of this type of trial being useful would be to investigate whether doctors could be replaced with nurses for performing a specific therapy or treatment. If the nurses were better it would be an added bonus. This is different to equivalence trials where, in effect, we would wish to show both that nurses were as good as doctors and that doctors were as good as nurses.

A visual representation of this trial is shown in Figure A.1. If the difference is less than $-d_{NI}$, non-inferiority cannot be assumed. Non-inferiority trials have similar problems to superiority trials, since the non-inferiority limit is based on professional opinion which is subjective.

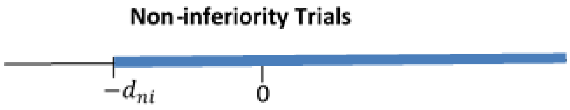


Figure A.1: Non-Inferiority Limit

The hypotheses for these trials are given by

- H_0: The given treatment is inferior with respect to the mean response,

- H_1: The given treatment is non-inferior with respect to the mean response.

This is usually written in terms of the clinical difference, d_{NI}

- H_0: $\mu_A - \mu_B \leq -d_{NI}$
- H_1: $\mu_A - \mu_B > -d_{NI}$.

In this context $-d_{NI}$ is the non-inferiority limit. Non-inferiority trials are tested using a one-sided hypothesis test, equivalent to testing just one part of the two parts of the two one-sided test (TOST) procedure in the equivalence trials. Realistically this is the same principle as getting a $(1 - 2\alpha)100\%$ confidence interval and concluding non-inferiority if the entire interval is greater than $-d_{NI}$.

A.1.1.1 Parallel Group Trials

The methods required to arrive at the sample size calculation follow closely with those from previous sections and have not been included here. These formulae are similar to the formulae for the superiority trials. The sample size for a parallel group non-inferiority trial is

$$n_A = \frac{(r+1)(Z_{1-\beta} + Z_{1-\alpha})^2\sigma^2}{r\big((\mu_A - \mu_B) - d_{NI}\big)^2} \tag{A.1}$$

The non-central t-distribution result is

$$1 - \beta = 1 - T^{-1}\Big(t_{1-\alpha, n_A(r+1)-2}, n_A(r+1) - 2, \tau\Big) \tag{A.2}$$

where

$$\tau = \left| \frac{\big((\mu_A - \mu_B) - d_{NI}\big)\sqrt{rn_A}}{\sqrt{(r+1)\sigma^2}} \right| \tag{A.3}$$

is the non-centrality parameter.

It is worth noting that if $d_{NI} = 0$ in Equation A.1, the formula becomes the same as Equation 2.1 for superiority trials. This shows that the parallel group superiority formula is a special case of the non-inferiority formula.

Appendix A.3 shows various sample sizes needed for different standardised effect sizes and percentage mean differences.

A.1.1.2 Crossover Trials

For the crossover design in non-inferiority trials, we need to include the within-subject standard deviation for the population (σ_w). The general formula for sample sizes in crossover trials is (Julious, 2004)

$$n = \frac{2\sigma_w^2(Z_{1-\beta} + Z_{1-\alpha})^2}{((\mu_A - \mu_B) - d_{NI})^2} \tag{A.4}$$

The non-central t-distribution formula for the power is (Chow et al., 2002)

$$1 - \beta = 1 - T^{-1}\Big(t_{1-\alpha,n-2}, n-2, \tau\Big) \tag{A.5}$$

where

$$\tau = \left| \frac{\Big((\mu_A - \mu_B) - d_{NI}\Big)\sqrt{n}}{\sqrt{2\sigma_w^2}} \right| \tag{A.6}$$

is the non-centrality parameter.

The quick result can be found in Appendix A.2

A.1.2 Equivalence Trials

Equivalence trials are carried out to determine whether two interventions produce the same results for the patients. An example of an equivalence trial is comparing dihydrocodeine and methadone in the treatment of heroin addiction. These are two different substances which are being tested to determine if they have the same effect on the patient. It could be that one method of treatment is cheaper than the other, or easier to formulate. The hypotheses being tested for equivalence trials are (Julious, 2004):

$$H_0 : \mu_A \neq \mu_B$$

$$H_1 : \mu_A = \mu_B$$

This is usually written in terms of the clinical difference, d_e

$$H_0 : \mu_A - \mu_B \leq -d_e \text{ or } \mu_A - \mu_B \geq d_e$$

$$H_1 : -d_e \leq \mu_A - \mu_B \leq d_e$$

Both parts in the null hypothesis need to be rejected in order for a complete rejection of the null hypothesis. This is an example of an intersect-union test. In these tests, each component is tested at an level and this gives a composite test which is also of significance level, α (Berger and Hsu, 1996). Usually we do the two one-sided tests

(TOST) which tests each component of the null hypothesis. This is operationally the same as a $(1-2\alpha)100\%$ confidence interval, where equivalence is established if each end of the confidence interval falls within the region $(-d_e{,}d_e)$. In other words the 95% confidence interval (CI) must lie within $(-d_e{,}d_e)$ in order for the treatments to be deemed equivalent at a 2.5% significance level (Jones et al., 1996).

One consideration for both equivalence and non-inferiority trials is the equivalence or non-inferiority margin. The setting of this can be rather controversial and has been defined as the largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice (for the Evaluation of Medicinal Products, 2000). Commonly used methods to establish the margin are clinical judgement and statistical reasoning. Usually the margin is set at a fraction of the limit of the placebo effect (Kaul and Diamond, 2006; Rothmann et al., 2003).

Often the decision on the equivalence limit is based on some comparison to placebo. The following steps (Agostino et al., 2003; Wiens, 2002) should be considered when determining the limit:

1. We must be confident that the active control would have been different from placebo had one been employed.
2. We should be able to determine that there is no clinically meaningful difference between investigative treatment and control.
3. Through comparing the investigative treatment to control we should indirectly be able to determine that it is superior to placebo.

This limit needs to be established on a study-by-study basis with advice from the relevant agencies involved in the trial. The issues raised for equivalence limits are the same as for non-inferiority limits discussed earlier in the chapter. Figure A.2 is a diagram of how confidence intervals can be used to test the different hypotheses of superiority, equivalence and non-inferiority trials (Julious, 2004).

Figure A.2 represents the area which the 95% confidence interval needs to lie in order to reject the null hypothesis. Notice for the equivalent and non-inferiority trials there is a pre-defined region or value which is close to zero.

A.1.2.1 Parallel Group Trials

General Case

For equivalence trials the sample size cannot be derived directly for the general case where the expected true mean difference is not zero. This is due to there being two one-sided tests being performed, as mentioned above. This results in two chances

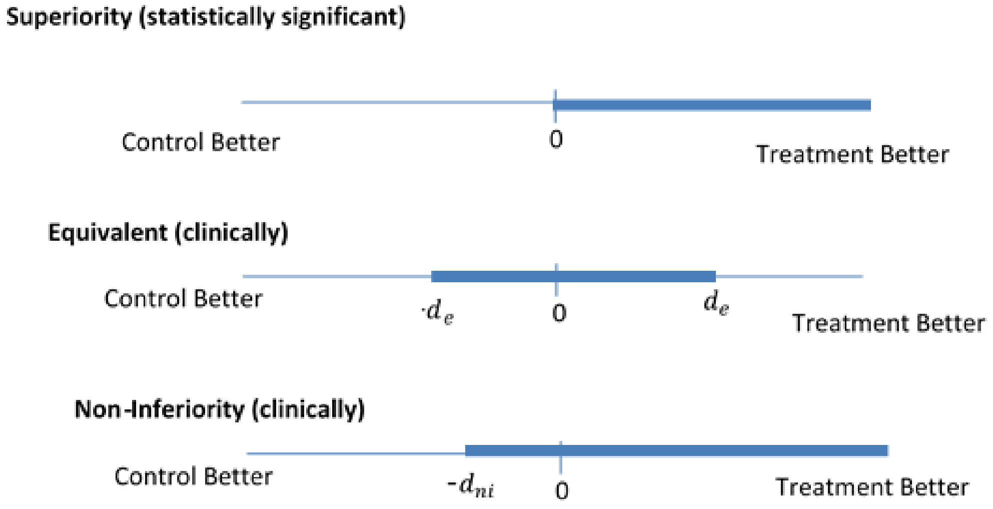


Figure A.2: Illustration of the aims for each trial design

of making a type II error. Hence, the sample size cannot be derived directly for the case where the expected true mean difference is not zero ($\mu_A - \mu_B \neq 0$) since the type II error must be split between the two tests.

The power can therefore be written as (Julious, 2004)

$$1 - \beta = \Phi\left(\sqrt{\frac{rn_A\Big((\mu_A - \mu_B) - d_e\Big)^2}{(r+1)\sigma^2}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{\frac{rn_A\Big((\mu_A - \mu_B) + d_e\Big)^2}{(r+1)\sigma^2}} - Z_{1-\alpha}\right) - 1 \quad \text{(A.7)}$$

This equation is then iterated until the desired power is achieved. As for the superiority trials, when the variance is unknown the $Z-$ values can no longer be used, so the equation is rewritten as (Julious, 2004)

$$1 - \beta = \Phi\left(\sqrt{\frac{rn_A\Big((\mu_A - \mu_B) - d_e\Big)^2}{(r+1)\sigma^2}} - t_{1-\alpha,n_A(r+1)-2}\right) + \Phi\left(\sqrt{\frac{rn_A\Big((\mu_A - \mu_B) + d_e\Big)^2}{(r+1)\sigma^2}} - t_{1-\alpha,n_A(r+1)-2}\right) - 1 \quad \text{(A.8)}$$

Again this uses the non-central Normal-distribution to calculate the type II error rate and power. For the sample size calculations, the approximation below is used. The non-central t-distribution gives (Chow et al., 2002; Hauschke et al., 1992; Owen, 1965)

$$1-\beta = T^{-1}\Big(-t_{1-\alpha,n_A(r+1)-2}, n_A(r+1)-2, \tau_2\Big) - T^{-1}\Big(t_{1-\alpha,n_A(r+1)-2}, n_A(r+1)-2, \tau_1\Big) \tag{A.9}$$

Where

$$\tau_1 = \frac{\Big((\mu_A - \mu_B) + d_e\Big)\sqrt{rn_A}}{(r+1)\sigma^2} \tag{A.10}$$

and

$$\tau_2 = \frac{\Big((\mu_A - \mu_B) - d_e\Big)\sqrt{rn_A}}{(r+1)\sigma^2} \tag{A.11}$$

are the non-centrality parameters.

If we simplify this calculation for an initial sample size for the iterations, we get (Julious, 2004)

$$n_A = \frac{(r+1)\sigma^2(Z_{1-\beta} + Z_{1-\alpha})^2}{r((\mu_A - \mu_B) - d_e)^2} \tag{A.12}$$

Special Case (no treatment difference)

If there is no treatment difference ($\mu_A - \mu_B = 0$), then Equation A.12 becomes (S. A. Julious, 2004)

$$n_A = \frac{(r+1)\sigma^2(Z_{1-\beta/2} + Z_{1-\alpha})^2}{rd_e^2} \tag{A.13}$$

The reason that $1 - \beta/2$ is now used as oppose to $1 - \beta$ is that the type II error is split between the two one-sided tests. However, as the mean difference is now zero the Type II error is split equally so we can have a direct estimate of the sample size. The non-central t-distribution result for the power is (Julious, 2004)

$$1 - \beta = 2 \times T^{-1}\Big(- t_{1-\alpha,n_A(r+1)-2}, n_A(r+1) - 2, \tau\Big) - 1 \tag{A.14}$$

Where

$$\tau = \frac{-\sqrt{n_A} \times rd_e}{\sqrt{(r+1)\sigma^2}} \tag{A.15}$$

is the non-centrality parameter.

A.1.2.2 Crossover Trials

A.1.2.3 General Case

As in previous sections, the power can be estimated using (Julious, 2004)

$$1 - \beta = \Phi\left(\sqrt{\frac{n((\mu_A - \mu_B) - d_e)^2}{2\sigma_w^2}} - Z_{1-\alpha}\right) \\ + \Phi\left(\sqrt{\frac{n((\mu_A - \mu_B) + d_e)^2}{2\sigma_w^2}} - Z_{1-\alpha}\right) - 1 \quad \text{(A.16)}$$

When the population variance is unknown, this equation can be rewritten as

$$1 - \beta = \Phi\left(\sqrt{\frac{n((\mu_A - \mu_B) - d_e)^2}{2\sigma_w^2}} - -t_{1-\alpha,n-2}\right) \\ + \Phi\left(\sqrt{\frac{n((\mu_A - \mu_B) + d_e)^2}{2\sigma_w^2}} - t_{1-\alpha,n-2}\right) - 1 \quad \text{(A.17)}$$

This shows that the methodology for crossover trials is similar to that for parallel group trials for equivalence studies (Chow et al., 2002; Hauschke et al., 1992; Owen, 1965). The non-central t-distribution result for the power is given by (Julious, 2004)

$$1 - \beta = T^{-1}(-t_{1-\alpha,n-2}, n-2, \tau_2) - Probt(t_{1-\alpha,n-2}, n-2, \tau_1) \quad \text{(A.18)}$$

where

$$\tau_1 = \frac{\Big((\mu_A - \mu_B) + d_e\Big)\sqrt{n}}{\sqrt{2\sigma_w^2}} \quad \text{(A.19)}$$

and

$$\tau_2 = \frac{\Big((\mu_A - \mu_B) - d_e\Big)\sqrt{n}}{\sqrt{2\sigma_w^2}} \quad \text{(A.20)}$$

are the non-centrality parameters.

For a quick calculation (for $\mu_A - \mu_B > 0$)(Julious, 2004)

$$n = \frac{2\sigma_w^2(Z_{1-\beta} + Z_{1-\alpha})^2}{\Big((\mu_A - \mu_B) - d_e\Big)^2} \quad \text{(A.21)}$$

For quick calculations at a 90% power and 2.5% type I error rate (Julious, 2004):

$$n = \frac{21\sigma_w^2}{\Big((\mu_A - \mu_B) - d_e\Big)^2} \tag{A.22}$$

Special case of no treatment difference

For the special case where there is no treatment difference ($\mu_A - \mu_B = 0$) the direct estimate of the sample size is (Julious, 2004)

$$n = \frac{2\sigma_w^2(Z_{1-\beta} + Z_{1-\alpha})^2}{d_e^2} \tag{A.23}$$

The quick calculation is (for a 90% power and a 2.5% type I error) (Julious, 2004)

$$n = \frac{26\sigma_w^2}{d_e^2} \tag{A.24}$$

A.2 Quick Results for Sample Size Calculations

A.2.1 Superiority Trials, Continuous Outcomes

A.2.1.1 Parallel Group

If a quick estimate is required for a sample size, we can use the quick results in each section of this chapter. It simply estimates the $(r+1)(Z_{1-\beta} + Z_{1-\alpha/2})^2$ part of the equation with their values from the Normal tables for the type I and type II errors.

For 90% power, equal allocation, 2-sided significance level the sample size per arm can be calculated by (Julious, 2004)

$$n_A = \frac{21\sigma^2}{d_S^2} \tag{A.25}$$

where d_S is the target difference between the treatments, σ^2 is the population variance. This equation comes from the following equation when $r = 1$:

$$n_A = \frac{10.5(r+1)\sigma^2}{rd_S^2}. \tag{A.26}$$

If there was not equal allocation, ($n_B = rn_A$), then this would be the quick result to use. The value 10.5 arises from the $(Z_{1-\beta} + Z_{1-\alpha/2})^2$ part of Equation 2.1, it is an estimate as oppose to the exact value when $Z_{1-\beta} = Z_{0.9} = 1.282$ and $Z_{1-\alpha/2} = Z_{0.975} = 1.96$.

For 80% power, equal allocation, 2-sided significance level (equal allocation) we can use

$$n_A = \frac{16\sigma^2}{d_S^2} \tag{A.27}$$

where d_S is the target difference between the treatments and σ^2 is the population variance. Again, the value of 16 arises from $r = 1$, $Z_{1-\beta} = Z_{0.8} = 0.842$ and $Z_{0.975} = 1.96$. A more accurate result would have 15.7, not 16, so this result would be a little more conservative.

Both the quick results give reasonable estimates of the sample size required and could be used as initial sample size estimates.

A.2.1.2 Crossover Trials

For a 90% power and a two-sided 5% type I error the sample size calculation is

$$n = \frac{21\sigma_w^2}{d_S^2} \tag{A.28}$$

Where σ_w^2 is the within-subject variance, d_s is the target difference and n is the total sample size.

A.2.2 Non-Inferiority Trials

The quick calculation at 90% power and a 2.5% type I error rate for a parallel group design is

$$n_A = \frac{10.5\sigma^2(r+1)}{r\Big((\mu_A - \mu_B) - d_{NI}\Big)^2} \tag{A.29}$$

The quick formula (at 90% power and 2.5% type I error rate) for a crossover design is (Julious, 2004)

$$n = \frac{21\sigma_w^2}{\Big((\mu_A - \mu_B) - d_{NI}\Big)^2} \tag{A.30}$$

A.2.3 Equivalence Trials

For a quick calculation with 90% power and type I error rate of 2.5% (Julious, 2004)

$$n_A = \frac{10.5\sigma^2(r+1)}{r\Big((\mu_A - \mu_B) - d_e\Big)^2} \tag{A.31}$$

This quick calculation only applies for differences which are close to d_e.

For a quick result with 90% power and a type I error rate of 2.5% (Julious, 2004)

$$n_A = \frac{13\sigma^2(r+1)}{rd_e^2} \tag{A.32}$$

This provides a direct estimate for the sample size per arm. Note that the coefficient in Equation A.32 is different to that in Equation A.31. This is because the Type II error is not allocated symmetrically when the population mean is non-zero.

A.3 Sample Size Tables

A.3.1 Non-Inferiority Trials

Tables A.1 (Julious and Campbell, 2012; Rothwell et al., 2018a) and A.2 show the various sample sizes needed for different standardised effect sizes and percentage mean differences for parallel group and crossover designs, respectively. As the target effect size increases the sample size decreases, as in previous tables. It is also clear that as the percentage mean difference increases, the required sample size increases.

	Percentage Mean Difference										
δ	-25%	-20%	-15%	-10%	-5%	0%	5%	10%	15%	20%	25%
0.05	5381	5839	6358	6949	7626	8407	9316	10379	11636	13136	14945
0.10	1346	1461	1590	1738	1908	2103	2330	2596	2910	3285	3737
0.15	599	650	708	773	849	935	1036	1155	1294	1461	1662
0.20	338	366	399	436	478	527	584	650	729	822	935
0.25	217	235	256	279	306	338	274	417	467	527	599
0.30	151	164	178	194	213	235	260	290	325	366	417
0.35	111	121	131	143	157	173	235	213	239	270	306
0.40	86	93	101	110	121	133	173	164	183	207	235
0.45	68	74	80	87	96	105	116	130	145	164	186
0.50	55	60	65	67	78	86	95	105	118	133	151

Table A.1: This table shows different sample sizes (n_A) for one arm of a **parallel group** design for a **non-inferiority** trial with an allocation of $r = 1$ (equal allocation) for various standardised non-inferiority limits ($\delta = \frac{d_{NI}}{\sigma}$). It shows the sample sizes for different true mean differences as a percentage of δ for a 90% power and type I error rate of 2.5%. These sample sizes are calculated from the non-central t-distribution.

	Percentage Mean Difference										
δ	-25%	-20%	-15%	-10%	-5%	0%	5%	10%	15%	20%	25%
0.05	5382	5840	6359	6949	7627	8408	9316	10380	11637	13137	14946
0.10	1347	1462	1591	1739	1909	2104	2331	2597	2911	3286	3738
0.15	600	651	709	774	850	936	1037	1156	1295	1462	1663
0.20	337	367	400	437	479	528	585	651	730	823	936
0.25	218	236	257	280	307	339	375	418	468	528	600
0.30	152	165	179	195	214	236	261	291	326	367	418
0.35	112	122	132	144	158	174	193	214	240	270	307
0.40	87	94	102	111	122	134	174	165	184	208	236
0.45	69	75	81	88	97	106	117	131	146	165	187
0.50	56	61	66	72	79	87	96	106	119	134	152

Table A.2: This table shows different total sample sizes (n) of a **crossover** design for a non-inferiority trial for various standardised **non-inferiority** limits ($\delta = \frac{d_{NI}}{\sigma}$). It shows the sample sizes for different true mean differences as a percentage of δ for a 90% power and type I error rate of 2.5%. These sample sizes are calculated from the non-central t-distribution.

A.4 Binary Outcomes

This section will discuss the sample size calculations required for the various trial objectives (specifically superiority, equivalence and non-inferiority) for trials with a single binary end-point. This end-point could be, for example, a pre-specified outcome of a condition (for example, pain or death). The main superiority trial design is included in this section, however the calculations for non-inferiority and equivalence trials are included in Appendix A.4. Each part of this section will discuss the hypotheses used for the sample size calculations along with a worked example. The examples will be completed both using the formulae and using sample size tables. For this outcome we have assumed that the reader is already familiar with Type I and Type II error rates, as discussed earlier in the chapter. One potential source of confusion for this section on binary outcomes is the alternating between π and p. To clarify the difference, π is used as the known population estimate for the absolute risk, whereas p is taken to be the sample estimate of π, which is estimated from the trial.

A.4.1 Superiority Trials

A.4.1.1 Parallel Group Trials

The parallel group superiority trial is investigating whether two different treatments/interventions are different in terms of their proportion of patients with a particular outcome.

Let π_A and π_B be the proportion of adverse events in groups A and B respectively.

The two hypotheses of interest would be

- H_0: There is no difference between the two treatment effects in terms of odds ratio ($\pi_A = \pi_B$)
- H_1: There is a difference between the two treatment effects in terms of odds ratio ($\pi_A \neq \pi_B$)

Consider Table A.3 (Julious and Campbell, 2012; Rothwell et al., 2018a), a summary table for a typical clinical trial with a binary end-point or outcome.

Let p_A be the proportion of responses in group A and p_B be the response in group B, with n_A and n_B be the total number of patients in groups A and B respectively such that $n = n_A + n_B$ is the total number of patients in the study and

$$\bar{p} = \frac{n_A p_A + n_A p_B}{n_A + n_B} \tag{A.33}$$

	Outcome		
Treatment	1	0	**Total**
A	p_A	$1 - p_A$	n_A
B	p_B	$1 - p_B$	n_B
Overall Response	$\bar{p}$	$1 - \bar{p}$	$n = n_A + n_B$

Table A.3: A summary table for a typical clinical trial with a binary outcome.

is the average response across the treatments.

There are two methods for getting a sample size calculation for this type of outcome, one of which is using the anticipated responses under just the alternative hypothesis and the other is using the responses under both the null and alternative hypotheses.

Method 1 using anticipated responses under the alternative hypothesis

This method is a relatively simple calculation to get an approximate sample size relatively quickly. With this method we get a sample size calculation of (Campbell et al., 1995; Julious, 2010b; Julious et al., 1999)

$$n_A = \frac{\Big(Z_{1-\beta} + Z_{1-\alpha/2}\Big)^2 \Big(\pi_A(1-\pi_A) + \pi_B(1-\pi_B)\Big)}{(\pi_A - \pi_B)^2} \tag{A.34}$$

This formula is for the case of equally sized groups (i.e. $n_A = n_B$)

Quick Results Method 1

Two formulae can be used to quickly calculate a sample size estimate for a superiority trial with a binary outcome. For a 90% power and type I error rate of 5% we get (Julious, 2010b)

$$n_A = \frac{5.25}{(\pi_A - \pi_B)^2} \tag{A.35}$$

For an 80% power and a two-sided type I error rate of 5% we can estimate a sample size using (Julious, 2010b)

$$n_A = \frac{4}{(\pi_A - \pi_B)^2}. \tag{A.36}$$

Both these quick methods will provide a conservative maximum sample size estimation. This is because the maximum sample size would be when $\bar{\pi} = 0.5$ (where $\bar{\pi} = \frac{\pi_A + pi_B}{2}$) (Julious, 2010b) and this is the response assumed in the derivation of the results. These results are conservative outside the rage of (0.3, 0.7) for $\bar{\pi}$ (Julious and Campbell, 2012). These results are for equal allocation between treatment groups, there are a number of other extensions to these results depending on the type of design, such as fixed allocation between groups (Campbell et al., 1995) or random allocation between groups (Ambrosius and Mahnken, 2010). These results also assume that there is a single end-point, multiple-end points are discussed elsewhere (Senn and Bretz, 2007; Sozu et al., 2010; Yeo and Qu, 2009).

Method 2 using anticipated responses under both the null and alternative hypotheses

The alternative method used for calculating the sample size uses the responses under the null and alternative hypotheses. This method arose because in method 1 the variances under each hypothesis are assumed to be equal, when in practice this is unlikely to be the case. Under the null we have $\pi_A = \pi_B$ and under the alternative we have $\pi_A \neq \pi_B$. The adjusted sample size calculation is as follows (Fleming, 1982)

$$n_A = \frac{\left(Z_{1-\alpha/2}\sqrt{\text{variance under null}} + Z_{1-\beta}\sqrt{\text{variance under alternative}}\right)^2}{(\pi_A - \pi_B)^2}. \tag{A.37}$$

This means the sample size can be estimated using the following formula (Julious and Campbell, 2012)

$$n_A = \frac{\left(Z_{1-\alpha/2}\sqrt{2\bar{\pi}_1(1-\bar{\pi}_1)} + Z_{1-\beta}\sqrt{\pi_A(1-\pi_A)+\pi_B(1-\pi_B)}\right)^2}{\left(\pi_A - \pi_B\right)^2} \tag{A.38}$$

where

$$\bar{\pi}_1 = \frac{\pi_A + \pi_B}{2} \tag{A.39}$$

.

One final consideration for parallel group trials with a binary end-point is whether a continuity correction needs to be implemented. This test was proposed when computing power was not as great as it is today. A continuity corrected chi-squared test was undertaken if the chi-squared assumptions did not hold, for example if there were small or zero cell counts. Today in the same circumstances a researcher may undertake a Fishers exact test. The formula for this sample size is used once the initial sample size has been calculated using one of the previous formulae, the corrected sample size is (Campbell et al., 1995)

$$n_{cc} = \frac{n_A}{4}\left[1 + \sqrt{1 + \frac{4}{n_A \times (\pi_A - \pi_B)}}\right]^2 \tag{A.40}$$

This result can also be used to provide an estimate for the sample size when the data will be analysed using a Fisher's exact test. This estimate is a little conservative (Julious and Campbell, 2012).

	Treatment B		
Treatment A	1	0	**Total**
1	n_{11}	n_{10}	n_{A1}
0	n_{01}	n_{00}	n_{A0}
	n_{B1}	n_{1B}	n

Table A.4: A summary table of a crossover trial. An example summary of a hypothetical crossover trial, where n_{xy} are the number of responses in cell xy. The end row and column give the total responses for each treatment. These overall responses are the numbers expected to be seen if it were a parallel group trial.

	Treatment B		
Treatment A	1	0	**Total**
1	λ_{11}	λ_{10}	p_A
0	λ_{01}	λ_{00}	$1 - p_A$
	p_B	$1 - p_B$	1

Table A.5: A summary table of a hypothetical crossover trial where $\lambda_{xy} = \frac{n_{xy}}{n}$, $p_A = \frac{n_{A1}}{n}$ and $p_B = \frac{n_{B1}}{n}$. The marginal totals are found in the end row and column and each cell shows the proportion of responses for each treatment combination or outcome.

A.4.1.2 Crossover Trials

Cross over trials with binary data are quite different to anything that has been discussed up to this point. However, they are very similar for each of the three trial types (superiority, equivalence and non-inferiority). There are two main summary measures used in these designs, which are odds ratios or the difference in proportions.

If we first consider Table A.3, for a crossover trial the only cells of interest are the discordant cells (the cells '01' and '10'). The reason for this is that the concordant cells in a superiority trial agree with the null hypothesis that there is no difference between the treatments, whereas the alternative hypothesis is stating that one treatment is preferable.

Table A.4 (Julious and Campbell, 2012; Rothwell et al., 2018a) can be re-written in terms of the proportions of responses, giving Table A.5.

The trial can then be summarised in an odds ratio using the discordant cells by $\Psi = \frac{\lambda_{10}}{\lambda_{01}}$. This can sometimes be difficult to interpret, therefore an approximate odds ratio can be gathered from the marginal totals of Table A.5 where (Royston, 1993).

$$\Psi = \frac{p_A(1 - p_B)}{p_B(1 - p_A)} \tag{A.41}$$

The discordant sample size can be estimated using the odds ratio, then from this value the total sample size required for the trial can be calculated. The discordant sample size, n_d, is deemed useful since it does not contain any unknown values, it is purely based on the odds ratio (Ψ), $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ (Connett et al., 1987; Fleiss and Levin, 1988; Julious et al., 1999; Royston, 1993; Schlesselman and Schneiderman, 1982)

$$n_d = \frac{Z_{1-\alpha/2}(\Psi + 1) + 2 \times Z_{1-\beta} \times \sqrt{\Psi})^2}{(\Psi - 1)^2} \tag{A.42}$$

The total sample size can then be estimated using (Connett et al., 1987; Julious et al., 1999; Royston, 1993)

$$N_{total} = \frac{n_d}{\lambda_{01} + \lambda_{10}} \tag{A.43}$$

A.4.2 Non-Inferiority Trials

A.4.2.1 Parallel Group Trials

As in Appendix A.1, recall that the null and alternative hypotheses for non-inferiority trials are (Julious, 2010b)

- H_0: A treatment is inferior in terms of the risk response ($\pi_A \geq \pi_B$)
- H_1: A treatment is non-inferior in terms of the risk response ($\pi_A < \pi_B$)

The two hypotheses for this type of trial can be rewritten in terms of a pre-specified clinical difference, d_{NI} (Chan, 2003; Chen et al., 2000; for the Evaluation of Medicinal Products, 2000).

$$H_0 : \pi_A - \pi_B \leq d_{NI}$$

$$H_1 : \pi_A - \pi_B > d_{NI}$$

where d_{NI} is the non-inferiority limit. This means that the null hypothesis, H_0, is that a given treatment is deemed inferior and the alternate hypothesis, H_1, states that the given treatment is not inferior. The setting of the non-inferiority limit is not easy, however it is defined as the largest difference that is clinically acceptable such that a larger difference than this would matter in clinical practice (for the Evaluation of Medicinal Products, 2000).

This type of study can be treated as a one-tailed study; therefore the α value we use is 0.025.

If $n_A = n_B$ then a direct estimate of the sample size is (Dunnett and Gent, 1977)

$$n_A = \frac{(\pi_A(1 - \pi_A) + \pi_B(1 - \pi_B))(Z_{1-\beta} + Z_{1-\alpha})^2}{\Big((\pi_A - \pi_B) - d\Big)^2} \tag{A.44}$$

where π_A is the assumed proportion of responses is expected in subjects on treatment A and π_B is the assumed proportion of responses expected on treatment B.

A quick method, using 90% power and two-sided significance level of 5%, is (Julious and Campbell, 2012)

$$n_A = \frac{5.25}{\Big((\pi_A - \pi_B) - d\Big)^2} \tag{A.45}$$

A quick method, using 80% power and a two-sided significance level of 5%, is (Julious and Campbell, 2012)

$$n_A = \frac{4}{\Big((\pi_A - \pi_B) - d\Big)^2} \tag{A.46}$$

Like for superiority trials earlier, both these quick methods will provide a conservative maximum sample size estimate.

A.4.2.2 Crossover Trials

A number of other articles have covered this issue (Julious, 2010b). The methods used are just extensions of the methods for superiority crossover trials and parallel group non-inferiority trials. It is recommended to use the same methodologies as for the parallel group trials to form an estimate of the total sample size for a non-inferiority crossover trial. This is done by taking the sample size per arm to be the total sample size.

A.4.3 Equivalence Trials

A.4.3.1 Parallel Group Trials

As discussed in Appendix A.1, there are occasions when it is not necessary to prove superiority. Sometimes it is preferable to show that two different treatments are equivalent. For example, comparing a drug with a cream when the cream is cheaper, it may be of clinical interest to show that the cream is as effective as the drug. Another example would be comparing a surgical technique with intensive physiotherapy.

General Case

The two hypotheses for this trial are (Julious and Campbell, 2012)

- H_0: The two treatments are different in terms of their risk difference. ($\pi_A \neq \pi_B$)
- H_1: The two treatments are not different in terms of their risk difference. ($\pi_A = \pi_B$)

Normally these hypotheses are written not in terms of the risk difference, but in terms of the clinical difference, d_e. They become

- H_0: $\pi_A - \pi_B \leq -d_e$ or $\pi_A - \pi_B \geq +d_e$
- H_1: $-d_e < \pi_A - \pi_B < +d_e$

These hypotheses, as with the continuous outcome equivalence and non-inferiority trials, are intersection-union tests. With these tests, as previous discussed, each component of the null hypothesis is tested at the level α (Berger and Hsu, 1996; Julious, 2004, 2010b). The sample size for a parallel group equivalence trial is (Julious and Campbell, 2012)

$$n_A = \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2 \Big(\pi_A(1-\pi_A) + \pi_B(1-\pi_B)\Big)}{(|\pi_A - \pi_B| - d_e)^2} \tag{A.47}$$

Special Case (No treatment difference)

If there is no anticipated treatment difference ($\pi_A - \pi_B = 0$), a direct sample size estimate can be obtained from the following formula (Julious and Campbell, 2012)

$$n_A = \frac{2 \times (Z_{1-\beta} + Z_{1-\alpha/2})^2 \times \bar{\pi}(1-\bar{\pi})}{d_e^2} \tag{A.48}$$

A.4.3.2 Crossover Trials

A number of other articles have covered this issue (Julious, 2010b) the methods used are just extensions of the methods for superiority crossover trials and parallel group non-inferiority trials. It is recommended that the same methodologies are used as for the parallel group trials to form an estimate of the total sample size for a non-inferiority crossover trial. This is done by taking the sample size per arm to be the total sample size.

A.5 Cluster Trials

Cluster parallel group trials for both continuous and binary data need to be considered separately, as a design effect needs to be included in the sample size calculation (ICC). Cluster trials occur when it is not always possible or feasible to randomise patients at an individual level only. Cluster randomisation is therefore undertaken when, instead of the individual patients being randomised to a treatment or therapy, the entire unit (for example the hospital, GP surgery or school) is randomised to the interventions. The reason that cluster trials may be used is to avoid contamination of the treatment group. This is particularly important, for instance with educational studies it may not be realistic to give the intervention to one subject without others at the same unit (school, surgery) being exposed as well. A vital consideration for a clustered-randomised trial design is the intra-cluster correlation coefficient (ICC). This quantifies the extent of similarity between individuals within a cluster, that is, the more similar individuals are in a cluster, the higher the ICC will be. The ICC can be calculated using

$$\zeta = \frac{\sigma^2}{\sigma_B^2 + \sigma_w^2} \tag{A.49}$$

where σ^2 is the overall response variance, σ_B^2 is the between-cluster component of the variance and σ_w^2 is the within-cluster component of the variance. Note that $\sigma^2 = \sigma_B^2 + \sigma_w^2$ and $\sigma_w^2 = \sigma^2(1 - \zeta)$.

A.5.1 Parallel Group Superiority Trials

The sample size calculation for superiority trials with a clustered parallel group design and a continuous end-point is as follows. n_A is the number of subjects per intervention.

$$n_A = \frac{2\sigma^2\Big(Z_{1-\alpha/2} + Z_{1-\beta}\Big)^2\Big[1 + (m-1)\zeta\Big]}{md^2} \tag{A.50}$$

This equation is similar to that from Section 2.3.1 but also includes ζ which is the ICC, m which is the average sample size per group or cluster. This equation is again an extension of the parallel group superiority case, with an inflation factor of $1 + (m-1)\zeta$. Another variation of this sample size calculation is to determine the number of clusters, k, for each intervention.

$$k = \frac{2\sigma^2\Big(Z_{1-\alpha/2} + Z_{1-\beta}\Big)^2\Big[1 + (m-1)\zeta\Big]}{md^2} \tag{A.51}$$

These methods can be extended to non-inferiority and equivalence designs.

B. Chapter 4

The extraction variables used in the HTA review in chapter 4 have been included in this appendix. The list is quite long and whilst it is useful to display the variables, it is not vital for them to be included in the main body of the thesis. There are also two plots which are interesting but not vital to the results displayed in section B.2.

B.1 Extraction Variables

A full list of the variables which were extracted from the reports is provided here.

- Study ID
- Study acronym (if provided)
- Full study title
- Lead Author
- Corresponding Author
- Publication Year, Volume, Issue
- ISRCTN
- Trial type and design
- is the trial randomised?
- is the trial multicentre?
- What clinical area is the trial investigating?
  - Full list in section 4.2.4.2
- number of arms
- trial population
- setting

  - Hospital
  - GP
  - Mixed
  - Community
  - Primary or Secondary Care
  - Other

- primary end point and measure type

- intervention type

  - Drug
  - Therapy
  - Surgical
  - Education
  - Complex
  - Other

- Control type

  - Missing
  - Active
  - Placebo
  - Not Applicable

- Target (unadjusted) and final target sample size

- Achieved and evaluable sample size

- Target power and significance level

- Target difference and standard deviation (if provided)

- Target effect size

- Elicitation method

  - Previous research
  - Pilot
  - Systematic review
  - Cochrane review
  - Expert consensus

  - Meta-analysis
  - Other
  - No mention

- Is the target effect size the MCID?
- If previous study used to elicit target, what was result observed in study?
- DELTA categories of elicitation
  - Anchor
  - Distribution
  - Health economics
  - Opinion-seeking
  - Pilot
  - Review of evidence
  - SES (standardised effect size [Cohen] (Cohen, 1988)
  - Mixed
  - No mention
  - Other
- Observed treatment effects in each arm
- Observed effect size
- Observed effect size type
  - Mean difference
  - Relative Risk
  - Odds ratio
  - Hazard ratio
  - Difference in proportions
  - Regression coefficient
  - difference in score
  - General Linear Model coefficient
  - Other
- *P*-value
- Is the *P*-value significant?

- Lower and upper 95% confidence interval boundaries

There were sections provided for free text about the following areas

- trial design,
- early comments from initial reading,
- trial population,
- clinical area,
- intervention type,
- primary end-point,
- sample size,
- target difference elicitation,
- treatment difference,
- observed effect size,
- if the *P*-value was non-signficant for a MCID, was MCID re-evaluated?
- further comments.

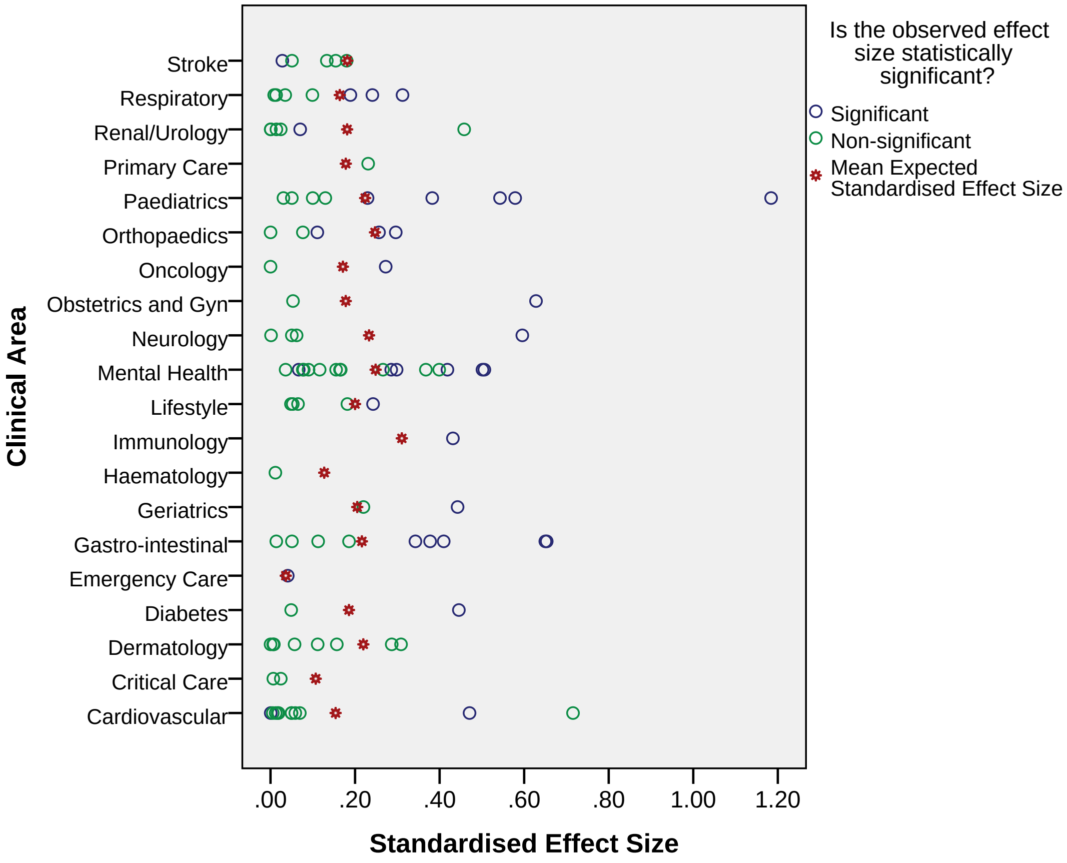


Figure B.1: The **individual observed** effect sizes with the **mean target** effect size for each clinical area.

B.2 Extra Plots

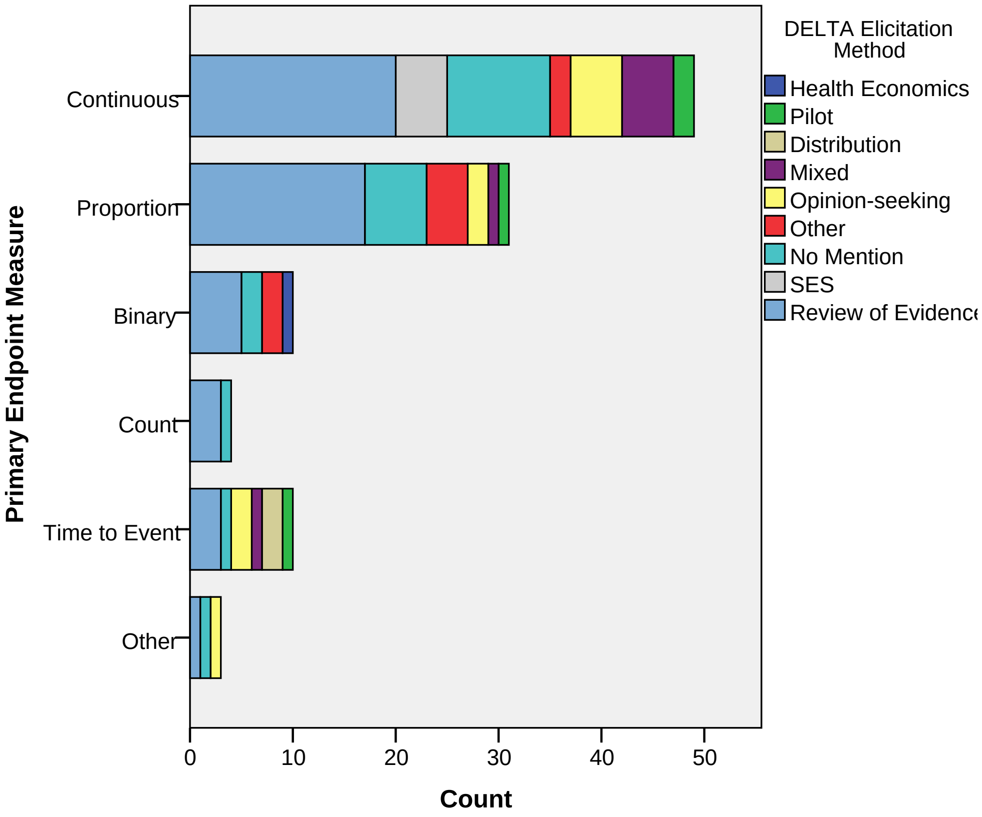


Figure B.2: The frequency of each primary endpoint type by elicitation categories.

C. Chapter 5

The research presented in this appendix was initially completed during the first year of the PhD. Since then, the research question was refined and this information was surplus to requirements. It did not enhance the thesis, yet it is representative of the quantity of work conducted during the course of this PhD.

C.1 Overview of Methods of Adjustment

There are a number of documented methods to adjust for regression to the mean which occur either in the design stage of the trial or the analysis stage (Barnett et al., 2005). There does not appear to be one that is superior to the other methods, with all the trials read for this review demonstrating use of a wide variety of methods. The use of a control group appears to be the most logical method, as it allows the trialists to observe how much of the treatment effect can be attributed to regression to the mean and the natural course of the illness or problem. However, there are a number of situations where having a control group is not deemed ethical, so other methods have to be applied. If a control group is used, this will help to control the within-study regression to the mean, however it will not adjust for the between-study regression to the mean. The between-study regression to the mean is the focus of this PhD. There are more methods for adjusting for this phenomenon in the analysis part of the trial compared to the design.

C.1.1 Design Stage Methods

The inclusion of a control group (or more specifically a non-intervention group) seems to be the simplest method used to adjust for regression to the mean. This is set out in the design of the trial and, subject to no ethical complications, works fairly effectively. The theory behind it is that the control group are representative of the intervention group if the intervention group had not received treatment. Any change in the control group is therefore solely attributed to natural fluctuation and regression to the mean (Greineder et al., 1999; Whitney and Von Korff, 1992; Yudkin and Stratton, 1996). The average change observed in the control group can then be subtracted from the average observed change to determine the true treatment effect (Cummings et al., 2004; Lipman et al., 2006; Maltby et al., 2005; Martinez-Yelamos et al., 2006; Moebus et al., 2006; Osterbrand et al., 2007; Rose et al., 1980; Sanders-van Wijk et al., 2014).

Since the main focus of this PhD is randomised controlled trials, there will always be a control group present. Therefore, the focus will be on analysis stage methods as oppose to design stage methods. The design stage methods work well for trials with before-after measurements, however they do not seem to be applicable to the context of this research.

C.1.2 Analysis Stage Methods

As previously discussed, there are two points at which an adjustment can be made, the design and analysis stages of the clinical trial. Sometimes it is not practical to make an adjustment during the design stage; in these situations one can make the necessary adjustments in the analysis stage. Consider a trial with measurements made at two time points and an experimental and placebo group. One method is to use a regression equation for the placebo patients to get expected untreated value at second time point conditional on baseline values, then subtract the expected value from the observed value to adjust for RTM (Lubsen et al., 2007). This method is similar to the adjustments made in regression techniques such as linear regression or analysis of variance. There are regression based methods like analysis of covariance (ANCOVA), analysis of variance (ANOVA) and multiple linear regressions which can also be used. These methods can be implemented a number of different ways but ultimately do the same thing. The idea behind these methods is to control for baseline differences between the treatment groups by including the baseline values as a covariate or variable in the analysis (Barnett et al., 2005; Bernstein et al., 2009; Crawford et al., 2012; Finney, 2008; Qouta et al., 2012; Twisk and de Vente, 2008; Vickers and Altman, 2001; Whitney and Von Korff, 1992). Twisk puts forward another method since ANCOVA has problems and other autoregressive techniques assume all groups are equal at baseline which is not always the case. Twisk describes residual change combination method and analysis of covariance combination method in general detail, but the methods are not applied to an example.

In a similar vein, these differences in baseline measurements are controlled in the method of multiple regression (Allison et al., 2009; Bernstein et al., 2009; SW, 1984; Victora et al., 1998; White et al., 1997). Multiple regression can be performed on both the treatment and control groups; the difference between the two equations is the effect of switching between treatment and control groups. The regression to the mean is controlled by the regression equations (Osterbrand et al., 2007).

The adjustment made to results is dependent on the type of outcome variable, whether it is continuous or binary. For continuous outcome variables there are three methods traditionally used, namely analysis of covariance (ANCOVA), the change from baseline to follow-up measurements could be used or a method called residual change which is not commonly used. ANCOVA is similar to a linear regression analysis, however the follow-up measurements are treated as the outcome variable and the baseline measurement is selected as a covariate. This provides an adjustment for the differences in baseline between the groups (Twisk and Proper, 2004). The advice given in this article is to use ANCOVA for continuous variables as the coefficients are easier to interpret than in the residual change method. For binary or dichotomous outcome variables, the preferred method to adjust for the statistical artefact is multinomial logistic regression adjusting for baseline effects. One is able to use logistic ANCOVA; however the interpretation of this is more complex and does not provide as much information as logistic regression.

C.1.3 Simulation and Graphical Methods

There are a number of simulations based methods which are used to adjust for regression to the mean. The methods are fairly complex though one is more intuitive than the others. The graphical methods vary in terms of their applicability to the adjustment of regression to the mean. The three main methods found in this literature review seem to focus on showing that regression to the mean is occurring, as opposed to being a method of adjustment. Due to this, they are unlikely to be useful in the context of this PhD.

C.2 Detailed Methods of Adjustment

This section will briefly discuss a method which is used to adjust for regression to the mean and would be applicable to this research area. A number of other methods were extracted and discussed in this literature review; however they did not appear to be applicable to the research problem.

C.2.1 Shepard and Finison Method

A method which was described in a paper by Shepard and Finison but not used in any trials in the review is based on the scenario where a potential participant to the trial must have an outcome measurement higher than a pre-defined threshold at multiple visits to the clinic (Shepard and Finison, 1983). This method appears to be analogous to the aim for this research, as it depends on the initial outcome measurement being higher than a specified value, such as having encouraging results at an initial trial. The method defines a model which aims to provide an average outcome measurement, having adjusted for regression to the mean. The model comes from Equation C.1

$$\hat{\hat{C}} = (\bar{x}_{full} - \bar{y}_{full}) - [(1 - \hat{G})(\bar{x}_{full} - \bar{x}_{pot})] \tag{C.1}$$

Where $\hat{\hat{C}}$ is the adjusted outcome variable,$\bar{x}_{full}$ and $\bar{y}_{full}$ are the average outcome measurements for baseline and follow-up time points, respectively. The full subscript denotes that these values are calculated using all the potential participants who completed the trial. The variable $\bar{x}_{pot}$ denotes the mean baseline measurement for all potential participants, so any participants who meet the entry criteria for the trial at the screening test. $\hat{G}$ is the coefficient of reliability for that measurement type, for example blood pressure. The interpretation of this value is dependent on the context of the intervention. In the paper they use the example of aiming to reduce blood pressure, so a positive value for $\hat{\hat{C}}$ would indicate that the intervention is successful and the average blood pressure has been reduced. Another extensive formula-based method which is used as the basis for other methods is by Gardner and Heady (Gardner and Heady, 1973). This method can be used to predict the expected regression effect, therefore providing a method to estimate the treatment effect on its own without the inclusion of the regression effect.

C.2.2 Other causes of regression to the mean

Another instance which can cause regression to the mean is where there are differences at the baseline value between the intervention and control groups. In theory, this should not occur when the patients are randomised because all the participants come from the same population. However, variations can occur by chance, and it is this difference in baseline measurement which can result in observing a response in both groups irrespective of which one was the intervention group (Twisk and de Vente, 2008). Regression to the mean commonly occurs when there are ceiling- or floor-effects in the data, for example a test score in education will be limited by the minimum mark of 0 and the maximum mark available. The reason that this effect occurs in these situations is simply because if a student gets an extremely high (or low) score then they only have one direction in which to go for the retest, so will normally get a lower (or higher) score upon the retest. This is an example of regression towards the mean as the subjects of interest with the extreme scores will likely regress towards the mean score of the population taking the test. It has also been discussed that regression to the mean will occur when there is a negative correlation between the baseline measurement and the change in measurement

(Cannito et al., 2012; Rocconi and Ethington, 2009). Negative correlation would occur if the baseline measurements were particularly high in a group. For example, Cannito et al. discovered a high negative correlation between pre-treatment baseline and change in Sentence Intelligibility Test (SIT) scores in patients with Parkinsons disease (Cannito et al., 2012).

This meant that the patients who scored higher pre-treatment showed less change than those who scored lower. Knowing this could be useful in tackling the issue and realising before the full analysis begins in a trial that there may be some other effects occurring. However, one problem with using this test is that there is an inherent negative correlation between the baseline value and the change, so this may not be the most appropriate solution (Campbell, 1990).

Missing data is another potential cause of regression to the mean. When the data is not missing at random, so there is the possibility that there is an association between the specific variables that are missing, this phenomenon is more likely to occur (Suissa, 2008). A number of other factors contribute to regression to the mean, such as having small sample sizes or investigating a disease which is cyclic in nature (Enck and Klosterhalfen, 2005).

C.2.3 What clinical areas does regression to the mean occur?

After looking through the literature to establish if there were any particular disease areas or intervention types for which regression to the mean was adjusted for, it became apparent that there was no obvious link between adjustment and disease or intervention type. Since it is most commonly discussed as affecting blood test measurements, a lot of cardiology interventions made various adjustments for it which are described in further detail later in the review (Asmar et al., 2001; Lubsen et al., 2007; Okin et al., 2001; Palmieri et al., 1999; Sanders-van Wijk et al., 2014; Stewart et al., 2000; White et al., 1997).

Other common disease areas were: diabetes (Manley et al., 2000; Persu et al., 2014); oncology (Sagar, 2008); pulmonary conditions (van Schayck et al., 1995); pain (Bjorkedal and Flaten, 2011; Mbizvo et al., 2015; van Schayck et al., 1995; White et al., 2004); osteoporosis (Chapurlat et al., 2001; Cummings et al., 2000); immunology (Lin and Hughes, 1995, 1996) and mental health (Allison et al., 2009; Lipman et al., 2006; McCall et al., 2011; Qouta et al., 2012). A number of different treatments were used in these specific randomised controlled trials, which varied from medical drugs, surgery, homeopathic remedies, and holistic approaches like acupuncture or lifestyle interventions such as diet and education.

C.2.3.1 In what areas is it adjusted for?

There does not appear to be a specific area or set of areas where regression to the mean is regularly adjusted for. It is only mentioned in the areas discussed above, and even then it is not always adjusted for in those areas. This was discovered during the literature searching, with a large number of articles mentioning it as a justification for why the placebo group in that particular trial also exhibited a treatment effect (Bjorkedal and Flaten, 2011; Martinez-Yelamos et al., 2006; Ryvicker et al., 2011).

Some papers adjusted for the effect but did not go into detail as to which method they used (Manley et al., 2000; Palmieri et al., 1999) whilst some interventions, by their nature, make it infeasible to have a control group such as educational interventions James (1973). This follows because once the education has been provided it cannot be unlearned and it could cause the participant to alter their behaviour.

C.2.4 Other Design Stage Methods of Adjustment

A consideration when adjusting for regression to the mean is the natural history of the disease or illness, or natural fluctuations in measurements of interest. This can be adjusted for in a similar way as including a control group, which is to include a no intervention arm. This is useful if there is already a control group which receives current or standard treatment, and an experimental group which is receiving the new treatment of interest (Conboy et al., 2006). Randomisation should mean that baseline variations are approximately equally split between the intervention and control groups; therefore the change observed in the control group can be used as an estimate for the effect of regression to the mean in both groups (Barnett et al., 2005).

A method used in one trial was to give patients a run-in period of no treatment, then a second period of usual treatment before randomisation to the treatment or placebo groups. The effect was analysed by taking the baseline responses from the usual treatment period, having had a wash out period. These baseline responses were then compared with the previous deterioration in the no-treatment period to give an estimate of regression to the mean (Burge et al., 2003). A similar method used in crossover trials is to give the patients a run-in period of no treatment, and then randomised to either treatment or no treatment groups. After that time period the two groups swap treatment. Regression to the mean is assessed by comparing the mean baseline values prior to randomisation with the end measurements of the no-treatment groups (Asmar et al., 2001). Variations of this method have been documented, though they are trial and intervention specific (Stewart et al., 2000).

Alternatively, one could design a trial which aims to take multiple measurements at baseline, then base participant selection on the average of these measurements. For example, if participation in a clinical trial is based on individuals presenting with an outcome of interest higher (or lower) than a pre-specified threshold value or cut-off point, taking the average of a number of measurements would reduce the chance of regression to the mean by adjusting for extreme values or measurements for each individual (Barnett et al., 2005; Denke and Frantz, 1993; Lin and Hughes, 1995; McDonald, 2003; Yudkin and Stratton, 1996). This method aims to get a better, more accurate estimate of the baseline values (Cannito et al., 2012; Chapurlat et al., 2001). This method is fairly common, for example in a hypertension trial by Okin et al, multiple measurements were taken prior to treatment and eligible patients were recorded as being hypertensive at all of these measurement time points (Okin et al., 2001).

In a similar vein, one could assess patients blood pressure or cholesterol levels to determine if they are potentially eligible for recruitment, then instead of using this measurement as a baseline value, one could take another measurement to use as the baseline measurement (Yudkin and Stratton, 1996).

A less commonly used method is to combine the baseline measurements with other risk factors for the outcome of interest, however this is not frequent as it could be rather expensive to implement since it could involve taking many more measurements or using more resources to extract the measurements (Chapurlat et al., 2001). For example, if the baseline measurement is a simple blood biomarker then that would be relatively inexpensive compared with the addition of minor surgery or further blood testing. One could compare the change in each measurement of interest between time points. The method used to adjust for the regression to the mean effect is described in a paper by Irwig; however this paper is one of those that could not be located (Irwig et al., 1990).

C.2.5 Detailed Methods of Adjustment

This section will detail a few different methods used for adjusting for regression to the mean. The purpose is to get a better idea of how a few of these methods work.

C.2.5.1 Value-Added Method

One method used is called the value added method (Heimendinger and Laird, 1983; Walker et al., 1996). The way this method works is to estimate the outcome measurement if there were no intervention, then use this to calculate how much value the intervention added. This is commonly seen in interventions where natural growth needs to be adjusted for as well as factors like regression to the mean. For example, if the intervention was investigating how much a supplement aided childrens growth, one would have to take how much the child would have grown naturally into account.

This leads onto a formula developed by Lord in 1956 which simplifies regression to the mean (Lord, 1956). The formula is given in Equation C.2.

$$\hat{Z}_2 = E(Z_2|Z_1) = \rho_{12}Z_1 \tag{C.2}$$

Where $\hat{Z}_2$ is the estimate of the outcome at time t_2, Z_1 and Z_2 are the observed outcomes at times and respectively, and ρ_{12} is the correlation between Z_1 and Z_2. This is under the assumption that both Z_1 and Z_2 are standardised outcomes or scores at times t_1 and t_2, and that the measurements were taken on one participant. By standardised, it indicates that these variables have a mean of zero and a standard deviation of 1, they are also known as Z-scores. This formula provides an estimate of the outcome measure at time point 2, having adjusted for the correlation between the baseline and subsequent measurements as well as the baseline measurement itself.

Another extensive formula-based method which is used as the basis for other methods is by Gardner and Heady (Gardner and Heady, 1973). This method can be used to predict the expected regression effect, therefore providing a method to estimate the treatment effect on its own without the inclusion of the regression effect.

C.2.6 Likelihood-based Methods

An alternative adjustment which has been used in the past is constructed using the Likelihood-based method (Lin and Hughes, 1996; Senn et al., 1985). It builds on the work by James (James, 1973) and adjusts for the fact that he has not taken the effect that the truncation point will have on the variance of the population into account in his model. He has also not acknowledged that there will be a covariance effect between the two values for the different time points or measurements.

C.2.6.1 Davis Method

A method which is also used is based on an equation by Davis which adjusts for regression to the mean (Davis, 1976; van Schayck et al., 1995). There are a series of equations involved in this adjustment. Let us assume that x_i is the i-th measurement of the outcome of interest, and that it is normally distributed with mean μ and standard deviation σ, ρ_{ij} is the correlation between measurements i and j. The intermediate equations have been omitted but the final estimate of the regression is given by Equation C.3.

$$RTM_{effect} = c_1\sigma(1 - \rho_{12}) \tag{C.3}$$

Where $c_1 = \phi(a_1)/[1 - \phi(a_1)]$, a_1 is the cut-off point, σ is the population standard deviation, ρ_{12} is the correlation between measurements 1 and 2, and μ is the observed mean value. Estimates of these variables can be gathered from the study data or using external estimates. The second method discussed in this paper is based on the results discussed in James article (James, 1973) which consists of equations to estimate the values for this equation using the current data as opposed to using external estimates.

C.2.6.2 Method of Moments

James describes a method called the method of moments (James, 1973) which has been used as the foundation for a number of other methods. It is rather complicated but is based around a series of formulae leading to a calculation to estimate the proportion of reduction in a variable (like cholesterol level) which is due to solely regression to the mean. This formula is given in Equation C.4.

$$\text{Proportion reduction} = 1 - \rho \tag{C.4}$$

This leads to being able to calculate the proportion of reduction which is based on regression as a fraction of the total proportion of reduction (Equation C.5.

$$\frac{prop.reduction_{regression}}{total.prop_{reduction}} = \frac{1-\rho}{1-\gamma\rho} \tag{C.5}$$

The details of this method can be found in the article, along with a worked example (James, 1973).

C.2.6.3 Other Methods

A method discussed in a paper on substance-use disorders is based on methods by Campbell and Kenny (Campbell and Kenny, 1999; Finney, 2008). They discuss that the expected regression to the mean effect without any intervention effect can be calculated using Equation C.6.

$$\text{effect} = b_{xy}(x - m_x) + m_y \tag{C.6}$$

Where m_x is the mean of the pre-intervention variable, m_y is the mean of the post-intervention variable and b_{xy} is the regression coefficient, x is the pre-treatment score for an individual or the average score for a group of individuals. One paper used a rather unconventional method to make an adjustment for the effect, which was to remove the two patients which were thought to be causing a skew in the data, so the two patients with the highest scores of a particular biomarker

This is the only mention of this method and as a statistician we are advised not to remove data without a good reason. It does not appear that this method is widely adopted, perhaps because it is not deemed to be sound methodology. One can adjust the initial measurement taken once a negative correlation has been established between the initial measurement and the follow-up measurement. Rocconi and Ethington have described an adjustment formula which alters the initial measurement, and then traditional methods of analysis can be performed on this new adjusted value. This method does not eliminate RTM but does reduce it (Rocconi and Ethington, 2009).

C.2.6.4 Simulation Based Methods

There are a number of simulation based methods which aim to adjust for regression to the mean (Bajard et al., 2009; Farlow et al., 2005; Krause and Pinheiro, 2007; Ross, 1995). They are complicated but one can see how they would be useful. The easiest method to understand is based on simulating to get a reference distribution of P-values (Krause and Pinheiro, 2007). he simulations are based on having a first measurement,x_0 , which is greater than a threshold value for entry to the trial, along with a second measurement, x_1, which is not guaranteed to be greater than the threshold value. This method emulates the outcomes of moving from Phase II to Phase III trials. These values are simulated many times to produce two distributions for the two time points. The mean for the x_0 distribution will be greater than the mean in the x_1 distribution. The method used in this paper determines a model for the treatment effect as a whole (including regression to the mean and other statistical artefacts) to describe the response of interest as best as possible. Using this model, a null model is created which removes treatment effect. The same believed cause of regression to the mean in the trial itself is then applied to this model and the simulated data (for example, the fact that selection is based on a threshold value could have caused regression to the mean). These simulated null data sets are they analysed in exactly the same way as the observed data were, which produce simulated test statistics and, by extension, a distribution of associated P-values. This is the reference distribution. The observed P-value is compared with this reference distribution. The resulting adjusted P-value is calculated as the percentage of simulated P-values which are equal or less than the observed p-value. This approach appears to be fairly logical, yet not widely adopted.

C.2.6.5 Graphical Methods

A useful way to see if regression to the mean is occurring is to look at plots or graphs. These provide a visual interpretation of the issue which can be illuminating

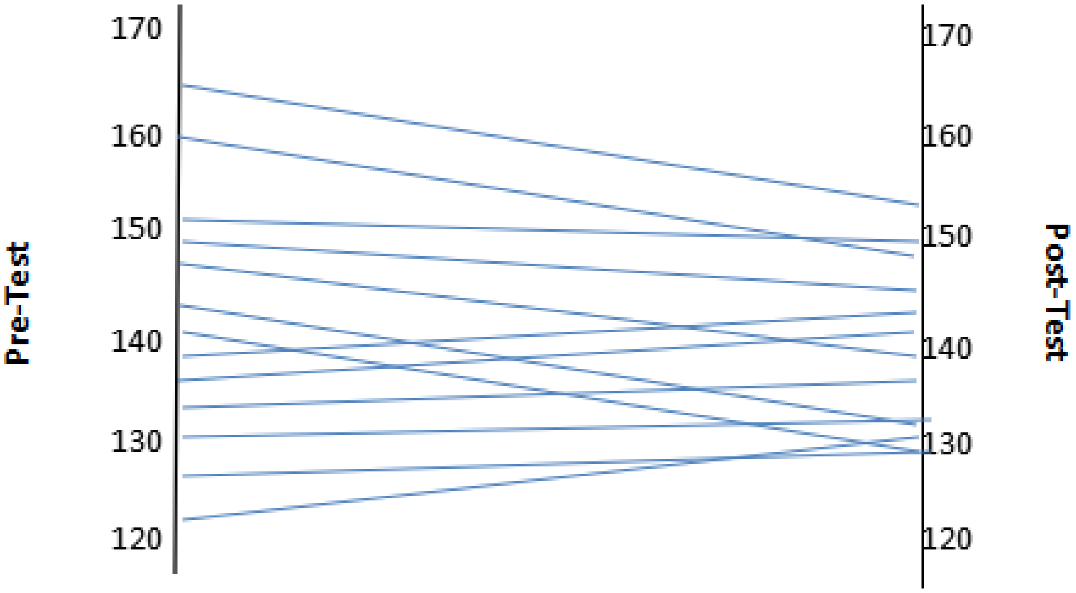


Figure C.1: An example of a Galton Squeeze diagram for a clinical trial testing the effect of a treatment.

for the reader. One type of plot is called a Galton Squeeze diagram, an example of which is shown in Figure C.1.

The Galton Squeeze diagram is plotted using the standardised difference between each individual observation and the overall mean value for that variable or measurement prior to randomisation, and also the difference at the end of the treatment period needs to be calculated. If the standardised difference before treatment is plotted against the standardised difference after treatment for each participant, one is able to assess whether regression to the mean is present. If the plot exhibits a funnel shape, regression to the mean is occurring (GL Burrell, 2010; McCall et al., 2011). Another graphical method is to plot the baseline measurement of an outcome variable against the mean change observed for each individual. If the plot exhibits a negative trend then regression to the mean is occurring (Bhorade et al., 2009; Whitney and Von Korff, 1992).

One final graphical method to illustrate regression to the mean is to calculate the difference between each time point for each individual, and use this to calculate the difference between the pre-treatment measurements and the post-treatment measurements. This is plotted against the mean change overall for the group. If there is a negative y-intercept then there is no change due to treatment, if there is a negative y-axis value when there is no treatment effect, then regression to the mean is evident. This method is based on having 4 different time points; 2 before treatment and two after the treatment (Ederer, 1972).

D. Chapter 6

This appendix provides extra plots which arose from the simulations conducted in chapter 6. Due to the quantity of plots and the fact that they are relatively similar in nature, it was decided to include them in an appendix for reference.

D.1 Initial Simulations

This section considers the same method of simulations as used in Section 6.2, however for these simulations the power levels and effect sizes are varied. First, the observed difference is fixed to be 10, with the same standard deviation of 50 and same significance level of 5%. If the power is changed to be 85%, 90%, 95% and 99%, it would be interesting to know what these distributions look like.

D.1.1 Varying Power

The plots for various powers can be seen in Figures 6.6 to 6.7. It can be seen in these Figures that as the power increases, the truncation point of the Normal distribution moves further from the mean. When the power is 99%, the truncated Normal distribution looks extremely similar to the standard Normal Distribution. This indicates that the effect of the bias decreases as the power increases also.

D.1.2 Varying Effect Size

A common effect size called Cohen‘s d (Cohen, 1973) (Cohen, 1973) can be calculated as

$$d = \left| \frac{x_1 - x_2}{s} \right| \tag{D.1}$$

Where x_1 and x_2 are the means for groups 1 and 2, and s is the pooled standard deviation for the groups given by

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{D.2}$$

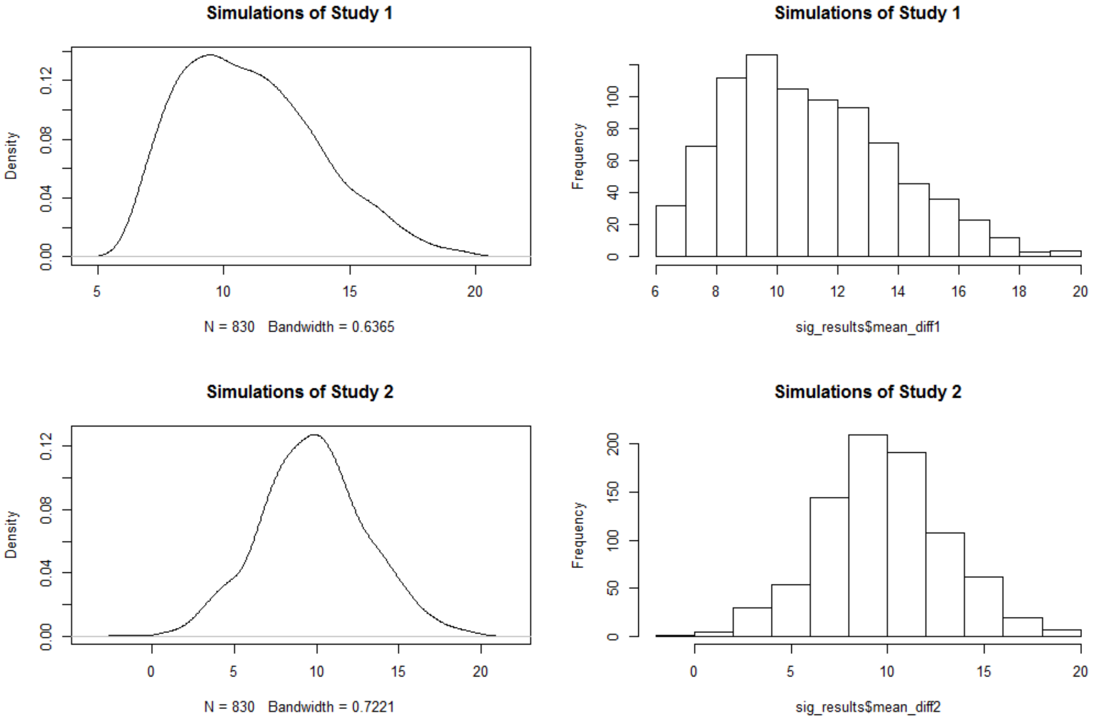


Figure D.1: The distributions for trials with 85% power. The truncated Normal distribution for study 1 appears to have a higher truncation point as the power increases.

where s_1^2 and s_2^2 are the sample variances for each group, n_1 and n_2 are the sample sizes in each group. Sometimes this formula for δ can be calculated without the -2 in the denominator, as other authors may prefer (McGrath and Meyer, 2006).

The standard levels of Cohen‘s d are 0.2 (small effect), 0.5 (moderate effect) and 0.8 (large effect). From the equation above it is clear to see that the $\bar{x}_1 - \bar{x}_2$ part of the equation is similar to the d_S part of the sample size equations from chapter 2. Also in that chapter there was mention of being able to substitute the standardised difference or effect size into the equation, effectively putting Cohens d into the sample size formula. Since these two d values are linked, it follows that they are both extremely sensitive in terms of sample size calculation. If we take the population standard deviation to be 50 for both groups, this results in being approximately 50 if the sample size in each group is equal. The effect size used in the simulations up to this point can be calculated as

$$d = \frac{10}{50} = 0.2$$

This effect size is classified as a small effect; however it is interesting to simulate what would happen if the effect size was larger. Consider now different values of the effect size, with $\bar{x}_1 = 0$ and $\bar{x}_2$ ranging from 20 to 40 and $s_1 = 50$, $s_2 = 50$ for 80% and 90% power. The corresponding effect sizes being simulated are 0.3,

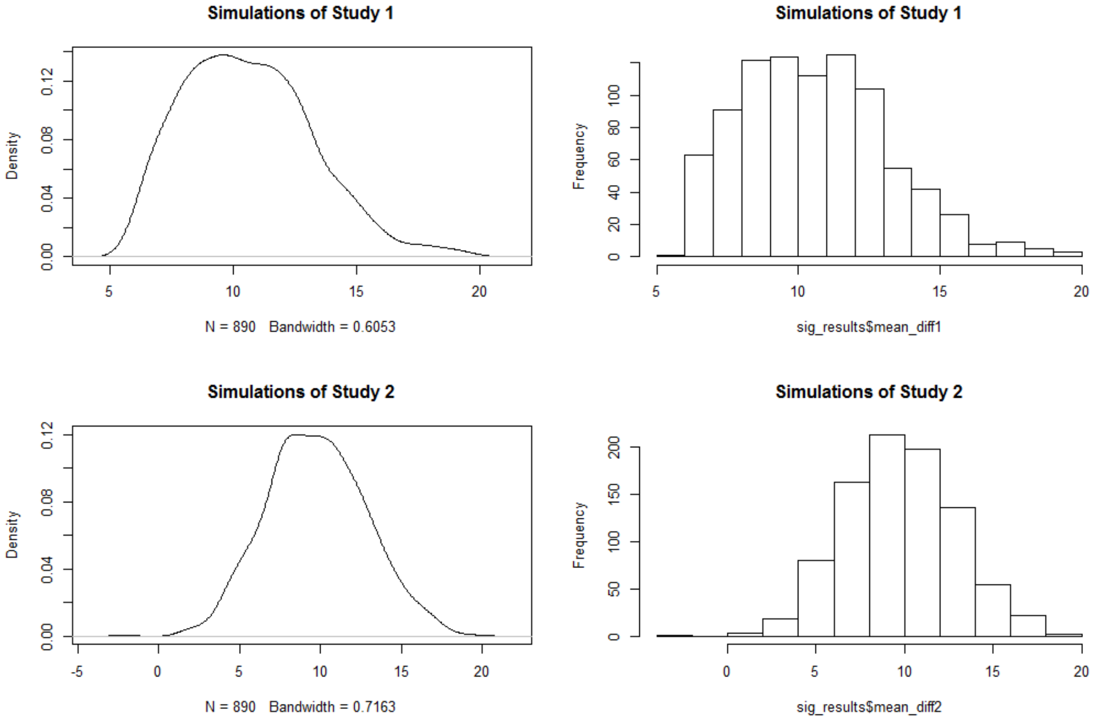


Figure D.2: The distributions for trials with 90% power. The truncated Normal distribution for study 1 appears to have a higher truncation point as the power increases.

0.4, 0.5, 0.6 and 0.8. These effect sizes have been included to show the change in distributions as well as including the standard levels of effect sizes for reference. The histograms of the distributions and the mean differences for the simulations shall be displayed in Figure D.5 for 80% power. As in Section 6.2, it is still noticeable that the difference in means decreases as the power increases. Due to this already being shown, the focus shall now be on 80% power. As the power remains constant it will allow further investigation of the impact the different effect sizes have on the means.

Figures D.5 to D.9 show the variation in the truncation point for changing effect sizes.

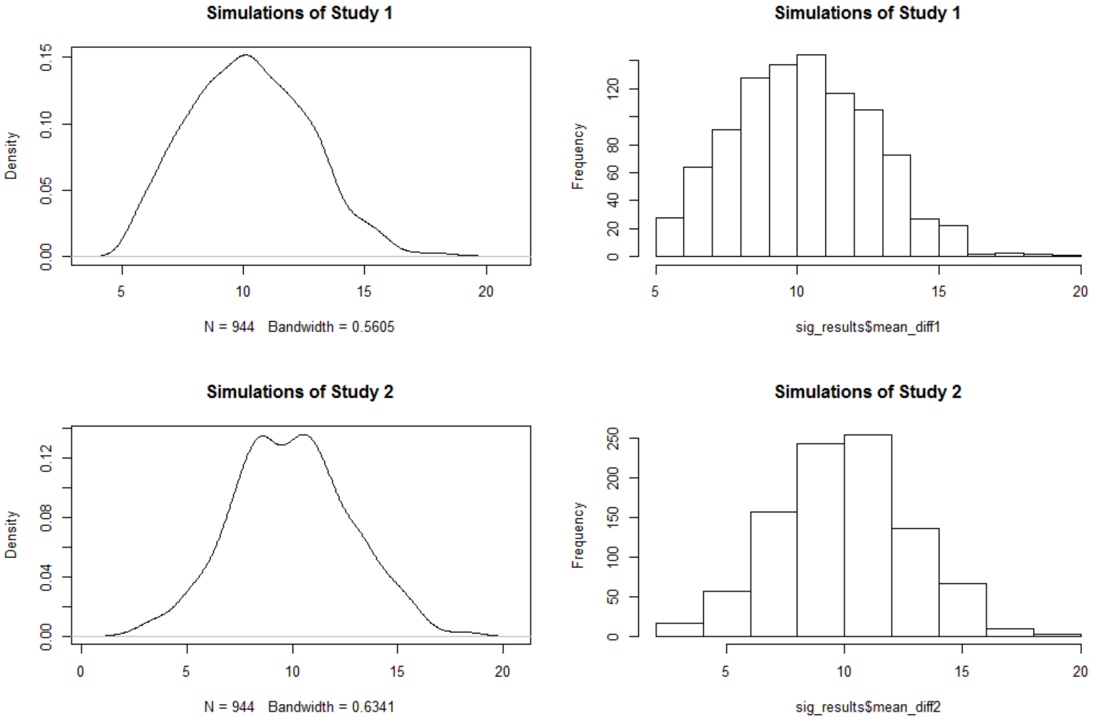


Figure D.3: The distributions for trials with 95% power. The truncated Normal distribution for study 1 appears to have a higher truncation point as the power increases.

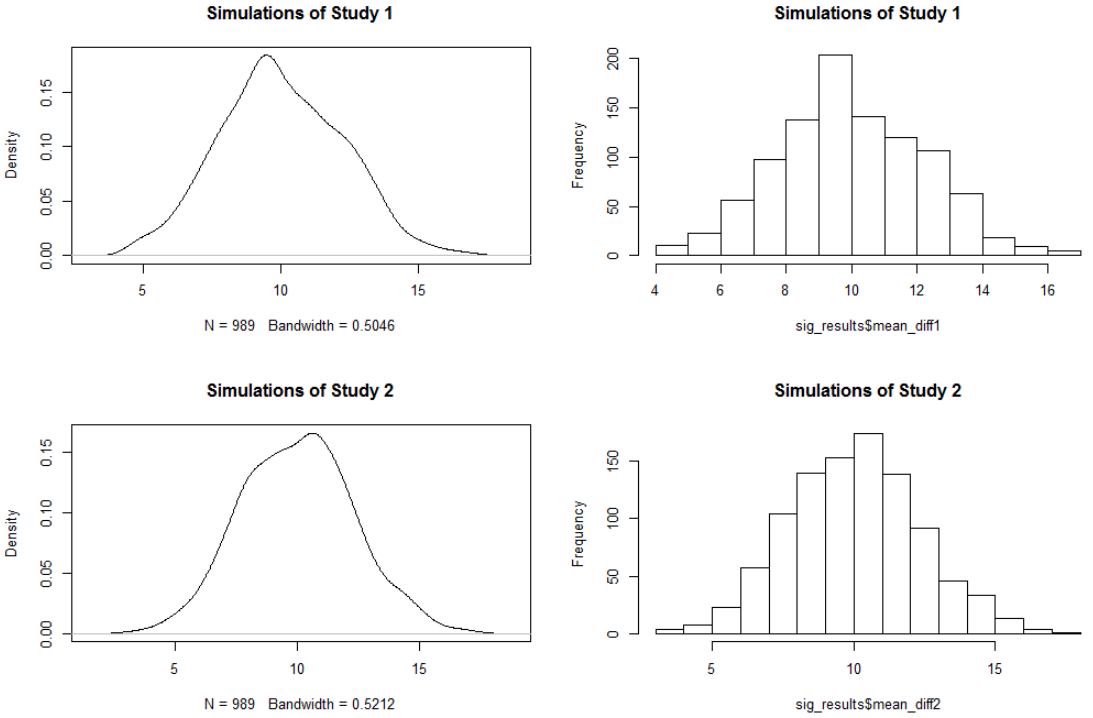


Figure D.4: The distributions for trials with 99% power. The truncated Normal distribution for study 1 appears to have a higher truncation point as the power increases.

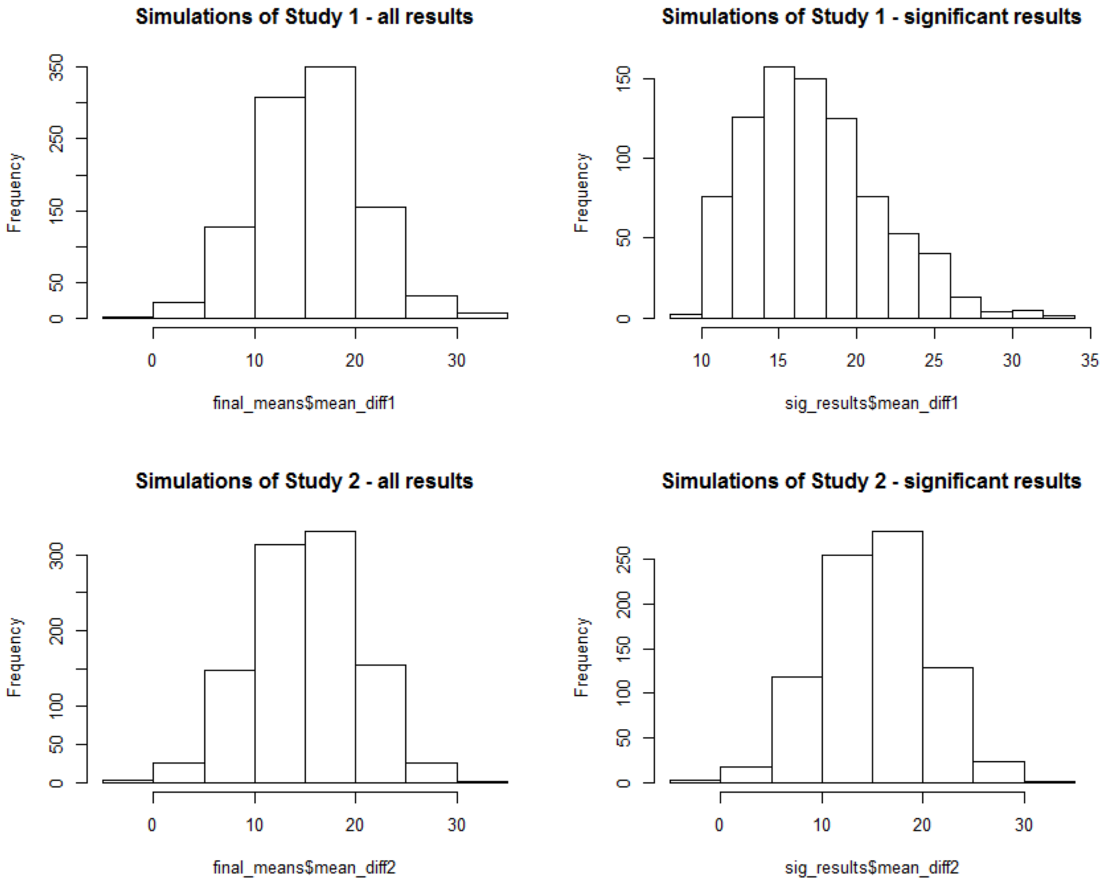


Figure D.5: The distributions for trials with 80% power and effect size of 0.3. The truncation point is noticeable.

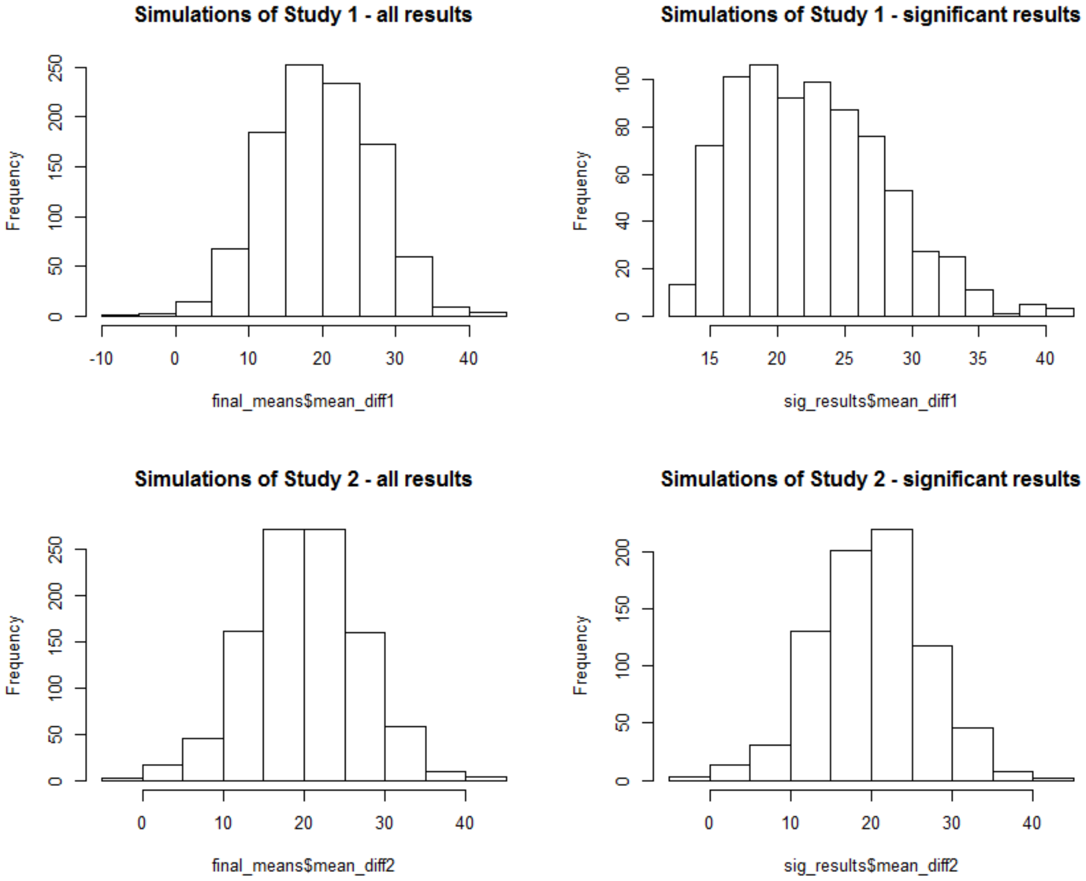


Figure D.6: The distributions for trials with 80% power and effect size of 0.4.

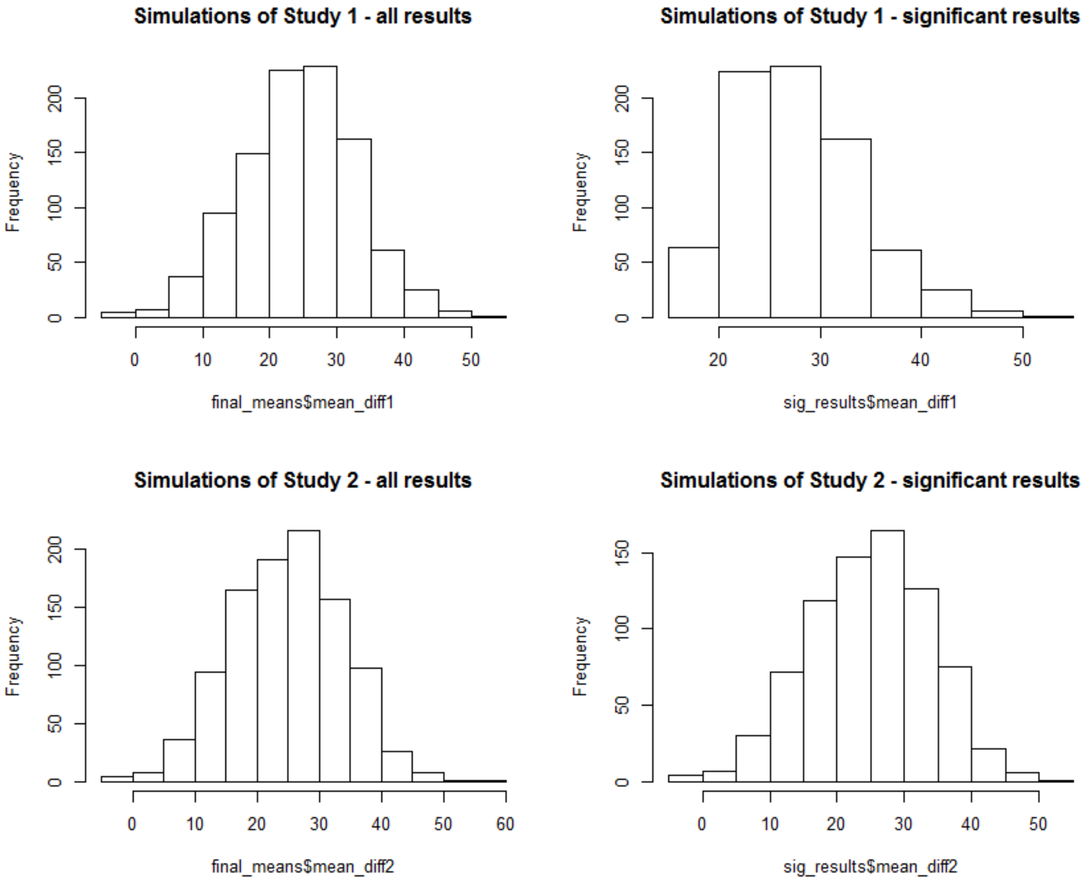


Figure D.7: The distributions for trials with 80% power and effect size of 0.5.

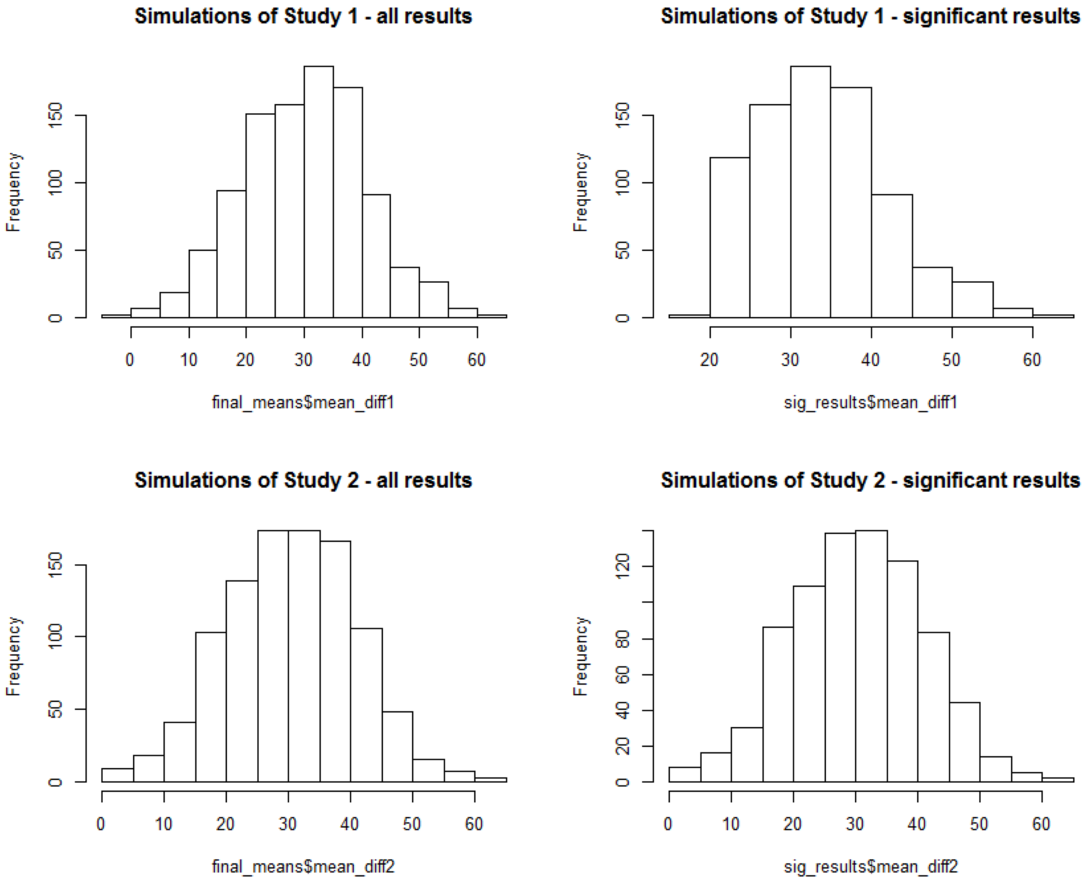


Figure D.8: The distributions for trials with 80% power and effect size of 0.6.

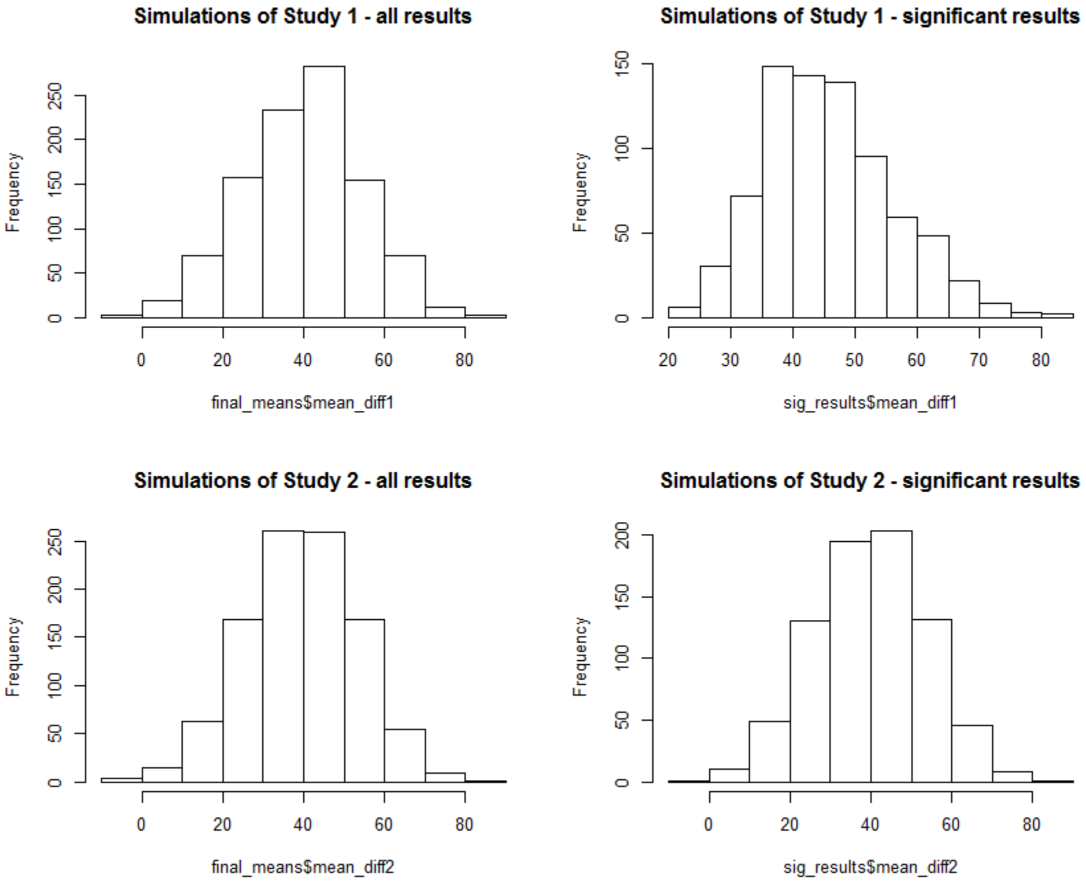


Figure D.9: The distributions for trials with 80% power and effect size of 0.8.

D.2 Sample Size Distributions for Part 2 Simulations

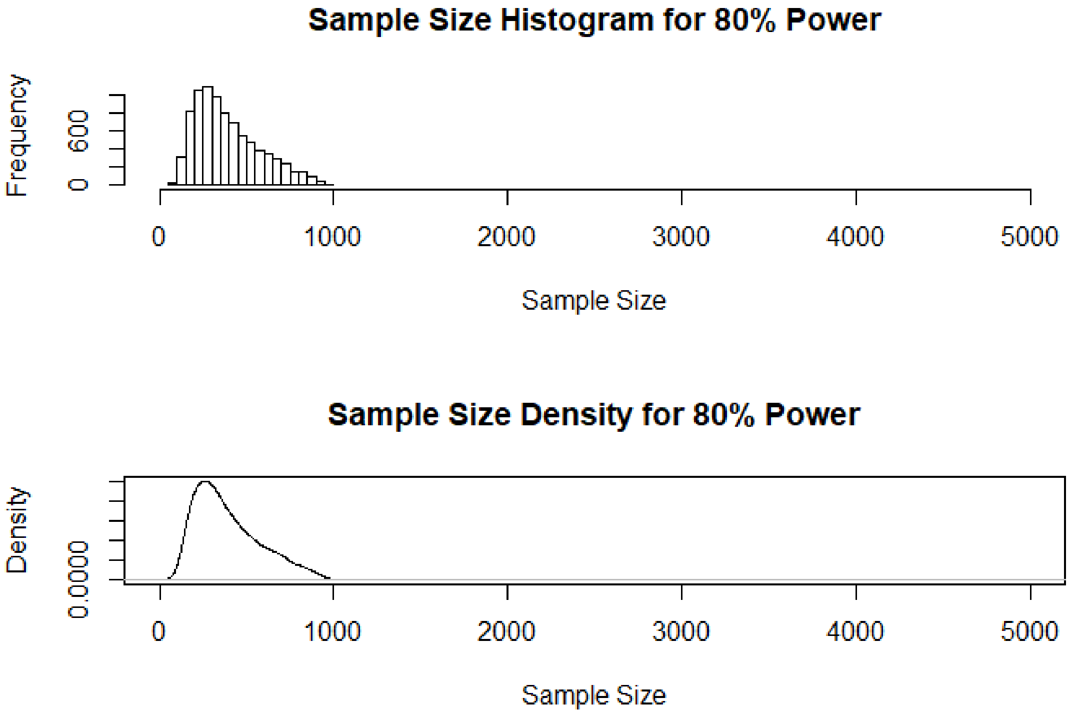


Figure D.10: The distribution and histrogram of Sample Sizes with 80% power and constant effect size.

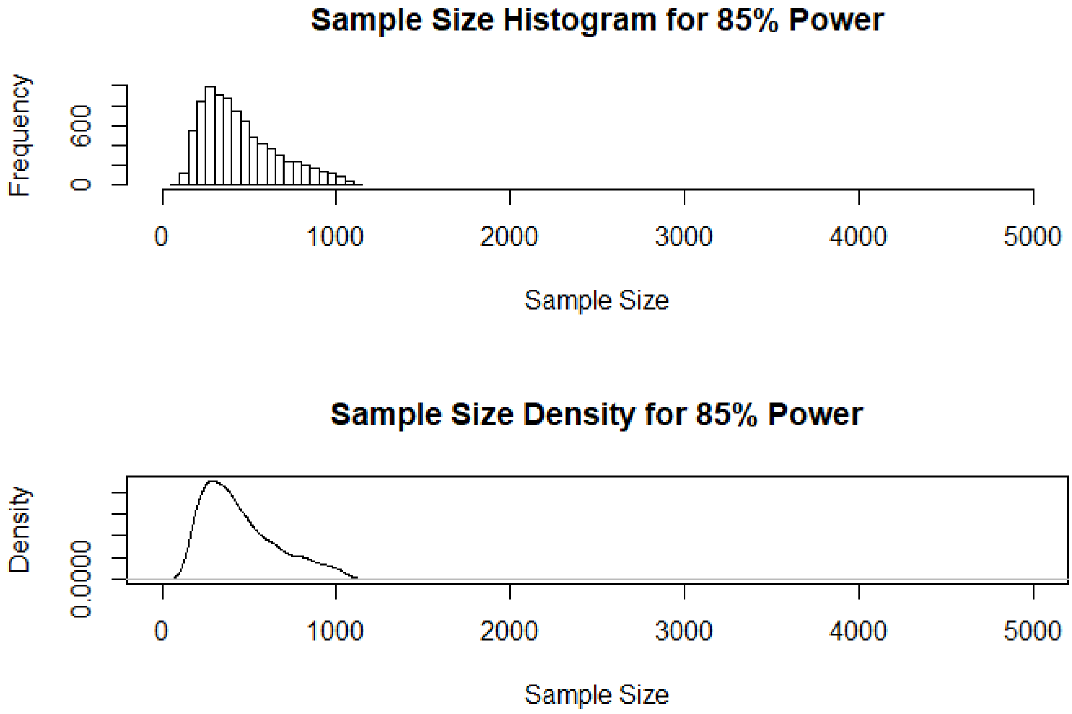


Figure D.11: The distribution and histrogram of Sample Sizes with 85% power and constant effect size.

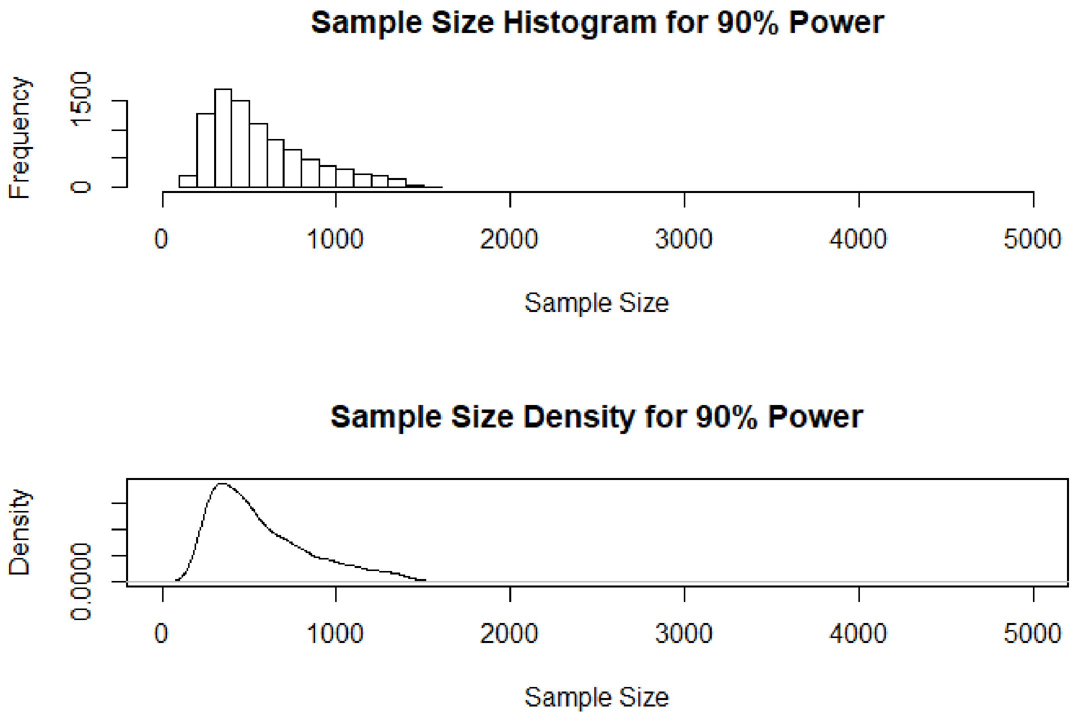
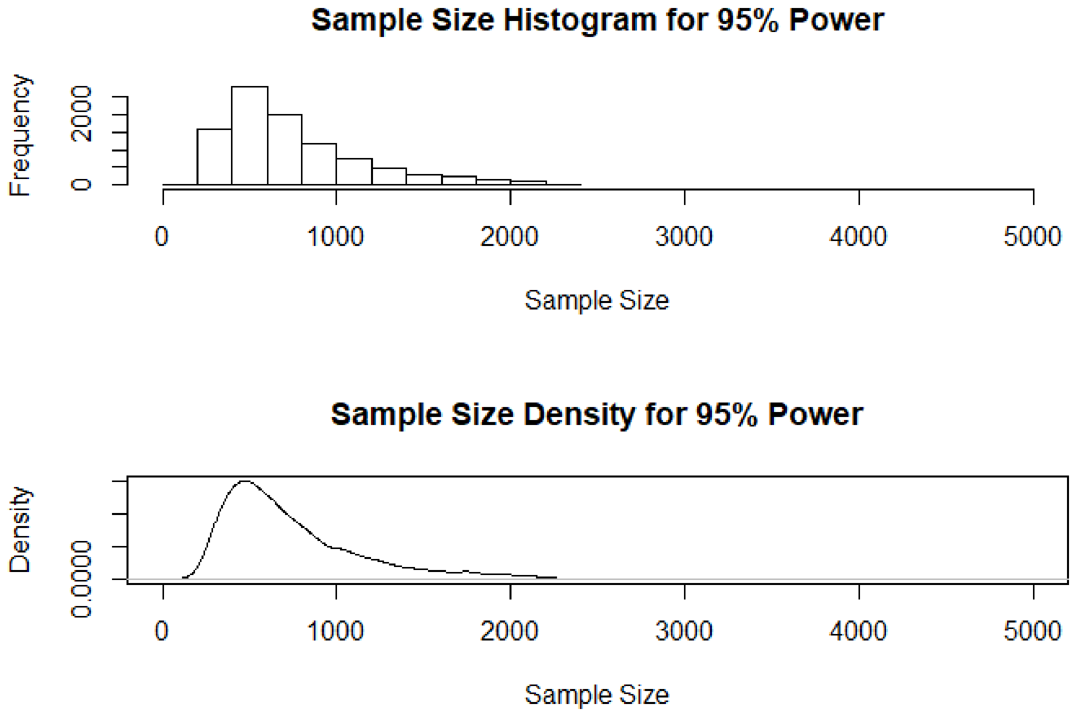


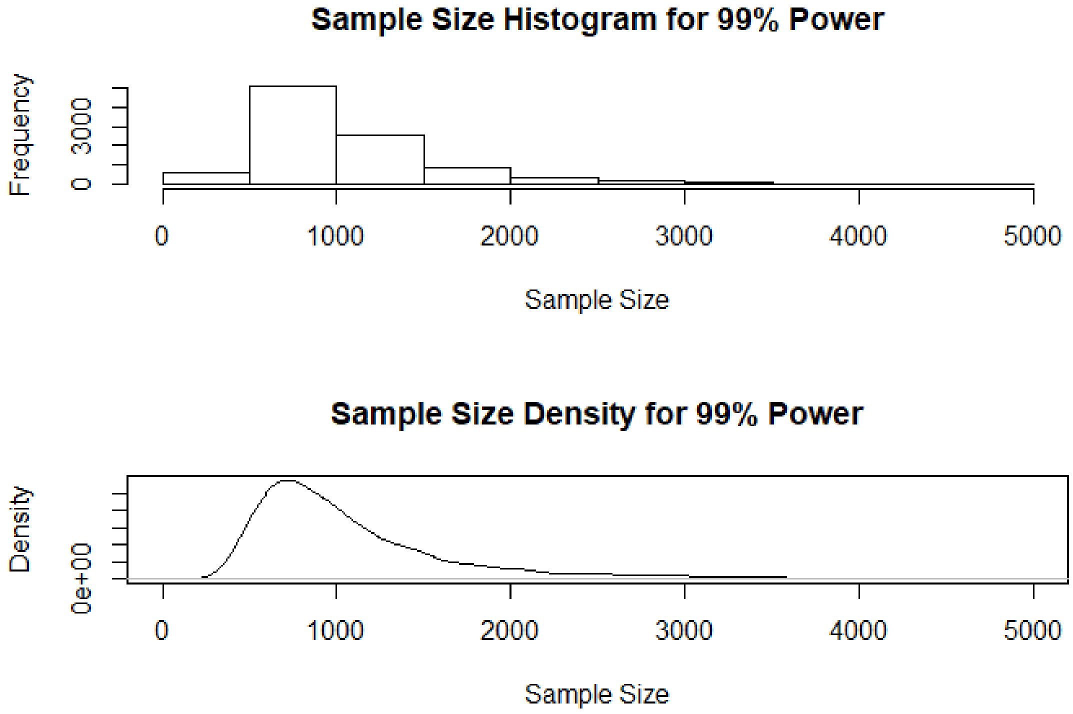
Figure D.12: The distribution and histrogram of Sample Sizes with 90% power and constant effect size.



Sample Size Histogram for 95% Power

Sample Size Density for 95% Power

Figure D.13: The distribution and histrogram of Sample Sizes with 95% power and constant effect size.



Sample Size Histogram for 99% Power

Sample Size Density for 99% Power

Figure D.14: The distribution and histrogram of Sample Sizes with 99% power and constant effect size.

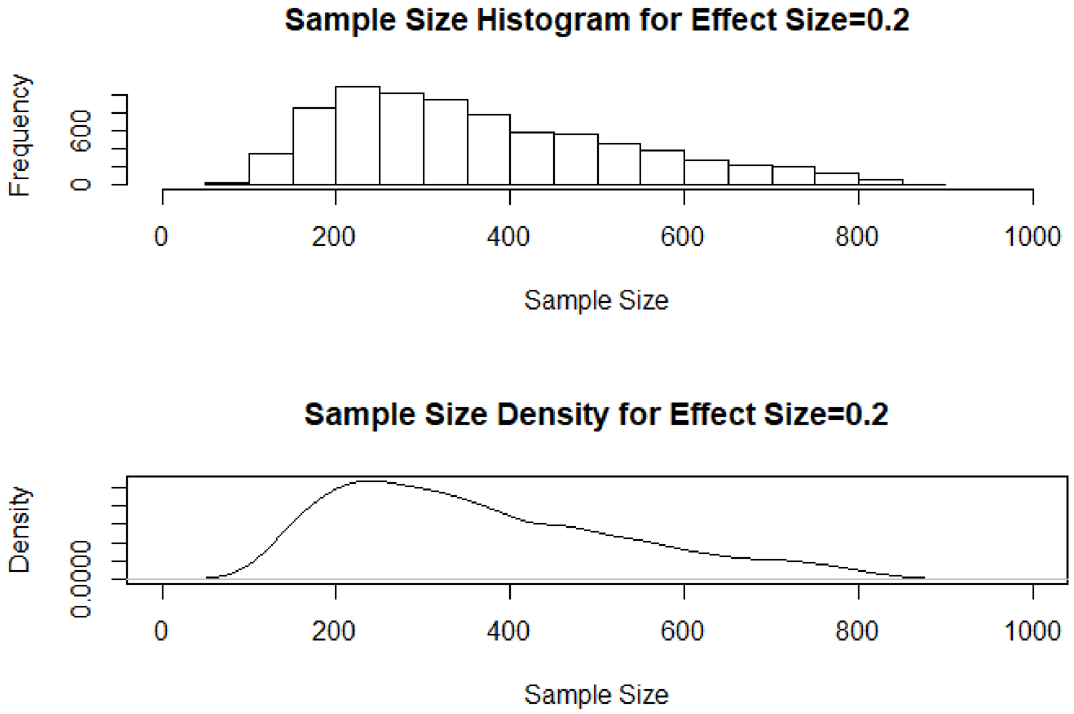


Figure D.15: The distribution and histrogram of Sample Sizes with constant power and effect size of 0.2.

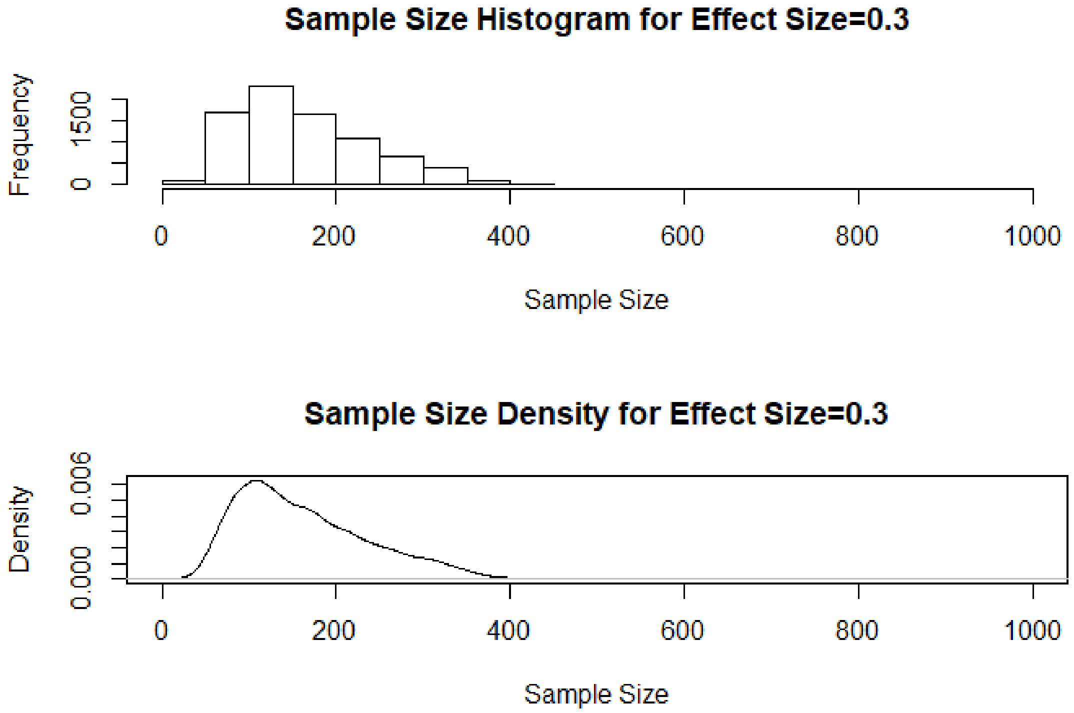


Figure D.16: The distribution and histrogram of Sample Sizes with constant power and effect size of 0.3.

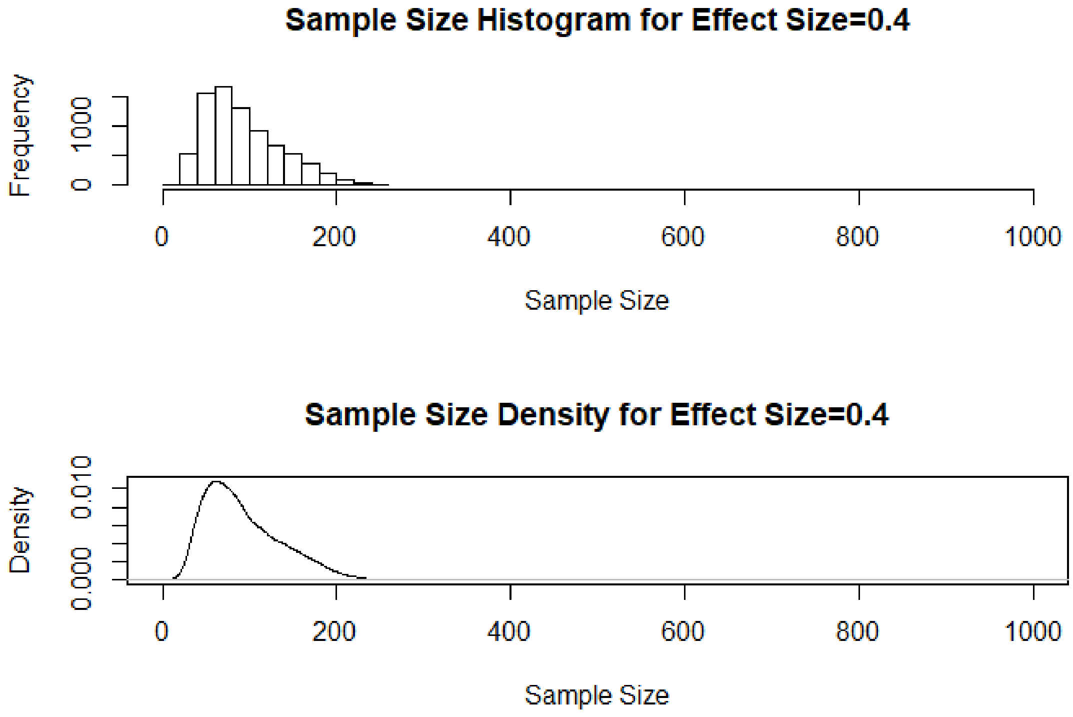


Figure D.17: The distribution and histrogram of Sample Sizes with constant power and effect size of 0.4.

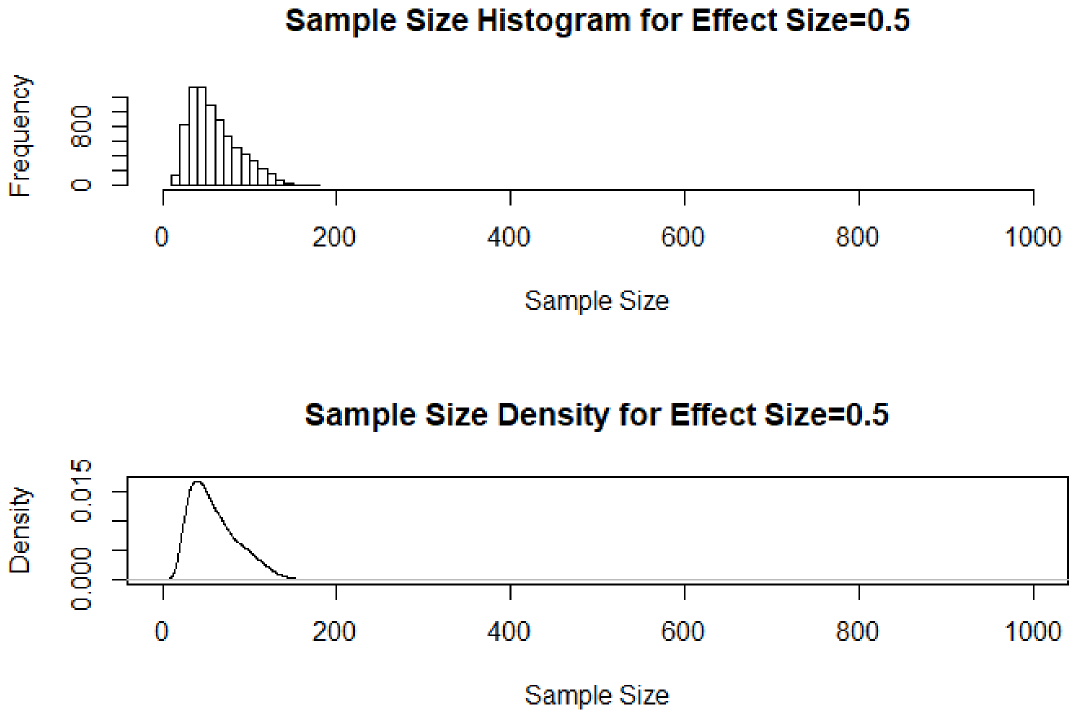


Figure D.18: The distribution and histrogram of Sample Sizes with constant power and effect size of 0.5.

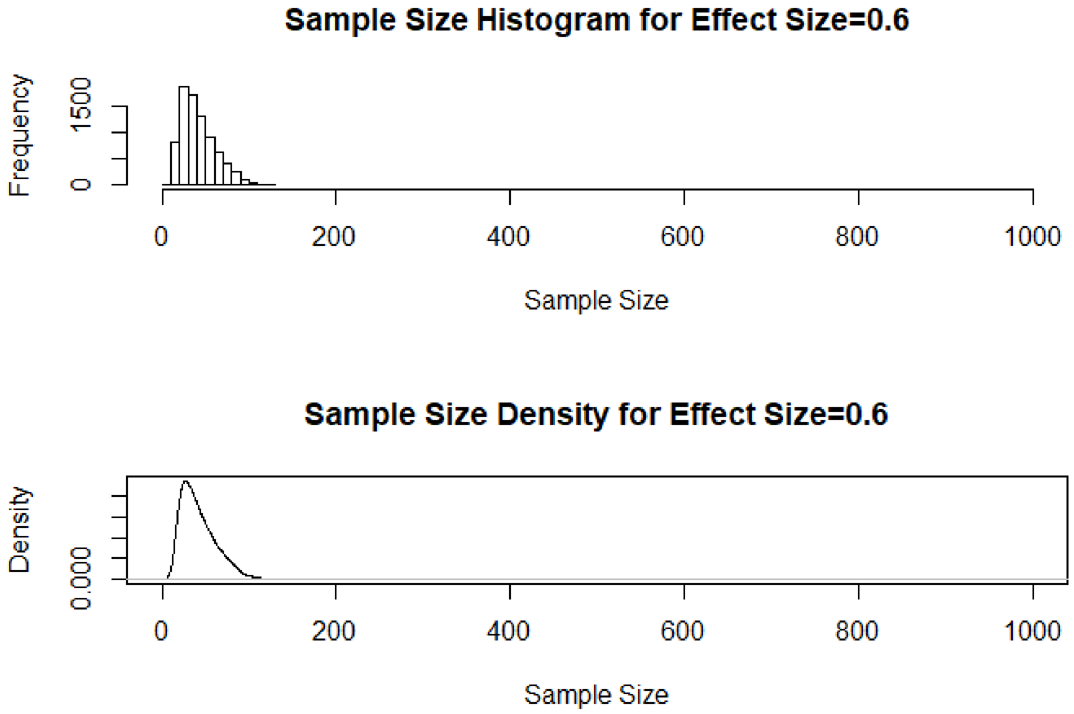


Figure D.19: The distribution and histrogram of Sample Sizes with constant power and effect size of 0.6.

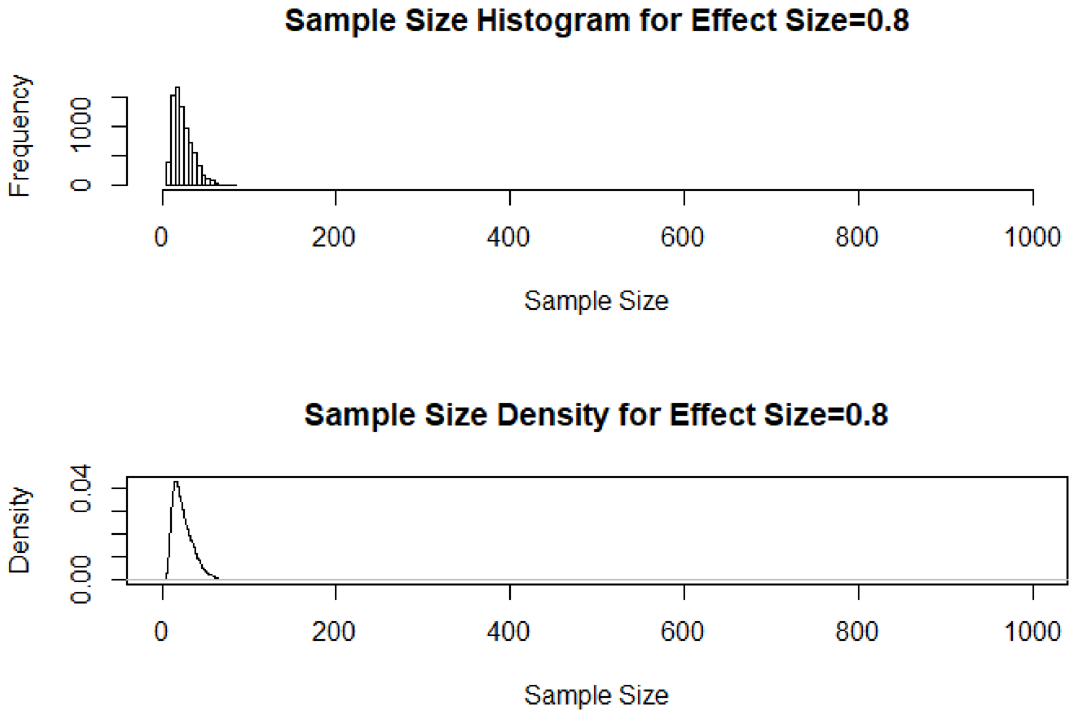


Figure D.20: The distribution and histrogram of Sample Sizes with constant power and effect size of 0.8.

E. Chapter 7

The tables included in this appendix are from chapter 7. They are the intermediary tables for the mathematical solution of the truncation point. Whilst they are useful, there was a concern that they would overshadow the true results tables and increase confusion. Thus they are placed in this appendix.

E.1 Mathematical Truncation Point Intermediary Tables

Below are two tables which show the intermediary results for section 7.2.4. Due to the use of the non-central t-distribution, the non-centrality parameter from Equation 7.18 is used as described in chapter 2. The standardised differences are being used, therefore the standard deviation, σ, equals 1. If the non-centrality parameter is used, then the variance becomes the unit variance and equals 1 and thus is easier to work with. These results are for when the variance is equal to 1, which makes the calculations more simple.

- σ is set to equal 1 for simplicity
- Sample size per arm (n) is set to the same values calculated in chapter 6
- μ is the mean difference, based on the non-central t-distribution. Thus $\mu = ES\sqrt{n/2}$ where ES is the effect size.
- a_{det} is calculated by $x.\mu$, where x is the associated value from Table 7.1
- a is the mathematical truncation point calculated as $a = t_{2n-2,1-\alpha/2}$
- $\mu*$ has been calculated using Equation 7.12.

The ratios for the chapter 6 bias are brought forward from Tables 6.3 and 6.4 and taking the inverse of the ratio. The ratio of $\mu/\mu*$ is calculated from the data in Table E.1 and Table E.2.

Effect Size = 0.2							
	Sample	**Truncation**		**Mean Difference**		**Ratio**	
Power	**Size (n)**	a_{det}	a	μ	$\mu*$	**Ch 6 Bias**	$\mu/\mu*$
80	393	1.962	1.963	2.804	3.154	0.885	0.889
85	450	1.962	1.963	3.000	3.274	0.926	0.916
90	526	1.962	1.962	3.243	3.439	0.945	0.943
95	651	1.963	1.962	3.608	3.717	0.970	0.971
99	950	1.960	1.961	4.290	4.316	0.994	0.994

Table E.1: A comparison of mathematically calculated truncation points and ratios of mean differences with simulated values for various powers.

Power = 80%							
	Sample	**Truncation**		**Mean Difference**		**Ratio**	
Effect	**Size (n)**	a_{det}	a	μ	$\mu*$	**Ch 6 Bias**	$\mu/\mu*$
0.2	393	1.962	1.963	2.804	3.154	0.886	0.889
0.3	175	1.962	1.963	2.806	3.157	0.891	0.889
0.4	99	1.962	1.963	2.814	3.164	0.884	0.889
0.5	64	1.962	1.963	2.828	3.175	0.880	0.891
0.6	45	1.962	1.963	2.846	3.189	0.892	0.892
0.8	26	1.962	1.963	2.884	3.220	0.896	0.896

Table E.2: A comparison of mathematically calculated truncation points and ratios of mean differences with simulated values for various effect sizes.