Unblinded Sample Size Re-estimation in Randomised Controlled Trials

By:
Julia M. Edwards

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy

School of Health and Related Research
Faculty of Medicine, Dentistry and Health
The University of Sheffield

November 2020

Abstract

Introduction: Sample size calculations require assumptions regarding treatment response and variability. However, there is often limited information prior to the trial and essentially a “best guess” is used. Incorrect assumptions can lead to either under or over powered trials, which poses an ethical concern. Unblinded sample size re-estimation (uSSR) designs allow treatment effect assumptions to be re-assessed at an interim analysis, and can modify the sample size if necessary. Those based on conditional power (CP) calculations rely on an assumption of the future treatment effect of the second stage sample size. Guidance for associated design features is unclear, and it is unknown which CP assumption should be used. Therefore this thesis aims to compare existing uSSR methodologies, to consider the associated design features, and make clear recommendations for future uSSR implementation.

Methods: The thesis is split into four main sections; a comprehensive review of the literature, retrospective data analysis applied to real world trial datasets, simulations, and a discussion on logistical implementation and future trial planning using uSSR. Promising Zone (PZ) and Combination Test (CT) designs are explored in detail. Four possible future treatment effect assumptions are investigated in the CP calculation. Maximum restrictions on sample size, timings of interim analyses and proportions of pipeline patients are also explored, as well as the incorporation of a futility boundary.

Results: The observed treatment effect was calculated for 21 retrospective case studies and found to be within $\pm 1^{*} SE$ of the end treatment effect from 57% through trial duration. The current trend assumption was best when a smaller than anticipated or no effect was observed. The hypothesised assumption was best when the observed effect was close to that planned. An 80% optimistic confidence limit of the observed current trend was shown to work well in either scenario: close to that planned, or smaller/zero and is a considered a good “middle-ground” between the two. The PZ design is the easiest design to implement, with an incorporated futility boundary potentially offering a funder to cut some of the losses if the effect is much smaller than planned. The CT design also offers a decrease in sample size, but should consider a minimum restriction on sample size in small trials, to avoid the inflation of Type I error.

Conclusions: This thesis has explored uSSR designs, associated logistical features, and the assumptions of the CP calculation used in these designs. This thesis explores a 10% futility threshold and recommends an interim timing between 60-70%, with between 50-100% maximum increases in sample size, and the use of the 80% optimistic confidence limit assumption for the CP calculation in this scenario. Further recommendations are found in Chapter 10.

Acknowledgements

Firstly I would like to thank my University supervisors; Professor Steven Julious and Professor Stephen Walters. Your support, guidance and comments over the last three years has been invaluable. Thank you for always making time for me, and for making me laugh in our supervision meetings.
I would like to thank my Industry supervisor, Dr Cornelia Kunz for introducing me to the industry side of trials, for the warm welcome I received during my visits to Biberach, and our useful discussions. Thank you for encouraging me to pursue the methodology side.
I would like to thank the University of Sheffield and Boehringer Ingelheim for their funding which allowed me to carry out this research, and to the University for the funding extension following the recent pandemic.

I would like to thank all those who provided real-world clinical trial data, especially:
Eisai, GSK, Sanofi and www.clinicalstudydatarequest.com
Professor Valerie Pomeroy and the FAST INdiCATE trial team
Professor Janet Powell, Dr Pinar Ulug and the IMPROVE trial study team
Professor Catherine Hewitt, Professor Simon Gilbody, the CASPER trial team and the CASPER PLUS trial team
Professor Linda Sharples, Christine Mills, Thomas Devine and the AMAZE trial team
Professor Stephen Walters, the SELF trial team, the Acupuncture trial team, and the Corn Plasters trial team
Professor Steven Goodacre, Mike Bradburn, the 3MG trial team and the RATPAC trial team

I would like to thank Dr Matthew Lee, Dr Matthew Bursnall and Professor Matthew Wilson for allowing me to use the RIPOSTE study to propose a prospective sample size re-estimation design, for giving up their time, and for providing incredibly useful comments and feedback. I would also like to thank Professor Goodacre for his time and comments on the study design.

I would like to thank my family for your encouragement, and for always believing I could solve “that maths problem”. Thank you to Janelle, for keeping me (relatively) sane in a somewhat insane time. Thank you to Scott for always being around for a pint and a distraction.

Lastly, but by no means least, I would like to thank my partner TJ for putting up with me over the last few years. Thank you for keeping me caffeinated, doing the endless piles of washing up, and of course for correcting that single comma! You believed in me even when I couldn’t see the end in sight and I am so grateful for all that you have done for me.

Contents

List of Figures

List of Tables

Abbreviations

AAA	Abdominal Aortic Aneurysm
AD	Adaptive Design
AdGSD	Adaptive Group Sequential Design
AED	Anti-Epileptic Drug
AF	Atrial Fibrillation
AMI	Acute Myocardial Infarction
ARAT	Action Research Arm Test
ASN	Average Sample Number
bSSR	Blinded Sample Size Re-estimation
CDR	Cognitive Drug Research
CHMP	Committee for Medicinal Products for Human Use
CI	Confidence Interval
CP	Conditional Power
CPT	Conventional Physical Therapy
CSDR	Clinical Study Data Request
CT	Combination Test
DIA	Drug Information Association
DMSC	Data Monitoring and Safety Committee
DRGSD	Delayed Response Group Sequential Design
EMA	European Medicines Agency
EME	Efficacy and Mechanism Evaluation
EORTC	European Organisation for Research and Treatment of Cancer
FDA	Food and Drug Administration
FFNS	Fluticasone Furoate Nasal Spray
FST	Functional Strength Training
GSD	Group Sequential Design
GSK	GlaxoSmithKline

HTA	Health Technology Assessment
ICC	Intraclass Correlation
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ITT	Intention to treat
MCID	Minimum Clinically Important Difference
MFNS	Mometasone Furoate Nasal Spray
MLE	Maximum Likelihood Estimate
MPT	Movement Performance Therapy
MRMC	Multi-reader multi-case
NICE	The National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
OR	Odds Ratio
PHQ-9	Patient Health Questionnaire-9 items
PhRMA	Pharmaceutical Research and Manufacturers of America
PZ	Promising Zone
RCI	Repeated Confidence Interval
RCT	Randomised Controlled Trial
SAP	Statistical Analysis Plan
SBA	Serum Bactericidal Antibody
SD	Standard Deviation
SE	Standard Error
SF-36	36 item Short Form
SF-12	12 item Short Form
SPADI	Shoulder Pain and Disability Index
SSR	Sample Size Re-estimation
uSSR	Unblinded Sample Size Re-estimation
VAS	Visual Analogue Scale

List of Symbols

Common symbols

H_0	- Null Hypothesis
H_1	- Alternative Hypothesis
α	- Type I Error
β	- Type II Error
p	- p-value
δ	- True Standardised Treatment Effect
δ_{plan}	- Target Standardised Treatment Effect
$\hat{\delta}_{obs}$	- Observed Standardised Treatment Effect
d	- True Mean Difference
d_{plan}	- Target Mean Difference
$\hat{d}_{obs}$	- Observed Mean Difference
σ^2	- True Variance
σ^2_{plan}	- Target Variance
$\hat{\sigma}^2_{obs}$	- Observed Variance
μ_A/μ_B	- True means (Group A/B)
π_A/π_B	- True Event Rates (Groups A/B)
n	- Total Planned Sample Size
n_A/n_B	- Planned Sample Size (Group A/B)
r	- Randomisation Allocation Ratio
d_{NI}	- Non-Inferiority Limit
d_{Eq}	- Equivalence Limit
n_1	- First Stage Sample Size
n_2	- Second Stage Sample Size
n^*	- New Total Sample Size
n_2^*	- New Second Stage Sample Size
n_{max}	- Maximum Sample Size Allowed

n_{req} - Sample Size Required

CP_{min} - Minimum Conditional Power Value

$\tilde{d}$ - Assumed Future Treatment Difference

Z_k - Observed Test Statistic at Stage k

$\hat{\delta}_k$ - Observed Treatment Effect at Stage k

$\hat{d}_k$ - Observed Mean Difference at at Stage k

$\hat{\sigma}_L^2$ - Gould and Shih Parameter: Lumped Variation Estimator

J - Number of Steps (Stepwise design)

v_j - Step value (Stepwise design)

M -Total Sample Size (Not Pre-specified) (Stepwise design)

τ - Information Fraction (n_1/n)

K - Total Number of Interim Analyses

I - Indicator Function

γ - Combination Test Design Parameter

$\phi()$ - Standard normal density function

$\Phi()$ - Cumulative normal distribution function

Alpha Spending Functions

$\alpha_*(\tau)$ - Alpha spending function

C_{Pocock} - Critical Value (Pocock Method)

C_{OBF} - Critical Value (O'Brien Fleming Method)

C_{WT} - Critical Value (Wang Tsiatis Method)

Δ - Wang Tsiatis Parameter

c_k - Critical Value at Stage k

c_{fut} - Critical Value (Futility)

c_{eff} - Critical Value (Efficacy)

c_α - Critical Value (α level)

$\tilde{c}_2$ - Required Critical Value to 'Spend' Remaining α

Controlling Type I Error

$A(z)$ - Conditional Error Function

$A_l(z)$ - Linear Conditional Error Function

$A_{circ}(z)$ - Circular Conditional Error Function

C_{max}, k_{lsw}, h_{lsw} - Li, Shih & Wang Design Parameters

α_1/α_2 - α value at Stage 1/2

α_{fut} - α value for futility

f - Inflation factor (Fisher Variance Spending)

n_k - Sample Size in Each Stage k

d_k - Mean Difference in Each Stage k

Z_1/Z_2 - Test Statistic at Stage 1/2

X_1/X_2 - Test Statistic at Stage 1/2

X_2^* - Test Statistic Stage 2 based on n^* Observations

X_C - Combined Test Statistic

X_{CHW} - CHW Test Statistic

X_W - Weighted Test Statistic

X - Unweighted Test Statistic

Delayed Response Group Sequential Designs

λ - Time Between Randomisation and Data Availability

$\mathscr{I}_k$ - Fisher Information at Interval k

$\tilde{n}_k$ - Sample Size at Stage k

S_k - Test Statistic at Stage k

U^g - Test Statistic Based on g Patients

$\hat{d}_k$ - Observed Treatment Difference at Interim Analysis k

$\tilde{d}_k$ - Observed Treatment Difference at $\tilde{n}_k$ Recruited Patients

Repeated Confidence Intervals

ρ_k - Sequence of k Repeated Confidence Intervals

d_L - Lower Confidence Interval of d

d_U - Upper Confidence Interval of d

Lan DeMets

$\zeta(\tau)$ - Lan DeMets Stopping Bound

$\zeta(\tau_j)$	- Lan DeMets Stopping Bound at Interim Look j
b_k	- Stopping Boundaries at Interim Look k
ε	- Conditional Rejection Probability

1 | Introduction

1.1 Background

Clinical trials aim to evaluate healthcare interventions and the effect of health related outcomes in human participants (WHO 2018). Randomised Controlled Trials (RCTs) are considered to be the "gold standard" design in evidence-based evaluation of interventions (Sackett 1996) as patient characteristics should be equally distributed between the groups due to the randomisation process (Roberts 1999). However, the cost of clinical trials is increasing in both publicly funded and industry settings (Shore 2012; Collier 2009). A number of reasons have been attributed to this growing cost, including longer and more complex trials, challenges in recruiting and retaining participants, and an increased regulatory burden (Lindsey 2009). Therefore, there is a need to improve the efficiency and reduce costs in clinical trial settings.

A sample size calculation is performed prior to the start of the trial and makes assumptions regarding the hypothesised treatment response and the statistical variability as the true values are unknown (Noordzij 2011). These can be estimated using similar studies, or any prior knowledge of the treatment. However, information is often limited, causing the quantification of these parameters to be very difficult and could lead to a questionable sample size calculation. Inaccurate sample size estimates can result in an under or over powered study and potentially waste valuable resources (Chen 2004). Although some historical data may be available to guide this estimate, this parameter is still essentially a "best guess" at the planning stage (Herbert 2000).

Adaptive Designs (ADs) have become increasingly popular in recent years due to their flexibility and improved efficiency of conventional clinical trials (Chow 2011). Researchers implementing adaptive clinical trials are able to make use of accumulating data at pre-specified points in the trial, and make decisions affecting the remainder of the trial based on data observed at the interim analysis. Methodology has been developed for a number of types of adaptation. Some key adaptation strategies include adaptive dose finding, seamless designs incorporating two phases into one trial, stopping rules for early termination of the trial, and a modification of the required sample size (Bhatt 2016).

ADs can increase the efficiency of a clinical trial, benefit both trial participants and future patients, and improve the allocation of available resources (Gallo 2006). However, they also present their own challenges, including complex designs, logistical problems, and the need to maintain statistical rigour (Gallo 2006). Whilst potentially advantageous in terms of maintaining power, using an AD can greatly increase the complexity of the study, and therefore researchers should weigh up the potential advantages and disadvantages before committing to an AD.

An AD using Sample Size Re-estimation (SSR) methods could offer researchers the opportunity to re-evaluate sample size estimates during the trial progression to obtain the necessary power at the final analysis, either by looking at treatment effect (unblinded) or without revealing treatment allocation (blinded) (Chow 2011). As an indication of the growing interest and general upward trend of SSR in the last 30 years, a graph to show the number of publications involving SSR or related terminology was plotted using number of publications included in Web of Science since 1992. Note that the final column is only an indication of work published so far this year, and is not yet complete.

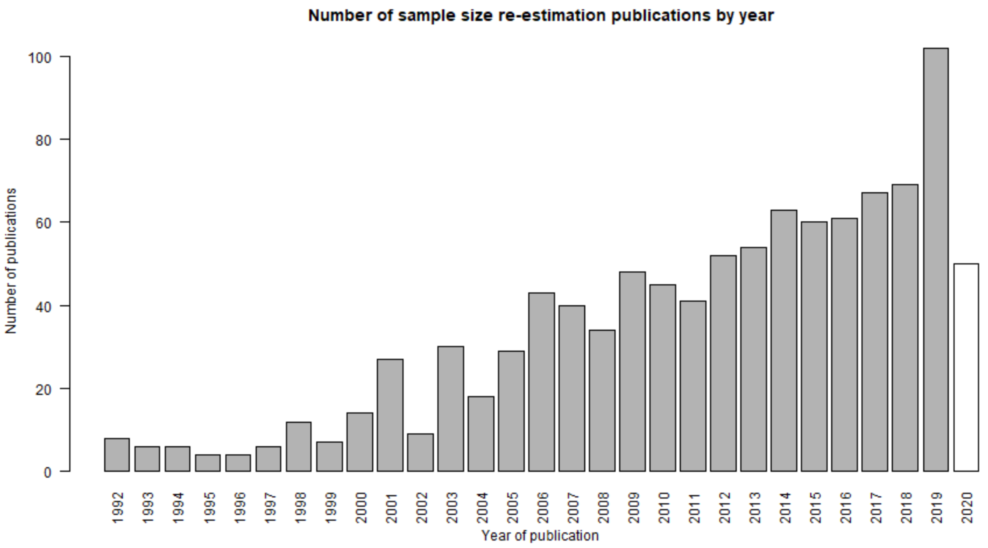


Figure 1.1: Number of publications on sample size re-estimation by year according to web of science

In 2011, Mehta and Pocock published the "Promising zone" methodology for Unblinded Sample Size Re-estimation (uSSR) (Mehta 2011). At an interim analysis, Conditional Power (CP) is calculated, representing the projected power at the end of the study having observed the data so far and assuming a future treatment effect to be observed in the remainder of the trial (Lachin 2005). If CP lies in a pre-determined "promising zone", sample size is increased according to a sample size rule; otherwise sample size remains the same as originally planned. Whilst the ultimate aim was to propose a 'simple' design for easy implementation for trialists, the methodology has been a source of controversy over the years for a number of reasons. Firstly, the use of an unadjusted critical value at the final analysis is straightforward for researchers and well understood, but many have criticised the efficiency of the design due to its conservatism (Glimm 2012). Additionally, the treatment effect can be highly variable at the interim stage, which can have a substantial impact on the sample size increase (Jennison 2015). Full details of the promising zone methodology and its strengths and limitations are described in Section 3. Jennison and Turnbull describe an alternative sample size rule in their 2015 publication (Jennison 2015). By using inverse normal combination tests, the sample size may be increased at any value of the interim observed test statistic, z_1, without inflating the Type I error, and results in a smoother sample size increase rule. Comparisons between the two designs are limited within the literature, and often impose different optimality criterion that may not necessarily be universal (Pilz 2019).

The knowledge of the new sample size following an interim analysis means that trialists may be able to back-calculate the treatment effect, and this knowledge may influence the remainder of the trial and therefore the validity of trial results. In 2016, Liu and Hu presented a stepwise function for SSR. Their rule for SSR increases in steps to a maximum point, and then decreases back to the original sample size in similar steps. Now, knowledge that the sample size has increased by 10% for example may indicate a particularly low CP value (say 15-20%), or a particularly high CP value (say 70-80%), which may be a useful design consideration if investigators have particular concerns about operational bias following the interim analysis. Full details of their methodology are provided in Chapter 3.

1.2 Publicly funded vs industry trials

Whilst publicly funded and industry trials both aim to improve medical treatments and patient care, there are some key differences where trials in the two settings can differ. Generally, the main focus in most industry trials is for the approval of an unlicensed treatment, with the ultimate aim of generating profits from the drug (Laterre 2015). A 'negative' trial in these settings means that the treatment under investigation is no longer developed in this setting, and is either scrapped completely or used in an alternative setting. Furthermore, there is an advantage of showing efficacy for industrial trials, as a non-inferior or equivalent drug will not be cost effective to develop in a saturated market. The ultimate aim is to show some benefit to the patient in terms of efficacy, decreased side effects or improved dosing schedule for example, in order for the patient to choose their product over an alternative (Mossialos 2005).

On the other hand, publicly funded trials largely investigate already licensed interventions and compare these to current standard care, which may have been placed in the care pathway with little or no evidence. In this setting, a 'negative' trial is still considered beneficial to researchers and to future patients as an ineffective treatment may be withdrawn from the standard care pathway. Although industry and publicly funded trials are not limited only to the cases in the above description, for the purpose of this thesis industry trials will refer to the approval of unlicensed treatments, and publicly funded trials will refer to comparisons of licensed trials to current standard practice.

1.3 Research Question

uSSR is a potentially valuable tool in trial design, with the ability to reduce the risk of an underpowered study and its capability in handling uncertainty in planning assumptions. With so much debate over uSSR methodology, it is important to gain a better understanding of both statistical and operational factors involved in designing and implementing a SSR. Specifically:

> *What sample size rule framework and associated design features need to be considered when using an uSSR in clinical trials with either a continuous or binary primary endpoint?*

The ultimate aim of the thesis is to make recommendations for uSSR implementation, which will assist the planning of future trials wishing to implement an uSSR trial design. Specific objectives include:

1. Compare existing methodologies for uSSR using CP calculations, with a focus on promising zone and combination test designs
2. Incorporate stopping boundaries in each methodological framework and compare interim decision making
3. Investigate the future treatment effect assumption used in the CP equation
4. Explore CP values when observed effect sizes are equal to, or different by some amount to the target effect size
5. Make recommendations for the planning of a future trial using SSR including operational considerations such as when an interim analysis should be carried out, and the maximum sample size increase to consider

To carry out the thesis aims described above, the thesis will employ a variety of methods, including a comprehensive systematic review of uSSR literature, retrospective data analysis to real clinical trial data, simulation work, and the application of designing a prospective clinical trial.

This thesis is limited to continuous and binary endpoints only- survival endpoints are not considered here. Both small and large sample size studies will be compared, as well as short, medium and long times to primary outcome data becoming available (e.g. 1-2 weeks, 3-6 months, and $\geq$ 1 year respectively). Whilst the priorities on the outcome of industry and publicly funded trials may differ, neither have an infinite source of funding, and so it is expected that both could have similar priorities regarding interim decision making. For completeness, both funding types are being considered within the thesis.

1.4 Overview of Thesis

Chapter 2 presents background statistical methodology that is relevant to the remainder of the thesis. This includes ‘fixed sample size’ calculations for superiority and non-inferiority trials with continuous and binary endpoints, CP calculations, and methods for SSR and Type I error control. Chapter 3 provides an in-depth review of the current literature on promising zone methodology, addresses limitations of the methodology, and presents the rationale for the specific research question for the thesis. A systematic review presents all trials that have reported using, or planning to use promising zone methodology. Trial characteristics are summarised, and issues surrounding the reporting of the methodology is discussed.

Chapter 4 contains a detailed retrospective data re-analysis plan for implementing a comparison of the three SSR methodologies. A summary of the data to be used for re-analysis is provided, including background information, a summary of original trial analysis and results, and information relating to patient and site recruitment. Some case studies will be presented within this chapter, and full details of remaining trials can be found in Appendix B.

Chapter 5 presents the results of the retrospective data analysis using the original observed treatment effect within each trial, and original sequential order of patients. Case studies of conditional power, new required sample size (n^*) in each design, and observed treatment effect results will be presented, with further details provided in Appendix C. Chapter 6 extends on the work of Chapter 5, using methods to adjust the observed treatment effect relative to the planned treatment effect. The level of misspecification of the treatment effect in the planning stage compared to observed will be compared, including zero and negative treatment effects.

Chapters 7 and 8 include details of the simulation protocol, and presents simulation results continuous and binary endpoints respectively. Chapter 9 provides an example of implementing SSR design in the design of a real study. Chapters 10 and 11 provide a discussion and conclusion, and detailed recommendations that have come from the results of the thesis.

1.5 Summary

RCTs require a pre-determined sample size using a “best-guess” of the treatment effect, of which the true effect is unknown. Using an AD with a uSSR looks at the accumulating data and can modify the sample size at the interim analysis if necessary, based on CP calculations. However, there is some debate over which uSSR design to use, and the future treatment effect assumption of the observed CP calculation. This thesis aims to compare existing methodologies, investigate CP assumptions, incorporate a futility analysis in the uSSR designs, and explore logistical factors that may impact the choice of SSR design such as timing of the interim analysis or maximum allowed sample size.

2 | Statistical methods

2.1 Introduction

Chapter 1 gave a brief introduction of ADs, the motivation behind SSR and an outline of the aims of this thesis. This chapter introduces key statistical concepts relevant to sample sizes, and SSR. The aim of this chapter is not to provide a detailed review of the current literature (which is presented in Chapter 3), but to provide the background statistical methods that have been used in the development of uSSR designs. Firstly, sample size calculations for fixed sample size trials are presented for superiority and non-inferiority trials with continuous and binary outcomes. Practical and ethical issues surrounding sample size estimates are discussed, and ADs and standard Group Sequential Designs (GSDs) are described. The motivation behind both Blinded Sample Size Re-estimation (bSSR) and uSSR is discussed and a brief introduction on methodology is presented, including CP calculations. Finally, methods for controlling Type I error are introduced, with strengths and limitations discussed.

2.2 Aims for the Section

This chapter ultimately aims to provide the background and motivation behind SSR methods. Specific objectives for this section include:

1. Introduce sample size calculations for fixed designs
2. Discuss ADs and their potential benefits and limitations in the future of clinical trials, with a focus on SSR methodology.
3. Present CP calculations more formally, and methods for controlling Type I error inflation in flexible sample size designs.

2.3 Sample size calculations

The choice of sample size in an RCT must be ethically balanced. Too many subjects could result in patients unnecessarily receiving an inferior treatment when there is already enough evidence to answer the research question available, as well as resulting in a waste of resources and additional costs incurred. On the other hand, too few subjects could mean that patients may be subjected to an experimental treatment in a study that is unable to detect any clinically important effect (Altman 1980).

For this reason, a sample size calculation should be carried out, to calculate the minimum number of subjects required to test a hypothesis with respect to a “control treatment" and an “experimental treatment". An experimental treatment may be, for example, a new drug, combination of treatments, or intervention of interest. A control treatment may be a placebo drug, a current standard practice on the care pathway, or even an effective treatment already available. There are certain ethical considerations concerning the appropriate choice of the control treatment. In cases where a treatment exists that has already been shown to be effective compared to placebo, this effective treatment should be chosen as the comparator. In this scenario, a placebo controlled trial provides no information regarding the effectiveness of the experimental treatment compared to the existing treatment (Streiner 2007). It has been argued that placebo controlled studies can be financially attractive to pharmaceutical companies as an active treatment is more likely to show a significant treatment effect when compared to a placebo, as opposed to a comparator treatment (Togo 2016).

In conventional RCTs, a sample size calculation takes place prior to the start of the trial, and will be referred to in this thesis as a conventional or ‘fixed sample size’ design. The trial then aims to recruit subjects until it reaches this target sample size, and does not take into account any accumulating data from the ongoing trial (Chen 2012). There are a number of different formulae widely available in literature, depending on study designs and outcome types. These calculations are described in more detail in Sections 2.3.3 - 2.3.4.

2.3.1 Choice of key design parameters

Assumptions of treatment difference and nuisance parameters must be made for sample size calculations, which can be highly subjective (Schulz 2005). Unfortunately, these estimates also play a vital role in the sample size estimate and can greatly influence the number of subjects required and the power of a study. Prior information may come from similar smaller trials or meta-analyses to assist in choosing parameters (Teare 2014), but often information is limited. Furthermore, even when similar studies have been carried out, it can still be challenging to use historical data for a number of reasons. These include, but are not limited to: small sample sizes, single centre studies, location of sites, a difference in patient populations due to varying inclusion/exclusion criteria, different length of follow up time, difference in monitoring procedures, varying diagnosis criteria, and healthcare improvements or change in practice over time (Viele 2013; Shih 1998) .

2.3.1.1 Statistical Significance and Power

In the true population of interest, a hypothesis may be true or false (i.e. there is a true difference in some hypothesised outcome or there is not). When testing this hypothesis based on a sample of the true population, the researcher may conclude one of two possibilities: to accept the null hypothesis (H_0), or reject the null hypothesis and instead accept the alternative hypothesis (H_1) (Biau 2010). One hopes that the trial correctly concludes whether there is a true difference or not; however this is not always the case. The Type I error (α) is the error associated with the outcome of incorrectly rejecting the null (i.e. concluding there is a true difference when in fact there is none, or, a “false positive result”). On the other hand, the Type II error (β) is the error associated with incorrectly accepting the null hypothesis (i.e. there is a true difference in the population but the test fails to conclude any difference, or, a “false negative result”) (Akobeng 2016). Table 2.1 summarises the four possible outcomes:

		Decision	
		Accept H_0	**Reject** H_0
Reality	H_0 **False**	β Type II error	✓ Correct Decision
	H_0 **True**	✓ Correct Decision	α Type I error

Table 2.1: A summary of the four possible decisions in a trial

The statistical significance threshold (α) is a key component of the sample size calculation and must be decided in advance of the trial, set at the desired level of Type I error rate. Typically, a two sided α of 0.05 is used (Fisher 1932; Sterne 2001), implying there is 5% chance that the observed difference has been seen by chance under the null hypothesis. Despite this typical value often being used in practice, the choice lies with the researcher, and certain scenarios could warrant a different level of statistical significance.

Type II error rate (β) is commonly expressed in terms of power ($1-\beta$); the probability of detecting a significant result, given that there is indeed a true difference in the population of at least a given magnitude (Noordzij 2011). Power is also a key component in sample size calculations and needs to be decided on in advance, and is typically chosen between 80 and 90% (Julious 2004; Noordzij 2011).

Whilst it is not possible to completely eliminate Type I or Type II errors, they can be minimised by increasing the sample size. The larger the sample size, the closer the sample estimates of the treatment effect will be to the true population values and therefore less likely for either error to occur (Sedgwick 2014; Akobeng 2016). A trade-off between minimising error and a realistic sample size should be considered when planning a trial. Julious (2004) notes that decreasing the power from 90 to 80% only decreases the sample size by 25%, but doubles the Type II error (Julious 2004). A Type II error rate of 10% is strongly advocated in this tutorial paper.

2.3.1.2 Treatment effect

Prior to the start of the study, researchers may set up a trial with the aim to show an intervention is a quantifiable amount different to the comparator/placebo (Cook 2014). This difference is often referred to as the target difference, or the effect size, and is used as the anticipated treatment effect in sample size calculations (δ_{plan}).

Current evidence and retrospective data can inform the choice of δ_{plan}, to ensure that the planned treatment difference is 'realistic' (Cook 2015). Choosing a larger target difference will require fewer subjects. When planning a trial, a small sample size is attractive due to decreased recruitment time and required budget. However, being too optimistic can mean that the planned treatment effect is much higher than the true difference, and the study may fail to achieve statistical significance, because the true treatment effect is much smaller than the one in which the planned sample size can reliably detect (Biau 2008). Similarly, choosing too small an effect size could result in a very large sample size, and even if a statistically significant result is obtained at the end of the study, the effect may be too small to be clinically meaningful and therefore not provide sufficient evidence to change clinical practice.

The CONSORT Statement (2010) states that small effect sizes are more likely to exist in reality than large differences (Schulz 2010; Yusuf 1984). However, a trial should aim to detect not only a 'realistic' difference, but also an 'important' difference too (Cook 2015).

Obtaining a clinically important difference is essential in order to implement the investigational intervention into standard care, or for approval of a new drug. Enrolling subjects into a trial that has no hope of improving care by detecting a clinically important difference poses an ethical concern. Therefore a researcher should choose a target difference that is also clinically meaningful. The minimum value that is still regarded as clinically relevant is known as the Minimum Clinically Important Difference (MCID) (Noordzij 2011).

The target difference should aim to be both important, and realistic, and is therefore greater than or equal to the MCID (Cook 2015). In order to change standard practice however, researchers may need to show a larger treatment effect than MCID, to counteract associated development or implementation costs, or to beat market competition (Fukunaga 2014). Implications of misspecifying δ in the trial sample size calculation may lead to an over or under powered trial, and the trial may no longer be able to reliably answer the research question of interest.

2.3.1.3 Variance

In 1940, Dantzig showed that there exists no power function that is entirely independent of variance (σ^2) for a fixed sample size single sample t-test (Dantzig 1940). Five years later, Stein extended this to two sample t-tests, and to linear hypotheses (Stein 1945). Therefore variance plays a big role in the determination of sample size and needs to be reliably estimated to obtain accurate sample size estimations. Variance can be estimated using historical data, or pilot studies in the population of interest.

2.3.2 Practical considerations

2.3.2.1 Choice of study design

The choice of study design varies on a case by case basis. To simultaneously recruit patients to two or more treatment groups and observe outcomes in parallel are called parallel group trials. The sample size is dependent on the allocation ratio, r. There may be reasons to recruit more patients in a particular group, and allocation ratio can be adjusted. Total sample size (n) is the sum of the the sample size in each group ($n = n_A + n_B$) and $n_B = rn_A$, where r is the randomisation allocation ratio. This sample size is usually minimised when r=1, i.e. equal randomisation (Julious 2004).

The simplest and most efficient design is a 1:1 parallel group trial comparing two treatments with equal number of subjects in each group. Crossover designs have the added benefit, where all patients receive all treatments, resulting in patients being their own control (Mills 2009). The ability to incorporate within-subject differences in the analysis decreases the required sample size while achieving the same precision compared to a corresponding parallel-group trial (Li 2015). These designs are particularly useful for trials for chronic conditions, where the condition returns once the treatment has worn off (Mills 2009). For example, treatments for asthma can alleviate symptoms, but the condition itself is not cured. Therefore, once the effect of one treatment has worn off, another can be tested on the same patient.

Factorial designs can be used to compare two treatments simultaneously, and patients are assigned to a combination of treatments. Both main treatment effects of every treatment under investigation, and their interactions can be analysed. The effect of receiving both treatments may not equal the effect of the two treatment effects added together (Moser 2015). For instance, an outcome may increase by “a" on Drug X, and increase by “b" on Drug Y, but the effect of receiving both Drug X and Drug Y is not equal to “a+b". This implies an interaction exists, and can be measured using this type of design. Treatments with multiple levels are considered, denoted by “n x m". A factorial design investigating Drug X and Drug Y, each against a placebo, will result in a 2x2 design, as each treatment has two levels (receive Drug X: yes/no; receive Drug Y: yes/no). This results in patients being assigned to one of the following combinations: Placebo X and Placebo Y, Drug X and Placebo Y, Placebo X and Drug Y, Drug X and Drug Y (Table 2.2).

		Drug Y?	
		No	**Yes**
Drug X?	**No**	Group 1 Placebo X & Placebo Y	Group 2 Placebo X & Drug Y
	Yes	Group 3 Drug X & Placebo Y	Group 4 Drug X & Drug Y

Table 2.2: A summary of the four possible group allocation in a 2x2 Factorial Design clinical trial investigating Drug X and Drug Y

Other designs include cluster trials (where groups (e.g. GP surgeries) are randomised rather than individuals), standard GSDs (allowing for stopping early for efficacy or futility) and ADs (discussed in Section 2.4).

2.3.2.2 Type of hypothesis test

A superiority trial tests the hypothesis that one treatment is superior to another. This usually aims to show that an experimental treatment is superior to a control treatment, such as a placebo, the current standard practice, or an alternative treatment (Lesaffre 2008) as opposed to the null hypothesis, that the treatments are assumed to be the same.

Non-inferiority trials test the hypothesis that one treatment is not inferior to, or not ‘substantially worse than' another. These trials are often used to test an investigational treatment with an already active treatment (i.e., not placebo). It may be useful for showing a treatment is not inferior to a treatment, but can offer some advantage such as fewer side effects, fewer doses or cheaper (Hahn 2012).

Equivalence trials test the hypothesis that two treatments are 'clinically equivalent' against the null hypothesis, that population means are different between treatment groups. These trials are similar to non-inferiority trials in that they aim to show a treatment is not substantially worse than another. However, they also aim to show that the treatment is not substantially better either (Greene 2008).

Figure 2.1 illustrates the difference between superiority, non-inferiority and equivalence trials in terms of some difference: d_{Eq} representing the equivalences limit, and d_{NI} the non-inferiority limit (Julious 2004).

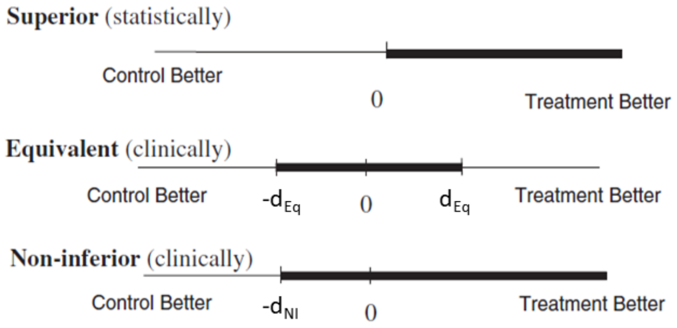


Figure 2.1: An illustration of the difference between superiority, equivalence and non-inferiority. (Julious 2004. Used with permission from John Wiley and Sons)

This thesis will focus only on superiority and non-inferiority hypotheses.

2.3.2.3 Type of data

In order to choose the correct sample size formula, the type of data collected in order to answer the primary outcome hypothesis needs to be chosen. For instance, the outcome may be collected on a continuous scale, such as blood pressure at 12 weeks, a rate or proportion, such as readmittance to hospital (yes vs no), or a survival outcome, such as time to disease progression.

2.3.3 Trials with Normal Data

2.3.3.1 Superiority Trials

For continuous data, a superiority trial tests if the means across the treatment groups are 'equal', against a hypothesis that the means differ by amount '*d*' (Julious 2004), chosen as the effect size the trial will detect (Section 2.3.1.2).

The null hypothesis (H_0) and the alternative hypothesis (H_1) for a two-sided superiority trial with means μ_A and μ_B in the control and intervention groups respectively can be written as:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

The minimum sample size required for Group A, with known variance, can then be calculated using Equation (2.1) (Flight 2016a):

$$n_A = \frac{(r+1)(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}})^2}{r\delta_{plan}^2} \tag{2.1}$$

where $\delta_{plan} = \frac{d_{plan}}{\sigma_{plan}}$ is the target effect size chosen for the superiority trial and Z_{1-x} represents the (1-x) percentage point of a standard Normal distribution.

2.3.3.2 Non-Inferiority Trials

Hypotheses for non-inferiority trials test that two treatments are within the non-inferiority limit d_{NI}, defined as the 'largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice' (Julious 2004). The null and alternative hypotheses for a non-inferiority study can be written as follows:

$$H_0 : \mu_A - \mu_B \leq -d_{NI}$$

$$H_1 : \mu_A - \mu_B > -d_{NI}$$

and the corresponding sample size for Group A can be calculated using Equation (2.2) (Flight 2016b):

$$n_A = \frac{(r+1)(Z_{1-\beta}+Z_{1-\frac{\alpha}{2}})^2\sigma^2_{plan}}{r(d_{plan}-d_{NI})^2}. \tag{2.2}$$

2.3.4 Trials with Binary Data

2.3.4.1 Superiority Trials

For a superiority trial with binary data to detect a difference in absolute risk d_{sup}, the hypothesis can be written as:

$$H_0 : \pi_A = \pi_B$$

$$H_1 : \pi_A \neq \pi_B$$

and the corresponding sample size can be calculated using:

$$n_A = \frac{\left\{Z_{1-\frac{\alpha}{2}}\sqrt{[(1+r)\overline{\pi}(1-\overline{\pi})]}+Z_{1-\beta}\sqrt{[r\pi_A(1-\pi_A)+\pi_B(1-\pi_B)]}\right\}^2}{rd^2_{plan}} \tag{2.3}$$

where

$$\overline{\pi} = \frac{(\pi_A+r\pi_B)}{(1+r)}. \tag{2.4}$$

The hypothesis and sample size calculation for instead detecting a difference in Odds Ratio (OR) can be written as follows:

$$H_0 : OR = 1$$

$$H_1 : OR \neq 1$$

and

$$n_A = \frac{6\left[Z_{1-\beta}+Z_{1-\alpha/2}\right]^2/(\log OR)^2}{\left[1-\Sigma_{i=1}^{2}\overline{\pi}_i^3\right]} \tag{2.5}$$

where $\bar{\pi}_1 = \frac{\pi_{1_A}+\pi_{2_B}}{2}$ and $\bar{\pi}_2 = 1-\bar{\pi}_1$. Note, Equations 2.3 and 2.5 are equivalent (Julious 2009).

2.3.4.2 Non-Inferiority Trials

For a non-inferiority trial with non-inferiority limit d_{NI}, the hypothesis can be written as (Julious 2009):

$$H_0 : \pi_A - \pi_B \geq d_{NI}$$

$$H_1 : \pi_A - \pi_B < d_{NI}.$$

When testing the absolute risk difference, the sample size calculation can be expressed as follows (Julious 2009):

$$n_A = \frac{\left(Z_{1-\alpha}\sqrt{\tilde{\pi}_A\left(1-\tilde{\pi}_A\right)+\tilde{\pi}_B\left(1-\tilde{\pi}_B\right)}+Z_{1-\beta}\sqrt{\pi_A\left(1-\pi_A\right)+\pi_B\left(1-\pi_B\right)}\right)^2}{\left(\left(\pi_A-\pi_B\right)-d_{NI}\right)^2}. \qquad (2.6)$$

Using Dunnett and Gent's method to estimate the variance under the null hypothesis, $\frac{\tilde{\pi}_A(1-\tilde{\pi}_A)}{n_A} + \frac{\tilde{\pi}_B(1-\tilde{\pi}_B)}{n_B}$, $\tilde{\pi}_A$ and $\tilde{\pi}_B$ can be expressed as follows:

$$\begin{aligned} \tilde{\pi}_A &= \frac{\pi_A + \pi_B + d_{NI}}{2} \\ \tilde{\pi}_B &= \frac{\pi_A + \pi_B - d_{NI}}{2} \end{aligned} \qquad (2.7)$$

When testing the difference in ORs, the hypothesis is written as follows:

$$H_0 : OR \leq d_{NI}$$

$$H_1 : OR > d_{NI}.$$

and the required sample size can be estimated using the following:

$$n_A = \frac{6\left[Z_{1-\beta}+Z_{1-\alpha}\right]^2}{\left[1-\sum_{i=1}^{2}\bar{\pi}_i^3\right]\left(\log(OR)-d_{NI}\right)^2}. \qquad (2.8)$$

2.3.5 Implications of underpowered trials

If prior information is limited in the planning stage, the resulting sample size may not be sufficient to answer the hypothesis in question. The CONSORT Statement (2010) states

“Reviews of published trials have consistently found that a high proportion of trials have low power to detect clinically meaningful treatment effects" (Schulz 2010)

Underpowered studies have sparked much debate in the trial community regarding the potential implications involved. A trial may not recruit enough subjects to fully answer the research question, and are susceptible to making false negative conclusions due to their low statistical power as a single study (Halpern 2005). Edwards *et al.* reason that some information is better than none at all (Edwards 1997) and these studies may still be ethical, as multiple underpowered studies may be combined in meta-analyses to improve power and estimate treatment effects (Halpern 2002). It should be noted however, that the trials included in the meta-analysis should have comparable research methods in order for the combined analysis to be useful (Halpern 2002).

In 2013, a large number of Cochrane reviews were examined and power per individual study was evaluated (Turner 2013). Their findings showed that underpowered studies contributed very little information at all to the meta-analysis when there were at least two sufficiently powered studies available (defined as $\geq$50% power to detect a 30% risk reduction). However, they also noted that in most Cochrane reviews, underpowered studies were the only studies available, highlighting the prevalence of underpowered studies. Despite so many underpowered studies available in literature, it is thought that this does not capture the full picture. Underpowered studies may not always be reported in the public domain and so will not be incorporated for later meta-analyses Griffiths 1997. Therefore, methods have been developed whereby the original sample size estimate may be modified during the study progression through incorporation of an interim analysis, taking into account accumulating data from the trial (Christy 2006). ADs using SSR methods could offer a potential solution for a trial to obtain the necessary power at the final analysis (see Section 2.4)(Chow 2011).

2.4 Adaptive Designs

ADs have gained much interest in the last 30 years (Bauer 2016a) due to their flexible nature and their potential for improved allocation of time and resources (Pallmann 2018).

ADs use the accumulated interim data at a pre-specified time-point in order to make a decision for the remainder of the trial (FDA 2010; FDA 2016). Decision rules for trial modifications should be pre-specified and adaptations should not be made on an 'ad hoc basis' to overcome poor planning at the design stage (Gallo 2006).

The term 'Adaptive Design' can cover a large number of trial designs. As this thesis focuses only on SSR, other individual designs will not be reported on in detail here. However, they can be broadly classified by four main rules; 'allocation rules' including adaptive randomisation designs, 'sampling rules' including SSR or drop the loser designs, 'stopping rules' including GSDs or adaptive treatment switching, and lastly 'decision rules' where changes can be made to the trial such as the primary endpoint, hypothesis, patient population or statistical methods (Mahajan 2010).

Despite much research on ADs in recent years, many researchers are still cautious in using these designs in practice (Pallmann 2018). Lack of experience is one major barrier in their implementation, and in some cases there is concern regarding bias and the interpretability of study results following adaptation (FDA 2010; Lin 2016).

In 2006, the Pharmaceutical Research and Manufacturers of America (PhRMA) working group recommended the use of a number of adaptive clinical trial designs, and advocate their implementation in industry, regulatory authorities and academia (Gallo 2006). There has been an increase in the use of ADs in recent years (Hatfield 2016). However, ADs are still not commonly used in clinical research (Hatfield 2016), and a number of logistical barriers must be overcome to fully achieve the benefits ADs have to offer (Coffey 2012). Kairalla *et al.* (2012) recommend developing “better methodology, infrastructure and software" in order to address the barriers opposing ADs currently (Kairalla 2012).

When reviewing ADs, Vandemeulebroecke (2008) recommends five key discussion points to consider; feasibility, validity, integrity, efficiency and flexibility (Vandemeulebroecke 2008). Any new methodology or design changes should consider these five points. This PhD focuses on enhancing methodology for just one type of adaptive design; SSR, described in Section 2.5.

2.4.1 The Cost of an Adaptive Design

With ever-increasing expenses associated with running a clinical trial (Shore 2012; Collier 2009), there is considerable interest to ensure trials are run cost-effectively. Adaptive designs can offer methods to allocate resources and time to trials in a cost-effective manner (Lin 2016; Chang 2016; Bauer 2016a). In the industry setting, ADs can speed up the time before product registration or marketing approval (Maca 2014).

Interim analyses can result in sample size savings if stopping boundaries are incorporated and the trial is allowed to terminate early. However, there is also a financial cost with carrying out interim analyses, additional Data Monitoring and Safety Committee (DMSC) meetings, and time allowed for the interim analysis to be carried out (Wassmer 2016; Chang 2016; Bauer 2016a; Mauer 2012). Furthermore, the number of subjects and recruitment time are directly linked to the cost of the trial and any sample size increase decided at the interim stage of an AD could incur huge costs for the trial team (Levin 2014; Koh 2017). However, it has also been shown that incorporating more interim analyses can yield a lower expected sample size (Koh 2017) and costs of recruiting additional patients should be carefully weighed against costs of interim analyses. Restricting the maximum possible number of patients ensures the trial is kept to a reasonable maximum cost set by the investigators (Bowden 2014). When sample size has been driven by logistical or budgetary issues as opposed to following a formula, these considerations may have already been taken into account (Wang 2012).

Complex designs may not be the most cost-effective solution (Lin 2016) and realistically, costs resulting from ADs can be extremely difficult to estimate in advance (Huskins 2018). The European Organisation for Research and Treatment of Cancer (EORTC) notes that neither an AD nor a standard GSD are useful when accrual is very quick and the time to the primary outcome is quick (Mauer 2012).

One additional logistical impact on the financial costs of an AD is the supply of the investigational product (Maca 2014). An increase in sample size requires an increase in resources and an added cost in distributing more product to sites after the interim analysis results. Any time delay between manufacturing additional products and their distribution should be carefully considered. Furthermore, capping the sample size can prevent highly significant findings that are clinically irrelevant. On the other hand, a trial that could result in early termination could mean a waste of resources if already manufactured and distributed to sites.

2.5 Sample size re-estimation

SSR is a form of AD whereby sample size can be recalculated during the study progression. Accumulating data can provide information about key parameters estimated prior to the start of the trial in the SSR. This design is attractive to researchers, especially when prior information is limited as it can avoid either an underpowered study, or an excessively large study due to over-optimistic or conservative prior estimates in the planning stage respectively (FDA 2010).

SSR can be carried out in either a blinded or an unblinded fashion. bSSR is when treatment group allocation is not revealed at the interim analysis and re-estimates nuisance parameters such as variance for a continuous outcome, or the probability of an event for a dichotomous outcome (Proschan 2005). Pooled variance can be used to highlight any significant difference in planning assumptions for variance, without unblinding the researcher to group allocation, and used to update the sample size in the second stage. For example, a superiority trial with a continuous endpoint is designed with a two-sided significance level $\alpha = 0.05$, 90% power and an assumed effect size of 0.3. There is uncertainty around the variance estimate, and is initially set as $\sigma^2_{plan} = 0.4$, but a bSSR incorporated to re-estimate this, using the methodology proposed by Wittes and Brittain in 1990 (Wittes 1990). The initial sample size is $n = \frac{2(1.960+1.282)^2 0.4}{0.3^2} = 93.4 \approx 94$ per arm. An interim analysis is conducted after 50 patients per arm have been observed, and the pooled variance estimate across treatment arms is $\hat{\sigma}^2_1 = 0.58$. The new sample size becomes $n^* = \frac{2(1.960+1.282)^2 0.58}{0.3^2} = 135.5 \approx 136$ per arm. Had the bSSR not taken place, too few subjects would have been recruited as the estimate of variance was too low, and the study would have been underpowered.

Re-estimation using just nuisance parameters can be performed in either a blinded, or unblinded manner. However re-estimating sample size using the treatment effect will always be performed using unblinded methods (Mütze 2018). A third approach, spanning the two key methods for SSR, can be considered. “Partially blinded SSR" updates nuisance parameters at the interim stage, but algorithms to calculate the new sample size required is not solely driven by the unblinded treatment effect (Pritchett 2015). “Partially unblinded SSR" refers to uSSR where group allocation is revealed as Treatment A or B only, and the researcher is unable to identify which specific treatment each refers to (Gould 2001).

Kairalla *et al.* (2012) note that researchers must exercise extreme caution when using treatment arm specific data, but these methods may be appropriate when access is fully restricted to very few people, and access only granted is absolutely essential (Kairalla 2012).

Stein (1945 & 1950) introduced a two-stage procedure for re-estimating sample size calculations, using stage one to provide the nuisance parameter information to update the sample size for stage two (Stein 1945; Stein 1950). A number of extensions to this methodology have been developed in both blinded and unblinded approaches (Wittes 1990; Birkett 1994).

Stein’s method however requires at least the statistician carrying out the interim analysis to be unblinded, as the pooled variance calculation (σ^2_{pooled}) depends on the sample mean of each treatment arm. Gould and Shih (1992) showed that a “lumped" variance estimator can be calculated using Equation (2.9) (Gould 1992), which does not require specific treatment group information:

$$\hat{\sigma}_L^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2 \tag{2.9}$$

where n_1 is the number of patients at the interim analysis, X_i are the individual responses, and $\bar{X}_1$ is the mean of all n_1 patients. However, it has been shown that this overestimates the true pooled variance (Proschan 2006). Additionally, if both $\hat{\sigma}^2$ and $\hat{\sigma}_L^2$ are known, the treatment effect at stage 1 ($\hat{\delta}_1$) can also be estimated by partitioning the sum of squares and therefore unblind any researchers with this knowledge (Proschan 2005). Gould and Shih overcome this with a method based on the EM (Expectation Maximisation) Algorithm, which has been seen as controversial in more recent research. The EM-approach is an iterative procedure, calculating conditional expectations of the unknown treatment allocations, computing maximum likelihood estimates (MLEs), updating conditional expectations and repeating (Friede 2002). Once the estimated within-group variance has stabilized, the procedure stops, as it is said to have converged. It was reported that this procedure estimates the true within group variance well, and was not influenced by the choice of initial values. However, Friede and Kieser (2002) showed that this is not the case: that the variance monotonically decreased when the initialisation parameter increased. Furthermore, Friede and Kieser also reported an inadequate convergence criterion, and uses simple randomisation - and so should be modified when the more commonly used block-randomisation is used (Friede 2002). Gould and Shih disputed these findings and presented further simulation work to support their work (Gould 2005). This procedure remains controversial (Waksman 2007).

Using total variance estimates for bSSR affects Type I error by only a negligible amount for parallel group superiority trials with continuous outcomes, and these methods are often able to maintain target power (Kieser 2003). This is also true for binary data (Friede 2004), longitudinal data (Wachtlin 2013) and count data (Friede 2010).

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 Statement considers bSSR methods to be generally well accepted by regulatory authorities (Lewis 1999). This corresponds with the initial 2010 Food and Drug Administration (FDA) draft guidance for industry, considering bSSR approaches as “well understood", as methods are able to control Type I error well (FDA 2010).

While blinded approaches not making any treatment effect comparisons are generally well accepted, there may be considerable uncertainty regarding the predicted treatment effect at the planning stage. Therefore, unblinded methods can be used in order to re-estimate the sample size based on the treatment effect at interim. These approaches are a lot more controversial due to requirement for knowledge of treatment allocation at the interim analysis (FDA 2010). Methods for uSSR will be discussed in more detail in Section 3.

uSSR methods were first classed as “less well-understood" according to the initial 2010 FDA guidelines, as these methods look at treatment group specific data. This gives the researcher the potential to increase the sample size if the treatment estimate was initially over-optimistic but still of clinical importance (Pritchett 2015). Knowing the specific group assignment of subjects at the interim analysis can introduce bias and can cause inflation of Type I error (false positive rate) (FDA 2010), making the methodology open to controversy. While there are statistical methods to control Type I error to adjust for decision making based on multiple looks at the data (Bauer 1994; Lehmacher 1999), operational bias, where unblinded results may cause analysts or investigators to conduct the trial differently, is a non-statistical source of bias with no statistical method of adjustment would be able to account for this FDA 2010. More recently however, the FDA have been generally more accepting of unblinded methods due to the increase in relevant research FDA 2016.

The PhRMA working group recommends routinely considering a bSSR in trials. They also state that unblinded methods exist and “can" be used (Gallo 2006). The European Medicines Agency (EMA) Committee for Medicinal Products for Human Use (CHMP) reflect this opinion, stating: “Whenever possible, methods for blinded sample size reassessment that properly control the Type I error should be used, especially if the sole aim of the interim analysis is the re-calculation of sample size. In cases where sample size needs to be reassessed based on unblinded data, sufficient justification should be made", for example when there is specific uncertainty regarding the true treatment effect, δ (European Medicines Agency 2007).

By unblinding at the interim analysis, bias is introduced and Type I error may be inflated. Therefore methods using uSSR tend to focus on methods for controlling for Type I error. This shall be discussed in more detail in Section 2.8.

The magnitude of the sample size re-estimation can be determined by a number of approaches. A calculation called CP, described in further detail in Section 2.6 can inform the researcher how many more patients would be needed to recover the planned power if the trial continues under the current trend observed so far (Gaffney 2017). One application using this method determines a range of CP values in which increasing the sample size is necessary, and applies a function to determine the new increased value of sample size at the interim time-point (Chen 2004; Mehta 2011; Jennison 2015). Regulatory agencies such as the FDA did not recommend a sample size reduction, and SSR rules are often constrained to an increase in sample size for this reason (Mehta 2011; FDA 2010). However, the FDA have been more accepting in recent years provided sufficient justification of the approach being used is also provided (US Food and Drug Administration (FDA) 2019). Knowledge of the sample size rule used and the final sample size can provide information regarding the observed interim treatment effect, and trialists may behave differently depending whether the treatment is "unfavourable" compared to "promising", whether knowingly or otherwise. A stepwise approach has been suggested to overcome this potential source of bias. Sample size can be increased in a step-wise manner to one of J discrete values, should a sample size increase be suggested (Wan 2015). For example, consider a trial that aims to recruit 200 patients. An interim analysis at 100 patients may suggest continuing to the originally planned 200 patients, or an increase to either 300, 400 or 500 patients, dependent on the interim results (J=3 discrete possible increased values). Now investigators who know the sample size rule used and the final planned sample size after the interim results may be able to work out a range (or ranges) in which the treatment effect lies, but is surrounded with more uncertainty.

2.6 Conditional Power

CP is the probability of rejecting the null hypothesis, H_0 at the final analysis, given the data observed at the interim analysis, and a future treatment effect for remaining patients $\tilde{d}$ (Mehta 2011). CP can be computed using the following formula (Denne 2001):

$$CP_{(\tilde{d})}(n|z_1) = 1 - \Phi\left\{\frac{c\sqrt{n} - z_1\sqrt{n_1} - \frac{n-n_1}{\sqrt{2\hat{\sigma}^2_{obs}}}\tilde{d}}{\sqrt{(n-n_1)}}\right\} \tag{2.10}$$

where z_1 is the observed test statistic at the interim analysis, $\hat{\sigma}^2_{obs}$ is the observed variance at the interim analysis, $\tilde{d}$ is the assumed future treatment difference, c is the final analysis critical value, n_1 represents the first stage sample size and n denotes the total sample size at the time of planning. For an unadjusted final analysis, the critical value c can be replaced with $z_{1-\frac{\alpha}{2}}$. Since the true effect δ is unknown, it is denoted in the CP equation by the assumption of the future treatment effect ($\tilde{d}$). Some common choices include assuming the the alternative hypothesis at the planning stage (δ_{plan}), or the current trend observed so far prior to the interim analysis ($\hat{\delta}_{obs}$) (Sully 2014).

If using the observed treatment effect as the future assumption, Equation 2.10 becomes:

$$CP_{\hat{\delta}_{obs}}(n|z_1) = 1 - \Phi\left\{\frac{z_{1-\frac{\alpha}{2}}\sqrt{n} - z_1\sqrt{n_1}}{\sqrt{(n-n_1)}} - \frac{z_1\sqrt{n-n_1}}{n_1}\right\}, \tag{2.11}$$

which is the same as the equation suggested by Mehta and Pocock (2011).

Denne (2001) suggests using the lower confidence limit of the observed treatment difference ($d_L(\alpha*) = \hat{d} - z_{\alpha*}\sqrt{(2\hat{\sigma}^2_{obs}/n_1)}$), the lower $\alpha*$ confidence limit for d (Denne 2001). Building from this, Herson *et al.* (2012) advocate the optimistic end of an 80% confidence limit of the observed treatment effect when a decision regarding futility is to be made at the interim time point (Herson 2012).

Some SSR designs use CP to aid decision rules, such as equating CP to planned power (Promising Zone deign), partitioning CP into discrete rules (Stepwise design), or as part of a combined objective function (Combination test design) (Mehta 2001; Liu 2016; Jennison 2015). These designs shall be discussed in detail in Chapter 3.

2.7 Sequential Designs

In 1943, Wald developed methods to analyse data sequentially for determining the outcome of an experiment sooner and with fewer observations (Wald 1947), known as the sequential probability ratio test. Due to its use in wartime logistics, the research wasn't available to the public until 1947. Using the same principles, a similar method known as the triangular test was developed in the late 1950s (Anderson 1960). Essentially, data is analysed sequentially to determine whether there is enough evidence for the acceptance of H_0, for H_1, or insufficient evidence to accept either. In the clinical trial setting, patients are enrolled until one of the hypotheses are accepted.

In the late 1970s, the methodology was extended to allow for grouped analysis of patients (Whitehead 1979). The number of interim analyses allowed is determined prior to the start of the trial. At each interim analysis, the hypothesis is tested and the outcome will terminate early due to either efficacy (the treatment has already been shown to be better than the comparison and recruitment can be stopped before the end of the trial) or futility (the treatment is very unlikely to be shown as effective and recruiting further patients is futile) (Bassler 2008), or neither hypothesis is accepted and recruitment continues until the next interim analysis, where the decision procedure is repeated.

Stopping boundaries are calculated according to the ‘alpha-spending function’ used (Section (2.7.1)) and the number of interim analyses planned.

While GSDs are considered as ADs themselves due to their use of interim analyses and decision rules for the continuation or early termination of the trial, for the purpose of this thesis, they will be considered as a separate design (standard GSD). This will better enable the distinction in comparisons between standard GSDs, Adaptive Group Sequential Designs (AdGSDs) and ADs.

AdGSDs incorporate a SSR in the final interim analysis of a standard GSD, provided the trial is not stopped at a previous interim time-point. Further applications of this method are known as Delayed Response Group Sequential Design (DRGSD), which include incorporating uncertainty about patients already enrolled at the interim analysis time-point, but who do not yet have outcome data available for analysis, known as pipeline patients (Hampson 2012).

2.7.1 Alpha Spending Functions

Due to the multiple testing nature of the hypothesis under investigation at each interim analysis, α will be inflated and so must be adjusted. To maintain α at the nominal level in a standard GSD, a number of different alpha spending functions have been suggested, and the choice lies with the investigator for each trial. Essentially, α is shared between all interim analyses in different weights, offering flexibility particularly when unforseen changes to the number or timing of interim analyses happens. When there is no wish to stop early unless there is overwhelming evidence, very small amounts of α may be spent in the interim analyses, and saved mostly for the final analysis. On the other hand, spending more

α earlier on may be advantageous if a treatment is expected to be beneficial and this effect can be shown earlier on. Critical values at each interim analysis depends on the method used, the overall significance level, power and number of interim analyses, K.

2.7.1.1 Pocock Method

In 1977, Pocock presented an alpha spending approach (Pocock 1977) of the form

$$\alpha_*(\tau) = \alpha ln(1+(e-1)\tau) \tag{2.12}$$

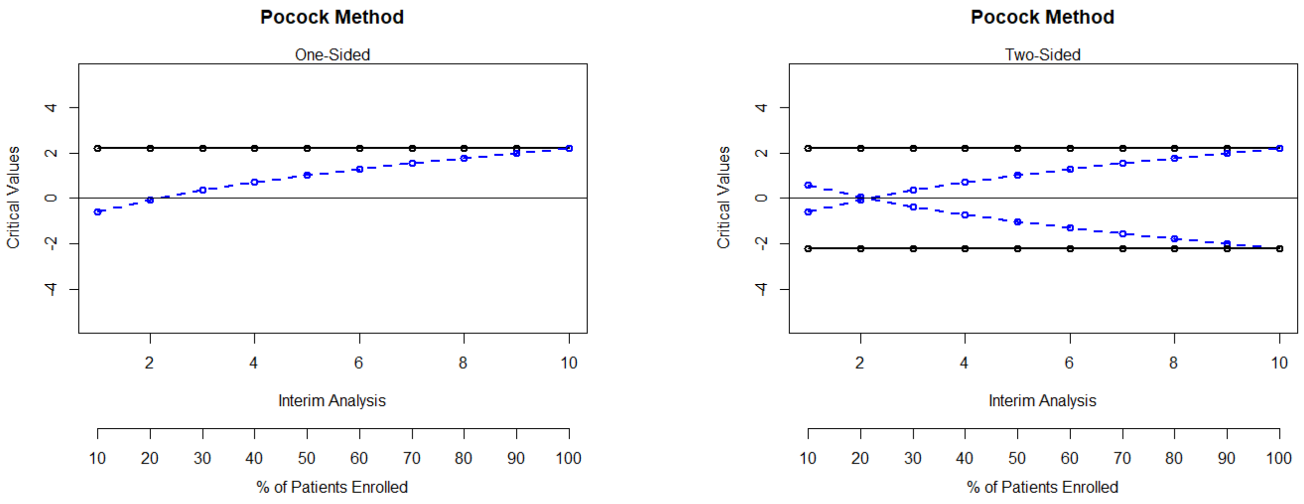
where $\tau = \frac{n_1}{n}$ represents the information fraction of patients at each interim analysis for continuous and binary endpoints, or the information fraction in terms of events for time-to-event endpoints (Proschan 1999). Alpha is 'spent' roughly equally, meaning that each interim analysis has roughly the same p-value cut-off point, where the trial would stop for efficacy.

The general procedure for the Pocock methodology is as follows (Chow 2008)

1. At interim analysis $k = 1,...,K-1$,
   - if $|Z_k| > C_{Pocock}(K,\alpha)$ the trial stops and H_0 is rejected;
   - otherwise, continue to group $k+1$
2. After stage K,
   - if $|Z_k| > C_{Pocock}(K,\alpha)$ then H_0 is rejected;
   - otherwise accept H_0

where $C_{Pocock}(K,\alpha)$ is the critical value using Pocock's Method for K planned interim analyses and significance level α. Z_k is the observed test statistic at the k^{th} interim analysis. The constants $C_{Pocock}(K,\alpha)$ can be found in tables available in the literature (Chow 2008).

This methodology can be extended to a futility boundary by using beta-spending functions, where the trial is stopped and H_0 is accepted, and are provided for illustrative purposes only (Pampallona 1994). Figure 2.2 illustrates both one-sided and two-sided efficacy and futility boundaries using the Pocock alpha spending function (Chow 2008).

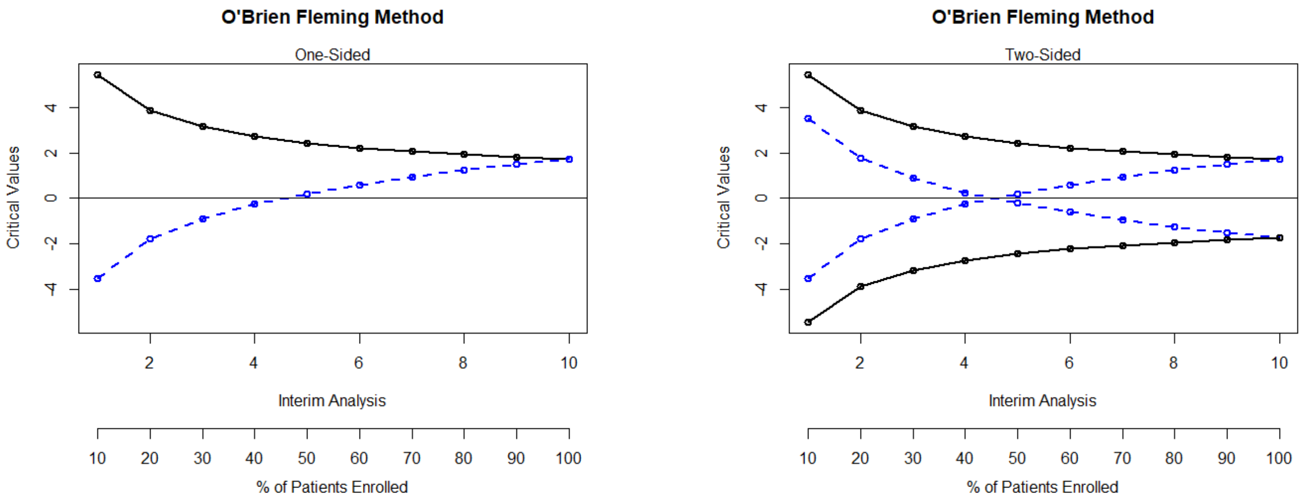


(a) *(b)*

Figure 2.2: Pocock stopping boundaries for 10 planned interim analyses with $1-\beta$*=0.9 and* α*=0.05. Figures (a) and (b) show one-sided and two-sided stopping boundaries respectively. If the black (solid) line is crossed, the trial stops for efficacy. If the blue (dashed) line is crossed, the trial stops for futility.*

2.7.1.2 O'Brien Fleming Method

While Pocock's method gives each interim analysis roughly the same weighting, O'Brien and Fleming designed stopping boundaries that give very small weighting to early interim stages, and larger weighting to later analyses. This means a very small *p*-value would need to be seen at an early stage in order to stop early. Therefore, less alpha is 'used up' in these early analyses, and there is more to use in later stages. This is particularly useful when there is no real wish to stop a trial early (O'Brien 1979).



(a) *(b)*

Figure 2.3: O'Brien Fleming stopping boundaries for 10 planned interim analyses with $1-\beta$*=0.9 and* α*=0.05. Figures (a) and (b) show one-sided and two-sided stopping boundaries respectively. If the black (solid) line is crossed, the trial stops for efficacy. If the blue (dashed) line is crossed, the trial stops for futility.*

The alpha spending function takes the form (Proschan 1999)

$$\alpha_*(\tau) = 2(1-\Phi(\tau^{-\frac{1}{2}}z_{\frac{\alpha}{2}})). \tag{2.13}$$

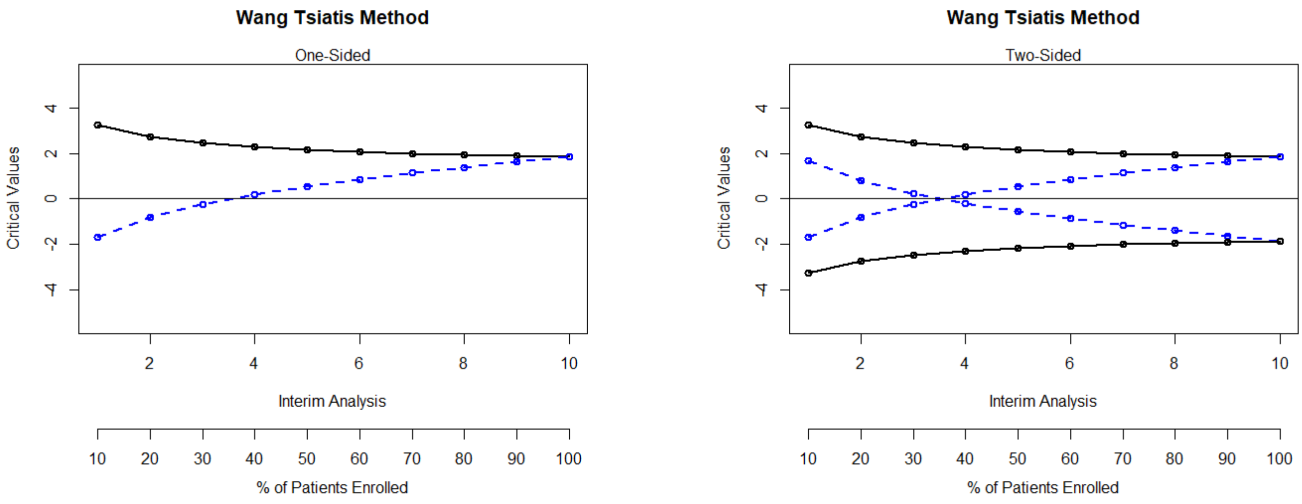
The general procedure for the O'Brien Fleming method is summarised below (Chow 2008),

1. At interim analysis $k = 1, \ldots, K-1$,
   - if $|Z_k| > C_{OBF}(K,\alpha)\sqrt{K/k}$ the trial stops and H_0 is rejected;
   - otherwise, continue to group $k+1$
2. After stage K,
   - if $|Z_k| > C_{OBF}(K,\alpha)$ then H_0 is rejected;
   - otherwise accept H_0

Similarly to the Pocock method, $C_{OBF}(K,\alpha)$ are critical values from the O'Brien Fleming method and are found in tables in the literature (Chow 2008).

Again, Pampaloon *et al.* extend this work to beta spending functions to allow the early termination of the trial for futility (Pampallona 1994). Figure 2.3 illustrates one and two sided efficacy and futility boundaries using the O'Brien Fleming method for 10 interim analyses.

2.7.1.3 Wang Tsiatis Method



(a) *(b)*

Figure 2.4: Wang Tsiatis stopping boundaries for 10 planned interim analyses with, Δ=0.25, $1-\beta$=0.9 and α=0.05. Figures (a) and (b) show one-sided and two-sided stopping boundaries respectively. If the black (solid) line is crossed, the trial stops for efficacy. If the blue (dashed) line is crossed, the trial stops for futility.

In 1987, Wang and Tsiatis a family of tests, depending on a new parameter Δ. The general procedure is as follows:

1. At interim analysis $k = 1, \ldots, K-1$,
   - if $|Z_k| > C_{WT}(K,\alpha,\Delta)(k/K)^{\Delta-\frac{1}{2}}$ the trial stops and H_0 is rejected;
   - otherwise, continue to group $k+1$
2. After stage K,
   - if $|Z_k| > C_{WT}(K,\alpha,\Delta)$ then H_0 is rejected;
   - otherwise accept H_0

This method is the same as the Pocock method when Δ=0.5, and the O'Brien Fleming method when Δ=0. Therefore a 'middle-ground' between the two methods discussed previously can be reached. Figure 2.4 illustrates efficacy and futility boundary for 10 interim analyses, α=0.05 and $\Delta = 0.25$.

2.8 Controlling Type I error

Chen *et al.* states "If the sample size recalculation is based on estimates of nuisance parameters such as within-group variance for Normal response or pooled event rate for a binary outcome, the Type I error rate will not be materially inflated. However, if the sample size recalculation is based on the observed treatment difference, the Type I error rate could be substantially inflated and an appropriate statistical adjustment may be needed to control it." (Chen 2004)

Therefore, a key focus in uSSR methodology has been around the preservation of Type I error at the point of the final analysis. Making a decision based on the accumulating data at the point of an interim analysis can inflate Type I error and a number of approaches have been developed in order to overcome this. These methods can be categorised into two broader frameworks: conditional error functions, and combination tests, both of which are described in this section.

2.8.1 Conditional error functions

2.8.1.1 Proschan & Hunsberger

As mentioned in Section 2.6, some SSR rules use CP to determine rules for SSR. Proschan & Hunsberger (1995) equate CP to a conditional error function $A(z_1)$ to determine the final critical value, based on observing interim results (Proschan 1995). Their approach allows flexible sample size and an adjusted final analysis, determining a critical value (c) and number of stage 2 observations required (n_2) using

$$CP_{\delta}(n_2, c|z_1) = 1 - \Phi\left(\frac{c\sqrt{2(n_1+n_2)} - z_1\sqrt{2n_1} - n_2\delta}{\sqrt{2n_2}}\right) = A(z_1) \qquad (2.14)$$

where $A(z_1)$ is an increasing function in the range [0,1] that determines how much Type I error to allow at the end of the study, given that $Z_1 = z_1$, and satisfies

$$\int_{-\infty}^{\infty} A(z_1)\phi(z_1)dz_1 = \alpha \tag{2.15}$$

where $\phi(z_1)$ is a standard normal density function. Equations 2.10 and 2.14 for CP are equivalent to each other, using $\tilde{d} = \delta$. Solving Equation (2.14) for c yields Equation (2.16):

$$c = \frac{\sqrt{n_1}z_1 + \sqrt{n_2}z_{A(z_1)}}{\sqrt{n_1 + n_2}} \tag{2.16}$$

, where $z_{A(z_1)}$ is the test statistic for the function $A(z_1)$ defined previously. An arbitrary choice of n_2 is allowed, provided this criteria is met (Posch 1999). When n_2 is close to 0 (i.e. almost no additional observations are required), the critical value, c, is close to z_1. For larger n_2, c approaches $z_{A(z_1)}$. Therefore, Type I error is controlled by only extending the trial if the interim *p*-value is sufficiently small. Furthermore, conditional error functions can adjust for the level of uncertainty on a case by case basis (Proschan 1995).

$A(z_1)$ can take on a number of classes of functions (Li 2002). Proschan and Hunsberger introduce the “circular conditional error function” of the form

$$A_{circ}(z) = \begin{cases} 0; & \text{if } z_1 < c_{fut} \\ 1 - \Phi\left(\sqrt{c_1^2 - z^2}\right); & \text{if } c_{fut} \leq z_1 < c_1 \\ 1; & \text{if } z \geq c_1 \end{cases} \tag{2.17}$$

where c_{fut} is the critical value chosen in which the trial would stop for futility (i.e. H_0 is accepted).

2.8.1.2 Linear conditional error functions

In 2001, Denne presented an alternative family of conditional error functions to the circular functions by Proschan & Hunsberger. The functions proposed by Denne are not required to be determined prior to the study start, as they depend on the sample variance ($\hat{\sigma}^2$) calculated at the interim analysis (Denne 2001). This improves unconditional power for cases where planned variance is smaller than the true variance (Denne 2001).

The "linear conditional error function" takes the form

$$A_l(z) = \begin{cases} 0; & \text{if } z_1 < c_{fut} \\ 1 - \Phi\left((a + bz_1)\right); & \text{if } c_{fut} \leq z_1 < c_1 \\ 1; & \text{if } z_1 \geq c_1 \end{cases} \tag{2.18}$$

Again, c_{fut} is the critical value that would stop the trial for futility and c_1 is the first stage critical value, determined at the interim analysis. Values a and b are determined by

$$a = \frac{\tilde{c}_2}{\sqrt{1 - \tau_1}} \tag{2.19}$$

$$b = -\sqrt{\frac{\tau_1}{1 - \tau_1}} \tag{2.20}$$

where τ_1 is the information fraction $\frac{n_1}{n}$ and $\tilde{c}_2$ is the required critical value in order to 'spend' the remaining α in the second stage given no sample size increase takes place (i.e. based on n_2 patients only).

2.8.1.3 Li, Shih & Wang

In 2002, Li, Shih *et al.* presented a modification to the work of Proschan & Hunsberger, with the purpose of relaxing the specific form of the circular or linear functions, $A(z_1)$. This results in a final critical value that is dependent only on the chosen design parameters, as opposed to the random first stage outcome (Li 2002).

The aim of the procedure is to determine the number of additional observations required (n_2) and a corresponding critical value (c) in order to maintain the overall Type I error rate at level α and conditional power at the final stage given the interim result at level $1 - \beta$, where α and β are pre-specified constants. Li *et al.* choose to select the first stage critical value, based on α_1, creating an 'alpha-spending' approach.

The overall Type I error can be expressed as:

$$\alpha = \alpha_1 + \int_{h_{lsw}}^{k_{lsw}} A\left(z_1\right) \phi\left(z_1\right) \mathrm{d}z_1 \tag{2.21}$$

where 'h_{lsw}' should be chosen such that $1-\Phi(h_{lsw}) > \alpha$. Specifically, the circular function becomes

$$\alpha = \alpha_1 + \int_{h_{lsw}}^{k_{lsw}} [1-\Phi\{\sqrt{k_{lsw}^2+z_1^2}\}]\phi(z_1)\mathrm{d}z_1 \tag{2.22}$$

and the linear function becomes

$$\alpha = \alpha_1 + \int_{h}^{k} \Phi(a+bz_1)\phi(z_1)\mathrm{d}z_1 \tag{2.23}$$

where k_{lsw} is a design constant dependent on α.

At the interim analysis, the final maximum critical value, C_{max}, can be determined from Equation (2.24):

$$1-\Phi(h_{lsw})-\alpha = \int_{h_{lsw}}^{k_{lsw_1}} \Phi\left[\frac{C_{max}(C_{max}+Z_{\beta_1}(u))-u^2)}{\sqrt{(C_{max}+Z_{\beta_1}(u))^2-u^2}}\right]\phi(u)du \tag{2.24}$$

where $k_{lsw_1} = \min(k_{lsw}, C+Z_{\beta_1})$. The number of additional observations can be calculated from Equation (2.25):

$$n_2 = \min\left[n_{max}, \left(\left(\frac{C_{max}+z_{\beta_1}}{\sqrt{z_1}}\right)^2-1\right)n_1\right], \text{ for } z_1 \in (h_{lsw}, k_{lsw}) \tag{2.25}$$

2.8.1.4 Bowden & Mander

Using the new sample size suggested by Li *et al.* can result in a large increase in sample size between stage 1 and stage 2. In 2014, Bowden and Mander suggested a slight adaptation to the LSW design, and go on to incorporate a maximum feasible sample size constraint (Bowden 2014).

Instead of choosing h_{lsw}, k_{lsw}, α and Z_{β_1} as with the standard LSW approach, a number of designs are now identified through the following algorithm:

1. Identify the fixed sample design with significance level α, power $1-\beta$ and hypothesised treatment effect δ_{plan}

2. Find all joint values $(h_{lsw}, k_{lsw}, Z_{\beta_1})$ consistent with α and $C = Z_\alpha$ from Equation

(2.25)

3. For each set of $(h_{lsw}, k_{lsw}, Z_{\beta_1})$ identified, find the minimum value of n_1 such that the unconditional power equals $1-\beta$.

Figure 2.5 illustrates this algorithm, identifying h_{lsw}, k_{lsw} and Z_{β_1}. The optimal first stage sample size, n_1, is determined by identifying the required power (e.g. 84% in this illustration, yielding an expected sample size $E(N)$ of 124, and corresponding design parameters k_{lsw} and l_{lsw}).

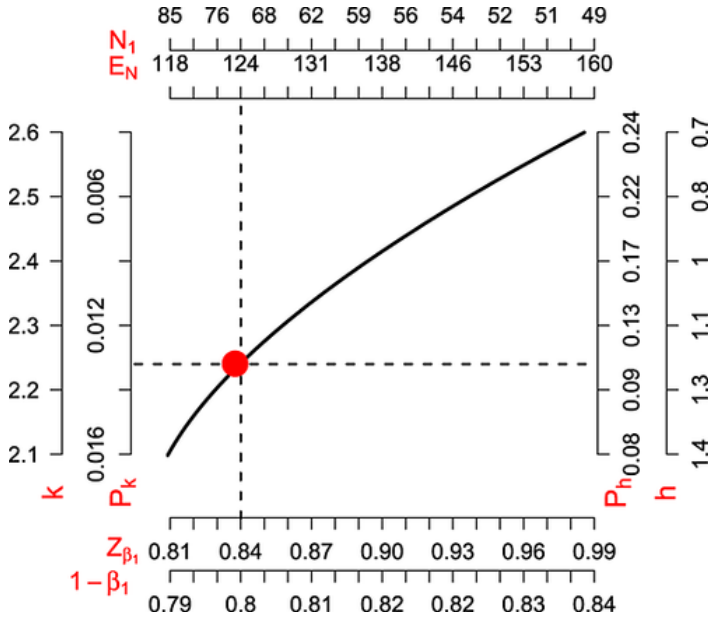


Figure 2.5: Illustration of the Reverse LSW algorithm for determining design parameters (Bowden and Mander 2014. Used with permission from John Wiley and Sons) (Bowden 2014)

This approach is also described when an additional restraint on the maximum sample size, n_{max} is to be specified.

The algorithm now becomes:

1. Identify the fixed sample size design with significance level α, power $1-\beta$, hypothesised treatment effect δ_{plan} and maximum value of $n_1 + n_2(z_1) = n_{T_{max}}$ say. Set $C = Z_\alpha$

2. Given $n_{max} = n_{T_{max}} - n_1$, find the set of values $(h_{lsw}, k_{lsw}, Z_{\beta_1}, n_{max})$ such that:

   (a) $(h_{lsw}, k_{lsw}, Z_{\beta_1}, n_{max})$ are consistent with α and $C = Z_\alpha$

   (b) n_1 is minimised, and

(c) unconditional power equals $1-\beta$

No illustration can be drawn for this reverse modified LSW approach, as not all parameters increase and decrease consistently.

2.8.2 Weighted combination tests

2.8.2.1 Fishers product combination test

In 1932, Fisher introduced a method of controlling Type I error based on the product of K independent p-values (Fisher 1932). H_0 is rejected if

$$\prod_{i=1}^{K}(p_i) \leq c_K \tag{2.26}$$

where c_K is a constant. As each p-value is uniformly distributed on (0,1),

$$-2\sum_{i=1}^{K} ln(p_i) \sim \chi^2 \tag{2.27}$$

with $2K$ degrees of freedom (Westberg 1985). Therefore it is possible to calculate c_K exactly, using:

$$c_K = e^{-\frac{1}{2}\chi^2_{2K,1-\alpha}} \tag{2.28}$$

where e is the exponential function, and $\chi^2_{2K,1-\alpha}$ is the $(1-\alpha)$ quantile of the Chi-Squared distribution with $2K$ degrees of freedom, and α is the significance level (Bauer 1994).

2.8.2.2 Modified combination test

In 1994, Bauer & Kohne extended the work of Fisher, specifically for testing an adaptive interim analysis such as incorporation of stopping boundaries. The general procedure for a two stage design using the modified combination test is described below (Wassmer 1998):

1. If $p_1 \leq \alpha_1$ (where $\alpha_1 > c_\alpha$), the trial stops for efficacy (H_0 if rejected).

2. If $p_1 \geq \alpha_{fut}$, the trial stops for futility (H_0 is accepted)

3. If $\alpha_1 < p_1 < \alpha_{fut}$, the trial continues to the second stage.

where α_{fut} is the futility boundary chosen by the investigator.

If the trial continues to the second stage, the value of p_1 may be used to redesign the study, such as modifying the number of subjects to recruit in the second stage (n_2). After stage 2, H_0 is rejected if $p_1 p_2 \leq c_\alpha$ While c_α can be computed directly using $c_\alpha = e^{-\frac{1}{2}\chi^2_{4,\alpha}}$, α_1 is determined iteratively, given an overall α level test, and using

$$\alpha_1 + \int_{\alpha_1}^{\alpha_{fut}} \frac{c_\alpha}{p_1} dp_1 = \alpha_1 + c_\alpha (ln\alpha_{fut} - ln\alpha_1) = \alpha \tag{2.29}$$

Bauer & Kohne also investigate the power loss of the combination test compared to the pooled sample test (both stage 1 and 2), as well as the additional power loss resulting from stopping the trial early. Compared to the pooled sample test, the largest loss of power resulting from the modified combination test occurred when the first stage sample size was 20% of the total sample size and a smaller overall significance level (0.01 compared to 0.05). Even so, this decrease was small (4.6% for $1-\beta = 0.5$, 4% for $1-\beta = 0.8$, and 2.7% for $1-\beta = 0.9$). These values were calculated using numerical integration. The loss in power due to early stopping compared to the classical test on the pooled samples is given as

$$P[\{p_1 p_2 \leq c_\alpha\} \cap \{p_1 \geq \alpha_{fut}\}] - P[\{p_1 p_2 \geq c_\alpha\} \cap \{c_\alpha \leq p_1 \leq \alpha_1\}]. \tag{2.30}$$

It can then be shown that an upper boundary for power loss from early termination is given by (Bauer 1994)

$$P[p_2 \leq \frac{c_\alpha}{\alpha_{fut}}] P[p_1 \geq \alpha_{fut}] - P[p_2 \geq \alpha_{fut}] P[\frac{c_\alpha}{\alpha_{fut}} \leq p_1 \leq \alpha_1] \tag{2.31}$$

Again, exact power loss can be calculated through numerical integration, and the power loss is relatively small compared to the pooled sample.

The main advantage of the modified combination test is that adaptations to the design can now be incorporated, such as early stopping for efficacy and futility, SSR, and even a change in hypothesis (Bauer 1994). However, Wassmer points out in 1998 that the “lack of concrete rules for calculating the sample size for the second part of the study” is a huge drawback of this method. However, he also states that by using CP procedures, this limitation can be overcome.

Wassmer compares the modified combination test to the work of Proschan and Hunsberger, described in Section 2.8.1.1, which also allows for a design modification at the interim stage but uses conditional error functions for Type I error control. While both methods are similar in terms of decision rules, power loss and expected sample size (Wassmer 1998), the modified combination test can handle more adaptation types than just SSR, such as a change in hypothesis at the interim stage (e.g. a dose response design, where only a subset of doses are explored in the second stage) (Bauer 1995). Furthermore, the methodology proposed by Proschan & Hunsberger is specifically for normal responses, with known variance and one interim analysis (Lehmacher 1999).

2.8.2.3 Fisher variance spending

In 1998, Fisher proposed a method of controlling Type I error, even when the interim analysis is unplanned by 'spending' a total variance of 1 between the two stages (Fisher 1998).

Denote the total sample size $n = n_1 + n_2$ in terms of the information fraction τ, so that $n = \tau n_1 + (1-\tau)n_2$, and let X_{Ai} and X_{Bi} denote the responses in two treatment groups A and B respectively. At the interim stage, the test statistic, Z_1 is defined as follows

$$Z_1 = \frac{\sum_{i=1}^{\tau n}(X_{Ai} - X_{Bi})}{\sqrt{n}} \sim N(\tau\sqrt{n}d, \tau). \tag{2.32}$$

Suppose the sample size is increased in the second stage by some inflation factor $0 < f < 1$, such that $n^* = \tau n + f(1-\tau)n$, where f is stage 1 data driven (Jennison 2003),

$$f = \frac{(\sqrt{(1-\tau)}z_{1-\beta} + z_{1-\frac{\alpha}{2}} - \tau\sqrt{n}\hat{\delta}_1)^2}{(1-\tau)^2 n\hat{\delta}_1^2}. \tag{2.33}$$

The second stage test statistic is defined as

$$Z_2 = \frac{f^{-\frac{1}{2}}\sum_{i=\tau n+1}^{n^*}(X_{Ai} - X_{Bi})}{\sqrt{n}} \sim N(\sqrt{f}(1-\tau)n\delta, 1-\tau). \tag{2.34}$$

However, no matter the value of f, Z_2 is independent of Z_1 and stage 1 data (Jennison 2003). The variance of Z_2 (1-τ) is the remainder of the variance to 'spend' in the second stage. Therefore the variance spending test statistic is

$$Z = Z_1 + Z_2 = \frac{\sum_{i=1}^{\tau n}(X_{Ai} - X_{Bi}) + f^{-\frac{1}{2}}\sum_{i=\tau n+1}^{n^*}(X_{Ai} - X_{Bi})}{\sqrt{n}} \tag{2.35}$$

Therefore, if H_0 is rejected if $Z > z_\alpha$, Type I error is maintained at level α.

2.8.2.4 Inverse normal combination tests

In 1999, Lehmacher and Wassmer introduce a method that combines standard GSDs with an adaptive sample size, even in the case of unknown variance and unequal sample sizes (Lehmacher 1999). Hedges and Olkin proposed the inverse normal method for combining p-values in 1985, where the test statistic is given by (Hedges 1985)

$$\frac{1}{\sqrt{k}}\sum_{k=1}^{K}\Phi^{-1}(1-p_k) \tag{2.36}$$

Lehmacher and Wassmer combine this test statistic, and classical group sequential boundaries in their proposed method, and show that Type I error is controlled at level α even for unequal sample sizes, n_k in each of the K stages. This is because the $\Phi^{-1}(1-p_k)$'s are independent and standard normally distributed. While an unweighted procedure can be used when each stage sample sizes are equal, it is recommended that a weighted procedure is used when sample sizes vary from stage to stage (Lehmacher 1999). However, the number of stages, K, must be specified in advance in order to use these methods.

2.8.2.5 Cui, Hung & Wang

In 1999, Cui, Hung and Wang proposed a weighted combination of test statistics (Cui 1999), since described in literature as the "CHW" method. Consider a trial with two stages (one interim analysis), with mean $\hat{\mu}_k$, number of observations n_k, k=1,2. Now consider scenario 1, where no sample size increase takes place at the interim stage. Conventional test statistics for each stage (1,2) and combined (C) are defined as follows:

$$\begin{aligned} X_1 &= \frac{\hat{\mu}_1\sqrt{n_1}}{\hat{\sigma}} \\ X_2 &= \frac{\hat{\mu}_2\sqrt{n_2}}{\hat{\sigma}} \\ X_C &= \frac{\hat{\mu}_C\sqrt{n}}{\hat{\sigma}} = \sqrt{\frac{n_1}{n}}X_1 + \sqrt{\frac{n_2}{n}}X_2 \end{aligned} \tag{2.37}$$

Now, consider scenario 2, where total sample size is increased from $n_C = n_1 + n_2$ to $n_C^* = n_1 + n_2^*$. In this scenario, conventional test statistics become:

$$\begin{aligned} X_1 &= \frac{\hat{\mu}_1\sqrt{n_1}}{\hat{\sigma}} \\ X_2^* &= \frac{\hat{\mu}_2^*\sqrt{n_2^*}}{\hat{\sigma}} \\ X_C^* &= \frac{\hat{\mu}_C^*\sqrt{n^*}}{\hat{\sigma}} = \sqrt{\frac{n_1}{n^*}}X_1 + \sqrt{\frac{n_2^*}{n^*}}X_2^* \end{aligned} \tag{2.38}$$

The CHW method uses the weights as though no sample size modification has taken place (i.e. Scenario 1), and the test statistic X_2^* relating to a sample size increase in scenario 2. It is however worth noting that X_2=X_2^* if no sample size increase occurs, and the final CHW test statistic will equal the conventional test statistic in this scenario. The CHW test statistic is (Mehta 2011; Cui 1999):

$$X_{CHW}^* = \sqrt{\frac{n_1}{n}}X_1 + \sqrt{\frac{n_2}{n}}X_2^* \neq \frac{\hat{\mu}_C^*\sqrt{n^*}}{\hat{\sigma}} \tag{2.39}$$

The trial is considered statistically significant if the CHW weighted test statistic exceeds the α level critical value, i.e. if $X_{CHW}^* > z_\alpha$.

The CHW statistic essentially downweights the second stage test statistic following the interim analysis. This has the added benefit that a SSR would not have to be specified in advance of the trial start and Type I error could still be controlled by using this method (Chen 2004). However, by assigning unequal weights to the two stages, the ‘sufficiency principle’ is violated. In a trial, every observation is equally informative. By unequally weighting the observations, this principle is infringed, and could lead to unreasonable results (Burman 2006; Basu 1969). In time-to-event trials, this situation is undesirable when the hypothesis is such that survival curves separate at first, and join together later on (Elsäßer 2014).

2.8.2.6 Dual test/Modified weighted method

Chen, DeMets and Lan (2004) suggest a modification to the weighted method described previously in Section 2.8.2.5. For the modified weighted test statistic approach having recruited n_1 patients, both the weighted statistic (X_W) and the unweighted statistic (X) must be greater than the critical value (z_α) for the final test to be considered statistically significant (Chen 2004). The Type I error for this method is actually less than the nominal α level. It has also been shown that the loss of power through using this method as opposed to the weighted approach is between 0 and 2% for up to a 100% increase in sample size increment. This method has also been termed as the 'dual test' in the literature (Shih 2016).

2.8.3 Unadjusted critical values

In 2004, Chen, DeMets and Lan ("CDL") showed that if sample size is only increased if CP is greater than 50%, then no adjustment needs to be made for the final analysis, and Type I error is still controlled at the nominal level, α (Chen 2004). Gao *et al.* (2008) and Mehta & Pocock (2011) further extended this range where Type I error is protected following SSR (Gao 2008; Mehta 2011). Researchers implementing this method must strictly adhere to these rules as any deviance may inflate Type I error. All rules must be pre-specified in advance. There is some concern from regulatory authorities in particular regarding compliance (Hung 2016a). This method is advantageous to researchers due to its comparatively straightforward statistical approach, and the added benefit that every observation is treated as equally informative to the final analysis (Mehta 2016a). This is also referred to as the 'one person, one vote' property in the literature. This method of Type I control, and the SSR rule used in the promising zone framework is discussed in more detail in Section 3.5.1.

2.8.4 Comparison

There have been a number of comparisons of the various methods to control Type I error in the literature. In 1999, Posch & Bauer performed a direct comparison between two combination test designs and showed that, in terms of CP, the modified approach performed better than the inverse normal approach for small values of z_1, and less well for large values of z_1 (Posch 1999). The inverse normal approach has the added benefit of using an unadjusted critical value for the final analysis over the two stages, and allows for uncertainty in variability. The modified weighted combination test and the circular conditional error function perform almost identically to each other in terms of power and expected sample size, but the modified approach also allows for additional adaptations (Wassmer 1998). In 2004, a method to find optimal conditional error functions was presented, which minimised expected sample size, and showed they obtained greater sample sizes for moderate treatment effects, and smaller sample sizes for small and large effect sizes (Brannath 2004).

Shi, Li & Wang compared four methods in 2016; CHW, Dual test, LSW and Promising zone. They note that the CHW method does not constrain a particular SSR rule, as it is not based on z_1 and therefore the choice of n_2 remains flexible. However, the choice of n needs to be planned carefully as it plays a vital role in the weighted test (Shih 2016). The Dual test was found to be less powerful than CHW, unless the SSR method was chosen specifically to match power, such as the methods of Burman and Sonesson (Burman 2006). The methods of Burman and of Mehta are opposing rules, with Burman's method losing power at the points in which Mehta's method would choose to increase sample size. Finally, the LSW method allows for a non-binding futility boundary, which could be advantageous particularly in industry settings (Shih 2016).

Whilst the thesis does not extend the work of these methods, they provide the understanding of the frameworks of the uSSR designs that are reviewed in detail in the next chapter, and the benefits and limitations provided by viewpoints in the literature. The promising zone design is able to control Type I error by only increasing sample size if z_1 falls between the lower and upper boundaries. An alternative approach however, has been presented using inverse normal combination tests, and is able to increase at any value of z_1, but use an inverse normal combination test approach to control Type I error. These designs will be discussed in more detail in Chapter 3.

2.9 Summary

This chapter presents the key statistical concepts that form the basis for the work of this thesis. Formulae for fixed sample size calculations have been presented for normal and binary data and superiority and non-inferiority hypotheses. Adaptive designs and the motivation behind blinded and unblinded SSR methods have been introduced. CP calculations can be used to inform a SSR according to some pre-specified rule, and the underlying assumptions for the future treatment effect have been discussed.

Adaptive designs can be beneficial to the future of clinical trials due to their flexible nature and ability to decrease time and resources. However, if multiple looks at the data occur, α inflation must be considered and adjusted for as appropriate. Methods to control Type I error have been presented and discussed, and this will aid the understanding of the frameworks discussed later in the thesis.

Chapter 3 presents a systematic review of the methodology of the 'Promising Zone' design and trials that have implemented this design. Alternatives to the 'Promising Zone' design for uSSR that are also based on CP calculations are discussed and areas of research deficit are summarised, which will be used to justify the aims of this thesis and the research question.

3 | Literature review

3.1 Introduction

Section 2.5 introduced the concept of uSSR, and background information on sample size calculations and control of Type I error. The promising zone design has been identified as a relatively straightforward approach to uSSR. A key motivation behind the PhD thesis is to ensure AD simplicity to promote the uptake of these designs where appropriate. This chapter systematically reviews the literature, focusing on the promising zone approach for this reason.

The aim of this chapter is to evaluate the current knowledge and usage of promising zone methodology in current trials, and aims to comprehensively synthesise and summarise the current literature available on promising zone methodology used in uSSR. More recent alternatives to this design will also be discussed. The review is split up into two sections; the first to capture research into the promising zone design, and the second to capture trials using this methodology in practice. Benefits and limitations of this design will be discussed, and the development of alternative uSSR designs in the literature will also be presented. The ultimate aim of this chapter is to inform the PhD research to be carried out in the remainder of the thesis.

3.2 Aims

The aim of this chapter is to systematically review the literature regarding methods of implementing a promising zone design in a clinical trial, as well as the current usage of the promising zone design in practice. Specific aims of the literature review include:

1. Review the methodology of the promising zone design, and highlight any advances in methodology since the original publication in 2011

2. Highlight both benefits and limitations of this design, and address responses from other statisticians/trialists to comparable designs
3. Summarise current uptake of promising zone implementation in recent trials
4. Identify related areas that would benefit from further research

The literature was systematically reviewed in order to find the current knowledge base on the following topics:

1. What is currently known about promising zone methodology?
2. When is this methodology appropriate to use?
3. What alternatives to promising zone design exist, and what advantages do they offer over the promising zone design?
4. Where has the promising zone design been implemented in practice?

3.3 Search strategy

3.3.1 Inclusion/Exclusion Criteria

Methodology Specific Criteria:

Inclusion Criteria:

- Literature including or mentioning promising zone methodology as a method for SSR, with or without case studies.

Exclusion Criteria:

- Insufficient mention of promising zone methodology (e.g. promising zone only mentioned in reference list, or mentioned once with no further details)

Trial specific Criteria:

Inclusion Criteria:

- Randomised controlled trial
- Use or plan to use promising zone methodology during the progress of the trial

Exclusion Criteria:

- Alternative SSR methodology used (i.e. no Promising Zone methodology)
- No details found of the SSR methodology used

3.3.2 Restrictions

The database search did not include any time restrictions, and includes all publications available up to 31st December 2018.

This review limited the search to publications or translations of publications available in the English Language.

3.3.3 Search strategy terms

3.3.3.1 Database search

A preliminary search was conducted in order to find out what terminology was used in the literature. A large number of variations were found for “sample size re-estimation", and “promising zone" methodology was sometimes described but not captured using standardised terminology. In light of this preliminary search, the following three search strategies were implemented individually and then combined, in order to find all relevant records.

1. “sample size re-estimation" OR “sample size reestimation" OR “sample size adjustment" OR “sample size readjustment" OR “sample size modification" OR “sample size recalculation" OR “sample size reassessment" OR “*creased sample size" OR “*crease in sample size" OR “adaptive sample size"
2. “promising zone" OR “promising region"
3. “promising" AND “results" AND “conditional power"

Final search: (1) AND (2 OR 3)

Online databases searched included Pubmed, Web of Science, Cochrane Database, CINAHL, OVID (including MEDLINE and PsychINFO) and clincaltrials.gov. Details of adaptations to the search strategy for each database are listed in Appendix A. Due to the structured reporting of clinical trials using online databases, SSR methods were rarely described in the short summary. For this reason, the clinicaltrials.gov database was searched only for strategy 1. Once a trial was identified as using SSR, further trial literature (e.g. full protocol, Statistical Analysis Plan (SAP), conference presentations or final publications) were sought and included only if promising zone methodology was implemented/planned to be implemented, and trial co-ordinators were contacted if necessary.

3.3.3.2 Pearl growing

Citation pearl growing, introduced by Hawkins and Wagers in 1982 (Hawkins 1982), is a method of searching backwards, starting with a relevant publication and finding other articles that have cited the publication (Schlosser 2005). Due to the lack of standardised terminology for promising zone methodology, even a comprehensive database search would not be able to identify all records. Therefore, a pearl growing technique was implemented in addition to the database search, using Mehta and Pocock's key paper as a starting point (Mehta 2011). Web of Science and google scholar were used to identify records that had cited their work.

3.3.3.3 Grey literature

In order to fully answer the primary objectives and fully understand when and how this methodology is used, a search for grey literature was also undertaken. FDA, EMA, The National Institute for Health and Care Excellence (NICE), PSI and Cytel websites were all searched for any documentation or other resources related to promising zone. Grey literature will inform both methodological records, or trials, as appropriate.

3.4 Search results

3.4.1 Database Search

A flowchart to describe the information found using the methods described above to search for literature is shown in Figure 3.1. Six databases were searched according to the search strategy described in Section 3.3.3.1, identifying a total 75 records. The key promising zone paper by Mehta and Pocock (Mehta 2011) had been cited by 215 papers, found through Web of Science and Google Scholar. No further records were identified through EMA or NICE websites. However, 4 records were discovered via FDA website, 9 through Cytel and finally an additional 15 through PSI. In total 324 records were identified, of which 133 were found to be duplicates. Of the 191 unique records, 63 were excluded for the following reasons: Insufficient mention of promising zone (n=32), Access issues (n=9), Trials where promising zone was not the SSR method used (n=17), Not an RCT (n=2) and no English translation available (n=3). The remaining 122 records were divided into either the methodology review (n=101) or the trials review (n=21). The two reviews are presented in detail in Sections 3.5-3.17.1 (methodological review) and 3.17 (trials review).

3.5 Promising Zone

3.5.1 Development of the 'Promising Zone' Framework

As discussed in Section 2.8, there has been much discussion on methods for controlling Type I error in ADs. However, Chen *et al.* (Chen 2004) showed in 2004 that conventional hypothesis tests could be carried out without inflating Type I error as long as an increase in sample size only occurs if the results at the interim analysis can be shown to be 'promising'. Chen *et al.* deemed the treatment effect to be promising if the CP assuming a future treatment effect based on the current trend was greater than 50%.

In 2008, Gao *et al.* extended the work of Chen *et al.* to expand the range of CP values within a K-stage group sequential design in which a researcher would be able to increase

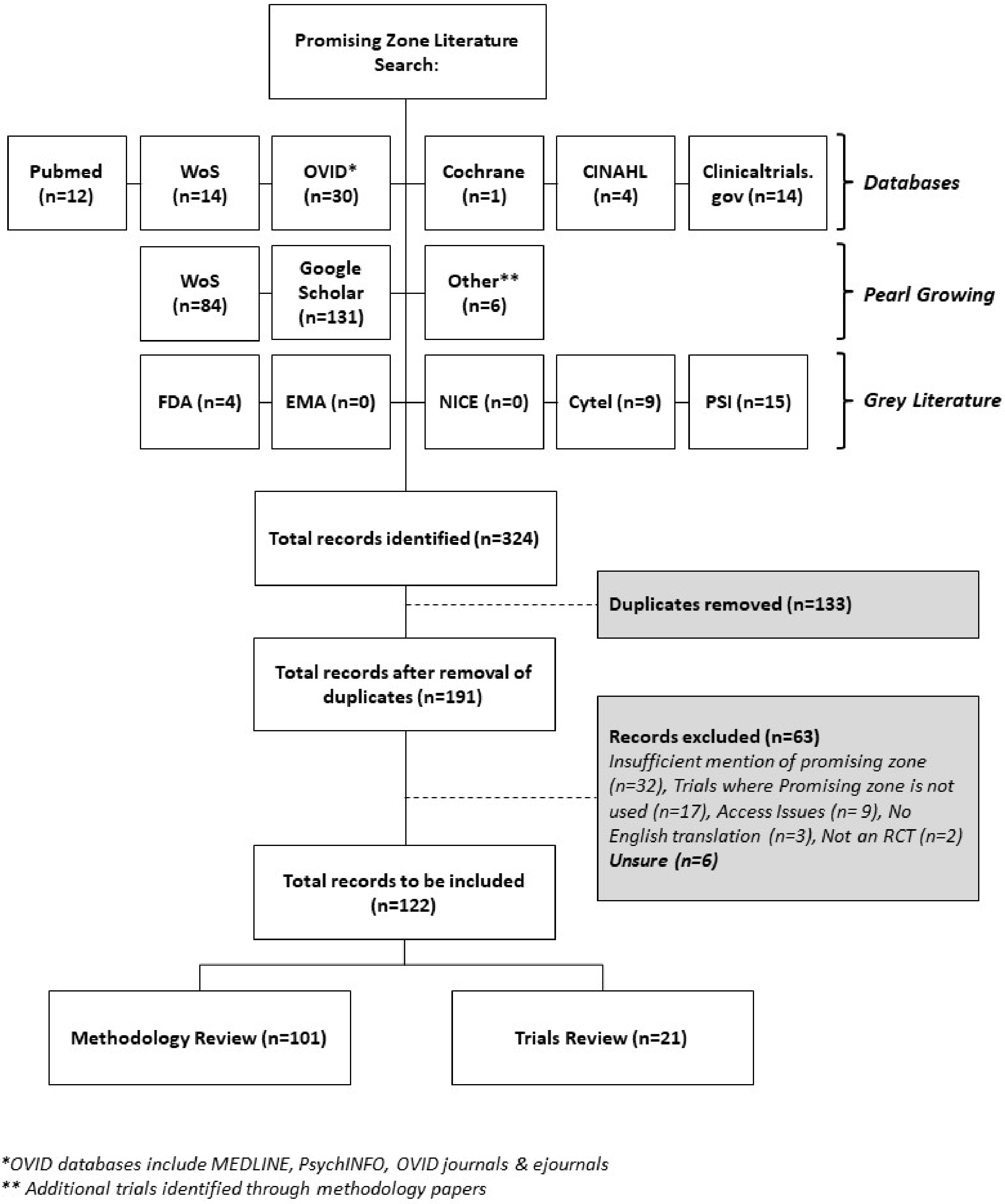

Promising Zone Literature Search:
Pubmed (n=12)
WoS (n=14)
OVID* (n=30)
Cochrane (n=1)
CINAHL (n=4)
Clinicaltrials. gov (n=14)
Databases
WoS (n=84)
Google Scholar (n=131)
Other** (n=6)
Pearl Growing
FDA (n=4)
EMA (n=0)
NICE (n=0)
Cytel (n=9)
PSI (n=15)
Grey Literature
Total records identified (n=324)
Duplicates removed (n=133)
Total records after removal of duplicates (n=191)
Records excluded (n=63)
Insufficient mention of promising zone (n=32), Trials where Promising zone is not used (n=17), Access Issues (n= 9), No English translation (n=3), Not an RCT (n=2)
Unsure (n=6)
Total records to be included (n=122)
Methodology Review (n=101)
Trials Review (n=21)
*OVID databases include MEDLINE, PsychINFO, OVID journals & ejournals
** Additional trials identified through methodology papers

Figure 3.1: Flowchart of Promising Zone literature search, detailing numbers included in each review and reasons for exclusion

the sample size, and still use a conventional final analysis without inflating Type I error (Gao 2008). Mehta and Pocock (2011) (Mehta 2011) further built on this work and made this design more accessible to researchers wishing to implement this in a two-stage design. Specific lower boundary CP values for the promising zone under a number of conditions such as maximum increase allowed, and the timing of the interim look were tabulated. The methodology of this paper is now described (Mehta 2011).

For a two stage design testing superiority of a continuous outcome in two groups, with significance level α and power $1-\beta$, the initial sample size (n) can be calculated using methods described in Section 2.3.3. After n_1 patients have been recruited, a decision is made to either keep going as planned and stop the trial after the original n patients have been recruited, or to increase the sample size according to the SSR rule, described later in this section. If the decision is to increase the sample size, a further n_2^* patients are recruited, such that

$$\begin{aligned} \text{Original sample size } &: \ n = n_1 + n_2 \\ \text{Increased sample size } &: \ n^* = n_1 + n_2^* \end{aligned} \tag{3.1}$$

At the interim analysis, CP assuming that the future treatment difference ($\tilde{d}$) is indeed the one estimated at the interim analysis ($\hat{d}_{obs}$) (i.e. under the current trend) is calculated according to Equation (2.11). This is equivalent to the conditional probability that the null hypothesis H_0 is rejected at the end of the trial, given that $Z_1 = z_1$ (Bauer 2006).

CP values may lie in one of the three pre-specified zones: unfavourable, promising, and favourable.

A result in the unfavourable zone indicates that the treatment effect is deemed 'disappointing', and it is simply not worth the increase in sample size to try and recover CP. Therefore, the sample size continues to the originally planned sample size (n). A result in the favourable zone indicates that the treatment difference is deemed 'sufficiently favorable' (i.e., $\hat{d}_{obs}$ is either greater than, equal to, or slightly smaller than the initial estimate of d_{plan} at the planning stage). No sample size increase is required, and the sample size therefore continues to the originally planned sample size (n). A result in the so-called promising zone indicates the treatment is neither disappointing, nor sufficiently favourable. The treatment effect is lower than expected, but not so low that CP cannot be recovered by a reasonable sample size increase. Therefore, the total sample size increases to n^*, having observed interim data. The increase in sample size is evaluated by the number required to ensure CP (given z_1 and $\hat{d}_{obs}$) equals $1-\beta$. In practice, logistical issues such as limited budgets or feasibility may mean that a maximum cap must be placed on the total sample size in the trial. For instance, increasing the sample size ten-fold may not be feasible in many cases, particularly if the increase in power in doing so may be small. Additionally, a cap in maximum sample size could prevent a statistically significant yet clinically irrelevant result. For this reason, researchers may place an upper limit constraint on the total sample size, n_{max}. Denoting the sample size required to maintain CP at the interim analysis as n_{req}, the new total sample size with a maximum sample size constraint can be described as

$$n^* = min(n_{req}, n_{max})$$

where

$$n_{req} = n_1 + \left[\frac{n_1}{z_1^2}\right]\left[\frac{z_{1-\frac{\alpha}{2}}\sqrt{n} - z_1\sqrt{n_1}}{\sqrt{n-n_1}} + z_{1-\beta}\right]^2 \tag{3.2}$$

So far, the three zones have only been described in terms of CP. These zones can also be converted to the z_1 scale, or indeed the scale $\frac{\hat{d}_{obs}}{d_{plan}}$, due to the following relationship (Mehta 2011):

$$\frac{\hat{d}_{obs}}{d_{plan}} = \left[\frac{z_1}{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}\right]\sqrt{\frac{n}{n_1}} \tag{3.3}$$

Figures 3.2 and 3.3 illustrate the key differences between the methodology of Chen *et al.* (2004) (Chen 2004) and Mehta and Pocock (2011) (Mehta 2011) in the development of the Promising Zone design.

3.5.2 Assumption of Future Treatment Effect

The choice of future treatment effect is a controversial subject, but an important decision for the CP calculations. In the literature, three main assumptions have been proposed; assuming the alternative hypothesis (H_1), the null hypothesis (H_0), or the current trend of data observed

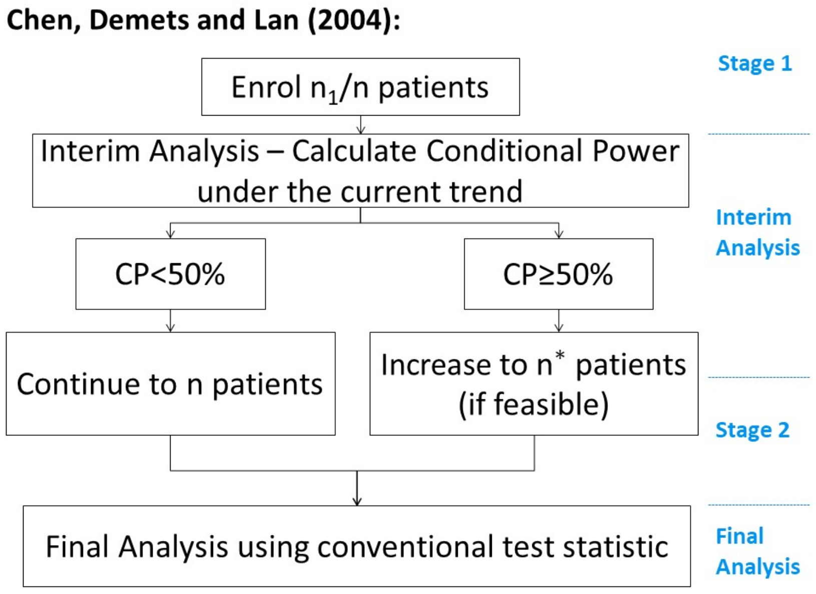

Chen, Demets and Lan (2004):
Stage 1
Enrol n_1/n patients
Interim Analysis – Calculate Conditional Power under the current trend
Interim Analysis
CP<50%
CP≥50%
Continue to n patients
Increase to n^* patients (if feasible)
Stage 2
Final Analysis using conventional test statistic
Final Analysis

Figure 3.2: Methodology for SSR proposed by Chen et al. (2004)

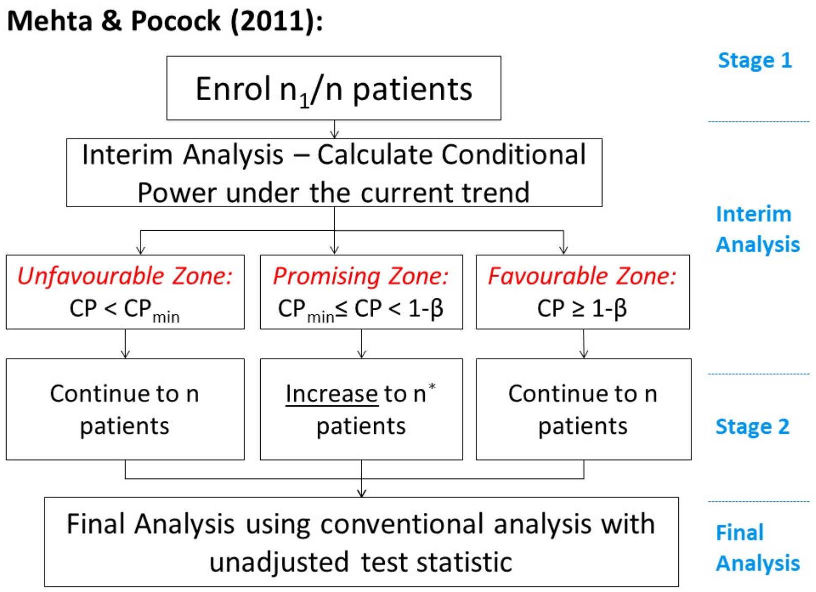

Mehta & Pocock (2011):
Stage 1
Enrol n_1/n patients
Interim Analysis – Calculate Conditional Power under the current trend
Interim Analysis
Unfavourable Zone: $CP < CP_{min}$
Promising Zone: $CP_{min} \leq CP < 1-\beta$
Favourable Zone: $CP \geq 1-\beta$
Continue to n patients
Increase to n^* patients
Continue to n patients
Stage 2
Final Analysis using conventional analysis with unadjusted test statistic
Final Analysis

Figure 3.3: Mehta and Pocock's general method

so far. While these have been used in practice to estimate the true value of CP, not all possible trajectories of the future data are described by just these three options (Herson 2012).

Glimm (2011) states that using the current trend assumption in the CP calculation yields an unstable estimate of the true CP value and argues that it is therefore unwise to base a SSR on this estimate.

As previously described, the cumulative Wald statistics are defined as:

$$Z_j = \frac{\hat{d}_{obs_j}}{\mathrm{se}\left(\hat{d}_{obs_j}\right)}, \quad j = 1,2 \tag{3.4}$$

where $\mathrm{se}\left(\hat{d}_{obs_j}\right) = 2\hat{\sigma}_{obs_j}/\sqrt{n_j}$. Now, at the interim analysis, the CP is calculated given the data observed so far (i.e. given that $Z_1 = z_1$). Now CP based on a future treatment effect under the current trend, given in Equation (2.11), is a function of z_1. Therefore, the estimated treatment difference, $\hat{d}_{obs}$ is effectively used twice. Any random deviation of $\hat{d}_{obs}$ from its true value d could consequently affect the value of CP (Glimm 2012; Bauer 2006).

Mehta and Pocock argue that the point of the CP calculation is merely to indicate whether results are ‘promising’ or not, and not to estimate d with precision (Mehta 2011).

Pepe and Anderson (1992) (Pepe 1992) and Herson *et al.* (2012) (Herson 2012) recommend considering the optimistic end of a confidence interval for the future trend for the purposes of stopping a trial for futility, provided the drift parameter throughout the life of the trial is believed to be constant. It has also been suggested for a DMSC to look at multiple future treatment effect assumptions, or to use a combination of the current trend and the planned treatment effect (Herson 2012; Bauer 2006). The current literature suggests that more research into this concept would be beneficial and the effect of various future treatment assumptions for the purposes of a SSR is needed to understand this in more depth.

3.6 Incorporating stopping boundaries

Chen *et al.* (2004) discuss the incorporation of stopping boundaries with their design; increasing the sample size if CP $\geq 50\%$, using the modified weighted statistic at the final analysis (Section 2.8.2.6). They state that Type I error is decreased when the addition of a futility stopping boundary is incorporated; stopping the trial if CP falls below a pre-specified lower limit, c_{fut}. They suggest three approaches in which this decrease in Type I error rate can be used. The first approach is to decrease the final critical value in the analysis such that Type I error is controlled at level α. For a stopping boundary of CP $\leq 10\%$, or equivalently $Z^{(n_1)} < 0.61$, the critical value to be used in the final analysis is calculated using the following:

$$\Pr\left(Z^{(n_1)} \geqslant 0.61, Z > c_{final} | \hat{d}_{obs} = 0\right) = \alpha. \tag{3.5}$$

The second approach is to also incorporate an efficacy boundary at the interim analysis, in which the trial is stopped at the interim analysis if the pre-specified upper boundary, c_{eff}, is crossed. c_{eff} is calculated using the following formula:

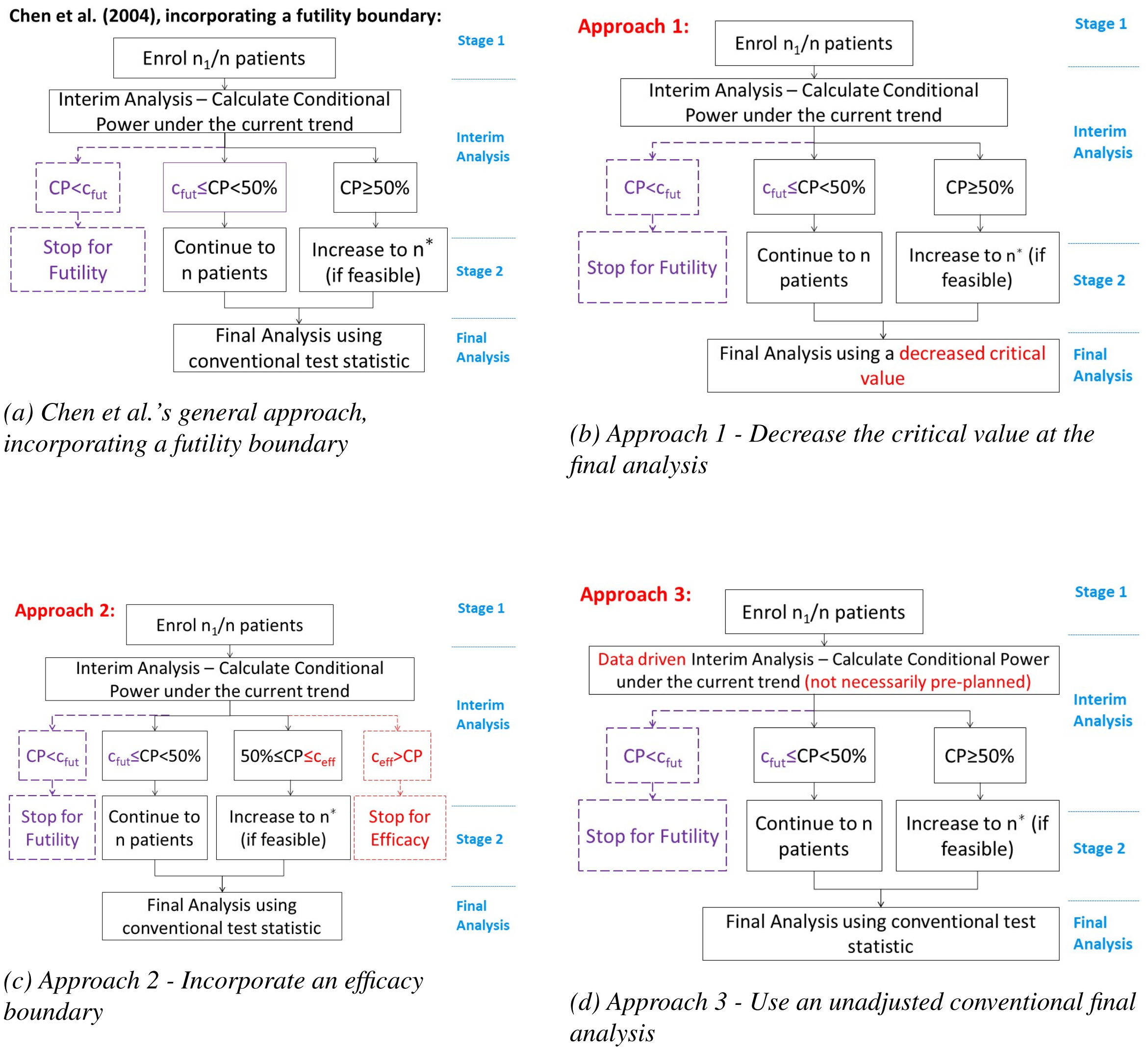
$$\Pr\left(Z_1 > c_{eff} | \hat{d}_{obs} = 0\right) + \Pr\left(0.61 \leqslant Z_1 \leqslant c_{eff}, Z > Z_{1-\alpha} | \hat{d}_{obs} = 0\right) = \alpha. \tag{3.6}$$

Finally, the third approach is to use an unadjusted final analysis at the end of the trial. The weighted statistic no longer has to be significant in addition to the unweighted, and Type I error is still controlled at the nominal level. Figure 3.4 illustrates the three approaches proposed by Chen *et al.* (Chen 2004).

Similarly to the third approach, Lan and Trost (Lan 1997) propose the following rules, incorporating a futility boundary and an unadjusted final analysis:

1. If CP $\leq c_{fut}$, stop the trial for futility and accept H_0
2. If CP $\geq c_{eff}$, continue to the originally planned number of patients. H_1 is accepted at the end of the trial if $Z > z_{1-\frac{\alpha}{2}}$
3. If $c_{fut} <$ CP $< c_{eff}$, increase the sample size to n^*, such that CP under the current trend is maintained at level c_{eff}. n^* is calculated using $\Pr\left(Z^* > z_{1-\frac{\alpha}{2}} | z_1, d = \hat{d}_{obs}\right) = c_{\text{eff}}$. H_0 is rejected at the final analysis if $Z^* > z_{1-\frac{\alpha}{2}}$

However, the above rules must be followed exactly in order to preserve Type I error still, and the literature suggests that there is much concern about researchers strictly following these rules (Hung 2016a).


Chen et al. (2004), incorporating a futility boundary:
Stage 1
Enrol n1/n patients
Interim Analysis – Calculate Conditional Power under the current trend
Interim Analysis
CP<cfut
cfut≤CP<50%
CP≥50%
Stop for Futility
Continue to n patients
Increase to n* (if feasible)
Stage 2
Final Analysis using conventional test statistic
Final Analysis
Approach 1:
Stage 1
Enrol n1/n patients
Interim Analysis – Calculate Conditional Power under the current trend
Interim Analysis
CP<cfut
cfut≤CP<50%
CP≥50%
Stop for Futility
Continue to n patients
Increase to n* (if feasible)
Stage 2
Final Analysis using a decreased critical value
Final Analysis
Approach 2:
Stage 1
Enrol n1/n patients
Interim Analysis – Calculate Conditional Power under the current trend
Interim Analysis
CP<cfut
cfut≤CP<50%
50%≤CP≤ceff
ceff>CP
Stop for Futility
Continue to n patients
Increase to n* (if feasible)
Stop for Efficacy
Stage 2
Final Analysis using conventional test statistic
Final Analysis
Approach 3:
Stage 1
Enrol n1/n patients
Data driven Interim Analysis – Calculate Conditional Power under the current trend (not necessarily pre-planned)
Interim Analysis
CP<cfut
cfut≤CP<50%
CP≥50%
Stop for Futility
Continue to n patients
Increase to n* (if feasible)
Stage 2
Final Analysis using conventional test statistic
Final Analysis

(a) Chen et al.'s general approach, incorporating a futility boundary

(b) Approach 1 - Decrease the critical value at the final analysis

(c) Approach 2 - Incorporate an efficacy boundary

(d) Approach 3 - Use an unadjusted conventional final analysis

Figure 3.4: Chen et al.'s method incorporating a futility boundary and three approaches of 'spending' the α reduction caused by the incorporation of a futility bound. Futility boundary stopping is shown in purple (dotted lines). Red highlights the differences between the three approaches.

The FDA did not recommend decreasing the sample size when uSSR is used, and suggested that a more conventional and "well understood" design should be used instead in their initial draft guidance (FDA 2010). However, in more recent guidance, they are no longer so prescriptive (US Food and Drug Administration (FDA) 2019). It can be argued that by carrying out an unblinded analysis and continuing to recruit to the trial can be unethical when there is already evidence to suggest that the trial can stop either for efficacy or for futility. AdGSDs offer a combination of the standard GSD and adaptive SSR methods; allowing for early stopping, and a SSR in the final stage if the trial has not stopped beforehand (Section 3.14).

Bowden and Mander (2014) implement the modified and reverse LSW method (Section 2.8.1.4) and incorporate stopping boundaries within their methodology. They note that it is very similar to a GSD; pre-specifying stopping boundaries and controlling for Type I error well, which is one of the key strengths of this design.

3.7 Timing Considerations

3.7.1 Time to Primary Outcome

When planning to use interim analyses in a trial, researchers must consider the time until the primary outcome can be recorded. If a large number of patients do not yet have primary outcome data at this time point, known as pipeline patients (Sully 2014), they will not be able to be included in the interim analysis. This could influence the outcome of this analysis, and a different decisions may be made, had these pipeline patients been included. The magnitude of pipeline patients depends on both the rate of the enrolment process and the length until primary outcome data is collected (Liu 2017a).

A primary outcome is defined as happening some time after the treatment is given (Committee on National Statistics 2010). GSDs are often used in situations where the length of time expected to observe the response is very short with respect to the length of the trial (Hampson 2012). In reality, some outcomes take a long time to occur and are generally not observed immediately. Therefore, GSDs may not be the appropriate design in these scenarios. In Mehta and Pocock's paper comparing the adaptive promising zone strategy with a standard GSD, Example 1 shows a trial with a 26 week primary endpoint, enrolling at 8 subjects per week. At the planned interim analysis (Week 52), there are 208 completers, but 416 patients have been enrolled. Therefore any decision to terminate early will hugely affect the savings in sample size expected, as the trial will terminate at 416 patients, rather than the 208 included in the analysis. This differs to the promising zone design as the decision to continue to the originally planned sample size or increase the sample size is not affected until the full originally planned sample size has been recruited (Mehta 2011). Furthermore, SSR based in censored patients at the interim analysis may inflate Type I error (Elsäßer 2014). Jennison and Turnbull (Jennison 2015) highlight the benefit of the Promising Zone design compared to a GSD for pipeline patients and agree that methodology incorporating information from these patients is needed and should be pursued.

Hampson and Jennison have since developed new methodology following the GSD framework, to be able to better deal with pipeline patients, known as DRGSDs. Full details of this methodology are given in Section 3.15.

3.7.2 Accrual Periods

The length of accrual time and rate of accrual also impact on the utility of an interim analysis and the type of design being used for the adaptation. Interim analyses have been said to be useful for trials that do not have a short accrual time (Deley 2012). One reason for this is the length of time required for an interim analysis. If it is expected that a large number of patients, or indeed all, are to be recruited whilst the interim analysis is in progress, then the impact of any decision is minimal. Whilst a sample size increase can still take place after recruitment has finished, resources still need to be found to be able to continue recruitment, which may be logistically very difficult.

However, it should be noted that in some situations, neither ADs and GSD methods necessarily offer the best solution and a search for alternative designs may be required. For instance, with a short accrual rate, and a lengthy primary outcome measure, neither Promising zone nor GSD offer any benefit (Kairalla 2012).

3.7.3 Timing of Interim Analysis

The timing of the interim analysis is important as it can have an impact on the amount of data the decision is based upon. For GSDs, the optimal timing in order to minimise the expected sample size can be derived (Togo 2013). However, there can be considerably more interim analyses in a GSD than in the promising zone design (Mehta 2016a). Within the promising zone literature, there is a general consensus that the interim analysis should not take place very early on in the trial. Liu & Lim (Liu 2017b) advise against carrying out the interim analysis too early, as the treatment effect is unstable early on in the trial. This lowers the ability to rescue an underpowered trial and thus retracts any benefits of using the Promising Zone design. In terms of efficiency, both Levin *et al.* (Levin 2013) and Gaffney *et al.* (Gaffney 2017) agree that the later the interim analysis, or closer to the minimal sample size, the more efficient the design becomes. However, logistical considerations should also be considered, as deciding close to the minimal required sample size may delay the continued recruitment while the adaptations are being implemented (Maca 2014).

3.8 Restricted Sample Size Increase

The extension of the promising zone range described in the methodology of Mehta and Pocock (Mehta 2011) compared to Chen *et al.* (2004) (Chen 2004) leads to a larger range where sample size should be increased. In this extended range, a larger increase in sample size is required to retrieve CP due to the decreasing function used to calculate the required number of subjects (Mehta 2011). When doubling or tripling of the original sample size is feasible, Chen *et al.* (2018) have said that this extension of the promising range can be useful in trials (Chen 2018a). However, there may be instances where promising zone methodology may suggest even more than 3 times the original sample size, and even doubling the sample size can be considered infeasible in many scenarios. When the sample size required in order to retrieve CP is very high, it may become infeasible either financially or logistically and it may become impossible to recruit the necessary number of patients (Bowden 2014).

One solution is to incorporate a maximum cap, n_{max}, that the sample size increase cannot go beyond. If the CP falls in the range where the sample size is required exceeds this value, the value of the maximum cap is instead taken. The flexibility in the design in allowing a maximum cap is appealing for researchers. For example, when looking at a portfolio of studies, it is inefficient to extend past a certain number of subjects as only small increases in power can be seen past this point (Antonijevic 2016). Similarly, Gaffney *et al.* found that introducing a maximum sample size restriction can restrict the increase in power, stating that the smaller the maximum cap of n, the smaller the expected increase in power (Gaffney

2017). Gaffney *et al.* also show the intuitive result whereby the larger the cap, the larger the expected gains in expected power increase, and the expected sample size. However, increasing the maximum cap from 1.5 times to 2.5 times the original planned sample size reduced the efficiency of SSR (defined as the expected power of the AD minus the power of the fixed sample size design) unless the true effect size is small (Gaffney 2017). Liu *et al.* (2016) found that if a conservative promising zone range was used, and sample size is restricted to no more than a 50% increase, then there is no loss in efficiency (Liu 2017a). Furthermore, any loss in power is counteracted by the substantial gain in CP when the treatment effect is found to be promising (Liu 2017a).

The value of n_{max} must be pre-specified and kept constant throughout the trial, as Type I error can be inflated if this value is lowered at any point after the start of the trial (Wang 2013).

3.9 Benefits of Promising Zone

The key benefit of using a promising zone is that it reduces the risk of an underpowered trial (Mehta 2012). Section 2.8.1 highlights the prevalence of underpowered studies and their underlying issues for research. Mehta also advocates the use of an unadjusted critical value at the final analysis, making the methodology approachable to researchers and allows for easy implementation and interpretation.

Type I error is well controlled, even when the algorithm triggers an increase in sample size (Brannath 2012). This corresponds with the research of Broberg *et al.* (2013), who derive Type I error inflation using the promising zone methodology and show that any inflation is very low/nominal (Broberg 2013).

In the original Mehta & Pocock paper (Mehta 2011), there is a large emphasis on the CP gain from using a promising zone design as opposed to a GSD or a fixed sample size design, particularly when the treatment effect is found to be promising. This is particularly advantageous to time-to-event trials when there is a delayed treatment effect (Freidlin 2017). However, this has no utility in reality unless this also corresponds to a gain in unconditional power (Freidlin 2017).

Promising zone designs are also useful for small biotech companies or medical device firms who may not be able to invest large amount funds for large trials (Mehta 2013). Instead, they can commit resources in stages, rather than having a large upfront cost for an unknown treatment effect (Pritchett 2015). If the treatment effect is promising, they are more likely to obtain the funding required for the necessary increase in sample size, as the effect has been shown to be promising (Posch 2013).

One team did a simulation study incorporating promising zone methodology in the development of biosimilars. Their idea was to apply the methodology within a seamless phase II/III trial and results suggested some appealing advantages, including time, cost and sample size savings, which could speed up the development of biosimilars (Uozumi 2017).

It has been indicated that Promising Zone designs could be particularly useful in early stage exploratory studies, where very little is known about the treatment effect (Wang 2012), and in clinical trials of rare diseases (Bayar 2016). However, more research is needed to fully investigate the benefits the Promising Zone design has in both areas.

3.10 Limitations of Promising Zone

Promising zone methodology has had a number of criticisms reported in the literature. This section summarises the main concerns.

3.10.1 Only Increases in Sample Size Allowed

The methodology of Mehta and Pocock (Mehta 2011) adds a constraint on the algorithm that the sample size must never decrease below the originally planned sample size. In 2014, Hung *et al.* pointed out that results falling in the favourable zone will have much higher CP than required, and suggested a decrease in sample size to maintain the minimal required CP would be advantageous to Promising Zone methodology (Hung 2014; Hung 2016b). However, further steps to control Type I error would need to be taken.

However, Mehta has since pointed out that the FDA guidelines clearly indicate that an adaptive SSR should only be used for an increase in sample size (FDA 2010), and therefore the constraint in their original methodology is to enable these guidelines to be followed (Mehta 2013). The FDA instead recommend well understood methodology of GSDs to be used if any decrease in sample size is to be considered, instead of a SSR design with scope for a decrease.

3.10.2 Using the Conventional Test Statistic

Glimm responded to the publication by Mehta and Pocock with two main criticisms (Glimm 2012); the conservative level of α using a conventional final analysis and the high variability in the estimate of the first stage treatment effect (Turnbull 2017). The conservative α level is described here (Glimm 2012; Turnbull 2017).

The conditional error function introduced by Proschan & Hunsberger (Proschan 1995) (discussed in Section 2.8.1.1) presents a function $A(z_1) \in [0,1]$ such that:

$$\int_{-\infty}^{\infty} A(z_1)\phi(z_1)dz_1 = \alpha \tag{3.7}$$

If this condition is satisfied, then any test at level $A(z_1)$ using the second stage data is able to maintain an acceptable overall significance level at α.

Müller and Schäfer (Müller 2001) suggest using $A(z_1)$ such that:

$$A(z_1) = P(Z_2 > z_\alpha | z_1) = 1 - \Phi\left(\frac{z_\alpha - \sqrt{\frac{n_1}{n}}z_1}{\sqrt{\frac{n_2}{n}}}\right) \tag{3.8}$$

Thus H_0 is rejected if the second stage p-value is below the $A(z_1)$ level. i.e:

$$1 - \Phi(z_2^*) \leq A(z_1) \tag{3.9}$$

which can be written as:

$$z_2^* \geq \frac{\sqrt{n_2^*}\Phi^{-1}(1 - A(z_1)) + \sqrt{n_1}z_1}{\sqrt{n^*}} \tag{3.10}$$

due to

$$z_2^* = \frac{\sqrt{n_2^*}z_2^* + \sqrt{n_1}z_1}{\sqrt{n^*}} \tag{3.11}$$

Using Lemma 1 proposed by Mehta and Pocock (Mehta 2011), then regardless of the SSR rule used, the following holds:

$$P_0\left(Z_2^* \geqslant b\left(z_1, n_2^*\right)\right) = \alpha \tag{3.12}$$

where

$$b\left(z_1, n_2^*\right) = \left(n^*\right)^{-0.5}\left[\sqrt{\frac{n_2^*}{n_2}}\left(z_\alpha\sqrt{n} - z_1\sqrt{n_1}\right) + z_1\sqrt{n_1}\right] \tag{3.13}$$

Plugging (3.8) into (3.10), the right hand side is the same as $b\left(z_1, n_2^*\right)$. Using the conventional test statistic z_α instead of $b\left(z_1, n_2^*\right)$ is therefore conservative and inefficient.

The issue with using an over conservative test statistic has been a key concern in recent literature (Bowden 2014; Chen 2018b; Tamhane 2012; Bauer 2016b; Bauer 2016a). The CHW statistic (discussed in Section 2.8.2.5) can be used as an alternative final analysis (Mehta 2016b; Hung 2014), but has shown to still be less efficient than deriving a critical value from sufficient statistics (Tamhane 2012).

The main argument by Mehta for using the unadjusted critical value in the final analysis stems from the ease of understanding and easy implementation in clinical trials (Mehta 2016b). However, it has been highlighted that there are a number of other studies with non-standard final analyses that are widely accepted and understood, such as in GSDs (Shih 2016).

Mehta and Liu indicated later that the use of either conditional error functions, or combination functions could be used with promising zone methodology in order to preserve Type I error (Mehta 2016a). Bauer *et al.* responded to this commentary publication with the following quote (Bauer 2016b):

> “We appreciate that Mehta and Liu are now proposing the combination test approach instead of using the conventional test statistic. This opens the possibility to use more efficient and flexible SSR rules and so, indeed, the broad framework proposed by Mehta & Pocock is valuable for clinical trial design".

Furthermore, it has the attractive quality that all patients are equally weighted, whether they were recruited before or after the interim analysis, unlike the CHW or other weighted test statistics. This has been referred to as the 'one patient-one vote' principle (Mehta 2016b).

3.10.3 Pre-specification of Design Parameters

The promising zone boundary must be pre-specified in advance of the trial, in order to ensure control of Type I error, and for regulatory purposes (Mehta 2011; Zhang 2016). This includes calculation of CP$_{min}$, which is determined by the maximum allowed sample size increase, timing of the interim analysis, and the targeted CP to be maintained. Table 1 of Mehta and Pocock's paper gives CP$_{min}$ values for several scenarios. Mehta and Pocock also advise that these are the smallest values CP$_{min}$ can be, and the researcher may choose a smaller interval in which to define the promising zone. However, decreasing this zone can decrease the trials overall power (Mehta 2011). One such scenario where changing this lower boundary for the promising zone may be to restrict increases in sample size to only happen where the treatment effect is above the MCID (Hsiao 2018). The calculation of the minimum allowed boundary, CP_{min} does not necessarily take this into consideration and should be considered by the researcher designing the trial on a case-by-case basis.

Glimm suggests that the rule for determining (n_2^*) does not need to be pre-specified as Mehta and Pocock state, but may be advisable for regulatory reasons (Glimm 2012).

3.10.4 Other limitations

Although this methodology could be beneficial for small biotech companies, there is also the opportunity for this methodology to be misused. Turnbull suggests that a small company that cannot afford a full trial may plan a trial with an overly optimistic treatment effect, with a SSR design built in, luring investors into funding the increased trial when the treatment effect is unsurprisingly found to be smaller than anticipated (Turnbull 2017). The risk however is that the treatment effect could fall into the unfavorable zone if they are too optimistic, leading to an underpowered trial.

Bioequivalence trials have found Promising Zone designs to have similar power when compared to fixed sample size designs, and GSDs. However, this design also has a higher expected sample size due to only being able to increase in sample size rather than decrease, which is a limitation to using this design in this scenario.

Finally, Jennison and Turnbull (2015) have presented findings that the greatest gain in power per patient lie outside the Promising Zone (Jennison 2015). They recommend an alternative sample size increase formula over a different qualifying boundary instead. Their methodology is discussed further in Section 3.11.1.

3.11 Increasing the efficiency of Promising Zone

3.11.1 Jennison and Turnbull Methodology

Promising Zone methodology has been under much criticism by many statisticians. The zones and the function for an increase in sample size is very sensitive to the estimated treatment difference ($\hat{d}_{obs}$), due to its high variability at the interim stage (Jennison 2015).

Jennison and Turnbull describe a method using inverse normal combination test instead of the conventional analysis, that moderates the increase in sample size, resulting in a smoother function of increased sample size, n^* over a broader 'promising' range.

Inverse normal combination tests are described in Section 2.8.2.4 and control Type I error if used in conjunction with a trial which has flexible rules on whether to adapt or not, provided the conditional Type I error is protected (Jennison 2003). The methodology described by Jennison and Turnbull (2015) is described below (Jennison 2015).

Using the inverse normal combination approach, the sample size may be increased at any value of z_1, or equivalently, $\hat{d}_{obs}$, without inflation of the Type I error. This avoids the conservatism employed in the Mehta and Pocock methodology, and extends the range in which sample size may be increased. Furthermore, a reduction of sample size is also considered for some values of $\hat{d}_{obs}$, but constrains this to a minimum sample size level, influenced by pipeline patients at the interim analysis.

To optimise the sample size increase, n^* is chosen to maximise

$$CP_{\tilde{d}}(z_1, n^*) - \gamma(n^* - n)\,, \tag{3.14}$$

where $\tilde{d}$ is the assumed future treatment difference used to estimate the true treatment difference, d. The choice of γ needs to be pre-determined, and can be chosen in conjunction with a treatment difference that might be considered lower than expected, but still relevant (such as the MCID if the sample size is based on a larger target effect size. It represents the "cost" of adding an additional patient to the trial, so that sample size is increased only up to the point where CP improvement is greater than γ (Jennison 2015).

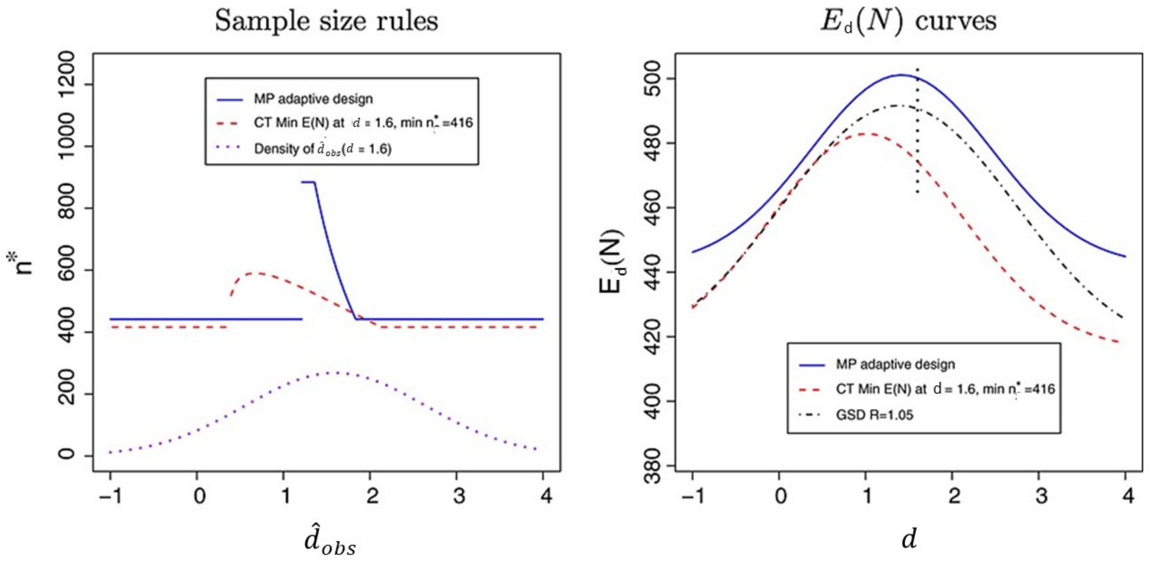


Figure 3.5: Illustration of the Jennison and Turnbull methodology compared to fixed sample size design and Mehta and Pocock methodology. Sample size rules are shown on the left, and expected sample size is shown on the right (Jennison and Turnbull 2015. Used with permission from John Wiley and Sons) (Jennison 2015)

The methodology of Jennison and Turnbull has the greatest increase in sample size at lower values of $\hat{d}_{obs}$ than in Mehta and Pocock's design, avoids the conservatism from the use of the conventional test statistic, avoids the sharp increase in sample size at CP$_{min}$, and allows for a slight reduction in sample size. This methodology opens up the broad framework of Mehta and Pocock for use in future clinical trials. An illustration of the sample size increase rule, and the expected sample size is shown in Figure 3.5.

Furthermore, using Bayesian methods, a sample size increase rule to optimise the weighted average of expected sample size, $E_{d_i}(n)$, as opposed to the expected sample size $E_d(n)$ at a single value $\theta = \tilde{d}$ can be found.

3.11.2 Bowden and Mander Methodology

Bowden and Mander implemented a 'reverse LSW approach' (described in detail in Section 2.8.1.4) and obtain a broadly similar result to the methodology of Mehta and Pocock.

The standard implementation of the LSW method has a lower expected sample size compared to the fixed sample design, but also has a lower overall power, and a lower non-standard critical threshold for the final analysis, which may cause some concern with many trialists (Bowden 2014). The reverse LSW approach allows implementation a SSR rule without being overly conservative; one big criticism of the work of Mehta and Pocock. This is due to allowing for early stopping at the interim stage and therefore decreases the expected sample size, without loss of power (Bowden 2014).

Due to the iterative procedure for identifying sets of h_{lsw}, k_{lsw} and Z_{β_1} required for the methodology of Bowden and Mander, and the number of corresponding design parameters required, the thesis will not investigate this design in any further detail. As mentioned previously, the main motivation for the thesis is to increase uptake of uSSR designs, in which design simplicity is thought to be a key driving factor.

3.12 Extensions to the Promising Zone Framework

Other methods for increasing the sample size at the interim analyses also exist. While the scope of the literature review is mainly on Promising Zone methodology, the following principles may be useful applications and may be considered when investigating methods of improving efficiency or logistical aspects within the promising zone framework.

3.12.1 Stepwise Increases in Sample Size

The increase in sample size following an interim analysis can inform investigators of the interim result. For instance, Liu and Hu provide an example whereby the investigators are told there is a 20% increase in sample size to be implemented following the interim results, and the investigators then know that CP is around 75% (Liu 2016), which can be used to back-calculate information about the treatment effect. This can cause operational bias, which could affect the running of the trial and the validity of the trial results. They then go on to recommend a stepwise function of sample size increase, where sample size is increased in steps up to a point, and then decreased in steps, which is based off the work in 2015 by Wan *et al.* (Wan 2015).

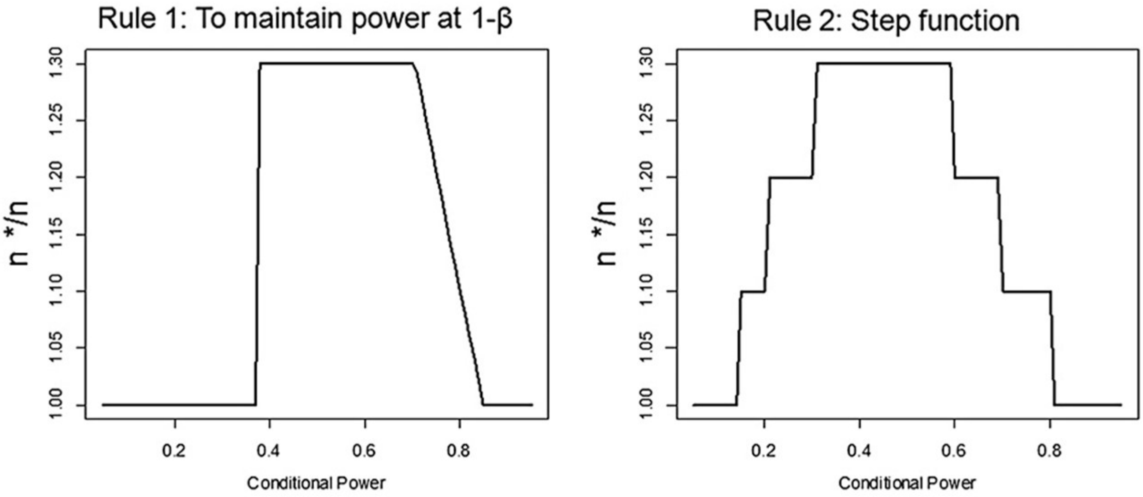


Figure 3.6: Illustration of the Promising Zone increase in sample size (Rule 1; Left) compared to the stepwise increase in sample size (Rule 2; right) (Liu & Hu 2016. Used with permission from John Wiley and Sons) (Liu 2016)

A 20% sample size increase is then less informative to the investigators, as now CP may be ‘unfavorable’ (20-30%), or ‘promising’ (60-70%). Figure 3.6 illustrates the difference in sample size increase rules using the promising zone design and the stepwise function for SSR.

Their methodology is briefly described below (Liu 2016).

The ratio of sample size increase, n^*/n, can be expressed in a stepwise form with J steps as follows:

$$n^*/n = \sum_{j=1}^{J} v_j I(z_1 \in l_j) \tag{3.15}$$

where $v_j \geq 1$ is the step value, chosen by the investigator in advance, z_1 is the test statistic at stage 1, l_j is the interval $l_j = (l_j^{(L)}, l_j^{(U)})$, and I is the indicator function. It can be shown that the Type I error is controlled provided the following condition holds for all $j \in M = j : r_j > 1, j = 1, ..., J$:

$$l_j^L \geq \frac{\left(\sqrt{1 - \frac{n_1/n}{v_j}} - \sqrt{1 - n_1/n}\right) z_\alpha}{\sqrt{n_1/n}\left(\sqrt{1 - \frac{n_1/n}{v_j}} - \sqrt{\frac{1 - n_1/n}{v_j}}\right)} \tag{3.16}$$

The step value, v_j, can be chosen depending on logistical or budgetary aspects (similar to the restriction of sample size in the promising zone framework). For example, if the maximum number of events is 1.45 times the original sample size, Liu and Hu suggest the following values may be appropriate: v_j = 1.2, 1.3, 1.45, 1.3, 1.2. With an interim analysis completed at $n_1/n = 0.5$, the corresponding lower boundary corresponds to a CP value of 0.45. Provided the first sample size increase step happens no lower than CP=0.45, there is no further constraint on CP values for the other steps.

However, if step rules are fully pre-specified in advance of the trial, there is scope to decrease the CP value of the first step, say to CP=0.3, and still control Type I error. One possibility, given in the paper by Liu and Hu is (Liu 2016):

$$\begin{aligned}(n^*/n) = I(CP(z_1) \leq 0.3) \\ + 1.2 * I(CP(z_1) \in (0.3, 0.35]) \\ + 1.3 * I(CP(z_1) \in (0.35, 0.4]) \\ + 1.45 * I(CP(z_1) \in (0.4, 0.6]) \\ + 1.3 * I(CP(z_1) \in (0.6, 0.75]) \\ + 1.2 * I(CP(z_1) \in (0.75, 0.85]) \\ + I(CP(z_1) > 0.85)\end{aligned}$$

Wan *et al.* also provide alternative methodology for increasing sample size in a stepwise manner and controlling Type I error (Wan 2015), similar to GSD methodology. Furthermore, Levin *et al.* (2013) developed alternative methods in order to choose optimal second stage sample size in a symmetrical stepwise manner. Their methodology investigates CP and predictive power ranges under both the alternative hypothesis at the planning stage ($\tilde{d} = d_{plan}$), and a Maximum Likelihood Estimate (MLE) ($\tilde{d} = \hat{d}_{obs}$) as future treatment effect assumptions (Levin 2013).

3.12.2 Surrogate Endpoints for Long Term Outcomes

When lengthy primary outcome measures such as overall survival are required for full regulatory approval, pharmaceutical companies often look at ways to speed up the approval process, in order to get the treatment to patients sooner. Liu and Hu (2016) have suggested an approach whereby a ‘surrogate’, shorter term, endpoint can be investigated in parallel to the long term required outcome (Liu 2016). This is more informative than the longer term outcome at this stage as more patients will have reached this endpoint and will be included in the interim analysis. If the surrogate endpoint, such as overall response rate, or progression free survival, shows to be effective at the interim analysis, accelerated approval may be recommended. The final analysis still looks at and tests the primary outcome required for full approval, but the shorter term co-primary endpoint may be able to provide evidence for conditional approval before the overall survival data is ready to be analysed.

Promising zone methodology may be used in conjunction with the surrogate endpoint methodology. At the interim analysis, the surrogate endpoint is tested for efficacy. If significant, the treatment is recommended for accelerated approval. Whether significant or non-significant, a SSR procedure is carried out based on the long term outcome. The SSR procedure can either be carried out to retrieve CP (such as in the promising zone methodology), or in a stepwise manner (discussed in Section 3.12.1). Either approach leads to sample size either remaining at the planned level, or increasing to some value. At the final analysis, both endpoints are tested for efficacy, applying methods to control Type I error for multiple comparisons as appropriate. Filing for full approval is only carried out if the longer term outcome, such as overall survival, is found to be significant (Liu 2016).

3.12.3 Adaptive Switch Designs

At a Drug Information Association (DIA) meeting in 2015, Cyrus Mehta discussed two very large cardiovascular trials that both failed to show superiority, but managed to show non-inferiority. As non-inferiority trials generally require fewer subjects and less time, he suggested that had these trials been designed as a non-inferiority, both trials could have substantial time and cost savings. However, it is easy to see this in hindsight, but this is difficult to know at the planning stage. Furthermore, this design may not benefit every scenario, for instance those that require superiority only for changing standard practice and where demonstration of non-inferiority is not appropriate for the context. He presents a design he co-authored in 2013 (Gao 2013a) whereby a test for non-inferiority may be followed by a test for superiority. There is also scope in this design to switch to the adaptive sample size design, if allowed by pre-determined rules illustrated in Figure 3.7, where promising zone methodology may be incorporated to increase the sample size to that required by the design switch to a superiority trial.

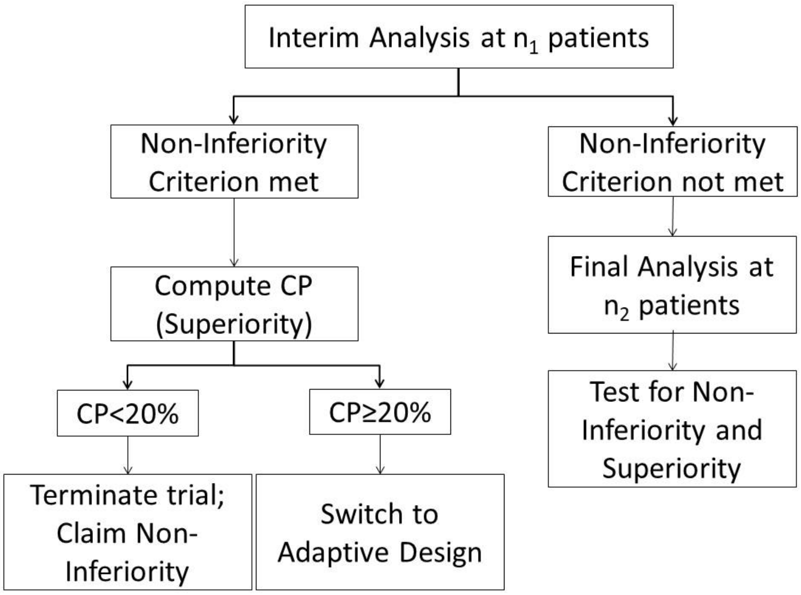

Interim Analysis at n_1 patients
Non-Inferiority Criterion met
Non-Inferiority Criterion not met
Compute CP (Superiority)
Final Analysis at n_2 patients
CP<20%
CP≥20%
Test for Non-Inferiority and Superiority
Terminate trial; Claim Non-Inferiority
Switch to Adaptive Design

Figure 3.7: Flowchart to illustrate the decisions made for changing a non-inferiority to a superiority trial - known as the 'Adaptive Switch' Design (Senchaudhuri 2015)

3.13 Promising Zone vs Group Sequential Designs

There has been much discussion in the literature about the use of Adaptive SSR and standard GSDs, and mixed views on the advantages of each.

Wan *et al.* state that using promising zone methodology reveals more information about the interim data and trial progress compared to a standard GSD (Wan 2015). This comes from the potential back-calculation of the treatment effect from knowledge of the sample size increase after the interim analysis (Maca 2014; Kairalla 2012). Some literature suggest logistical solutions such as restricting access to 'closed' documents only available to the DMSC and recording who accesses the information and reasons (Huskins 2018). If back-calculation is expected to be a big issue, the study team could consider one of the stepwise sample size increase rules discussed in Section 3.12.1 (Liu 2016; Wan 2015; Levin 2013). Mehta argues that GSDs require more interim looks and therefore also could give away information at various stages throughout the trial (Mehta 2016a). To fairly compare the two designs, the GSD should be chosen to have comparable number of interim looks. Alternatively, if the trial team can find a way around this issue, such as only revealing the maximum sample size and telling sites to recruit until told to stop, such methods may not be required.

Another argument is the way each design is used, and their comparisons. In 2013 Mehta responded to one particular critical paper to adaptive SSR (Levin 2013) with the following quote (Mehta 2013):

> “Group sequential designs and adaptive SSR designs play different roles in clinical drug development. Attempts to compare them based on traditional metrics such as power and expected sample size are fruitless and often misleading"

Sample size can never decrease in the standard promising zone framework (i.e. not incorporating a stopping boundary) and will always be at least the sample size of the fixed sample size design. Their use however, is to retrieve CP and so will be increased if necessary. Therefore, if the aim is to minimise expected sample size, the adaptive SSR would never be implemented as it will always give a higher expected sample size. Mehta recommends the use of adaptive SSR as an insurance policy for seeing lower than expected treatment difference that is still considered as clinically meaningful (Mehta 2013). A similar viewpoint to that of Mehta, is the work published by Chen *et al.* (2015). They agree that adaptive SSR and GSDs aim to address different issues in large Phase III trials (Joshua Chen 2015). They also advocate the use of an AdGSD, stating that well planned GSDs can benefit from the data dependent SSR (Joshua Chen 2015).

Many publications have aimed to compare GSDs and adaptive SSR as pointed out by Mehta (Mehta 2013), and some agree that it is often very difficult to compare these designs due to their different efficiency goals (Mehta 2013; Gao 2013a; Zhang 2016; Menon 2013). Zhang *et al.* (2016) point out that some literature compares one specific GSD, and then generalises to all (Zhang 2016). Menon *et al.* suggest more research is required in order to better critically evaluate the promising zone methodology (Menon 2013).

3.14 Extension to >2 stages

ADs offer flexibility in a trial and can modify a trial in progress based on the accumulating data (Müller 2001). Standard GSDs do not offer such flexibility in ranges of sample size increase, but do offer well established methods for the control of Type I error and efficient designs (FDA 2010). AdGSDs, a combination of the two methods, have been developed (Bauer 1994; Müller 2001). The promising zone design can be thought of as a simple two-stage AdGSD. However, other AdGSDs can incorporate multiple interim looks, and stopping boundaries (Liu 2017a). The key change is that at one or more of these interim analysis, a SSR method may be incorporated, and instead of terminating the trial or continuing on to a pre-specified number of patients, the trial either terminates, or continues to the required number of patients (n) or (n^*), based upon the interim results. Müller and Schäfer (2001) presented methodology where alpha-spending functions could be modified adaptively at any interim analysis, as well as the number and timing of future interim analyses (Müller 2001).

Gao *et al.* (2008) also present AdGSD methodology, incorporating a SSR only at the penultimate analysis of a standard GSD. Similarly, they determine the timing of the penultimate look stochastically (Gao 2008).

More recently, Cui *et al.* (2017) present an AdGSD based on the CHW statistic, and methodology for optimizing the AdGSD.

Consider a two-arm GSD with K-1 interim analyses, and 1 final analysis. The information fraction at each analysis $0 < \tau_1 < \tau_2 < ... < \tau_{K-1} < \tau_K < 1$ is pre-specified. The critical value, C_k, corresponding to each information fraction τ_k can be calculated depending on the alpha spending function to be used. The total sample size per group is not pre-specified and is denoted M, determined at interim analysis g, with $0 < g < K$ and n_g subjects per treatment arm. Before the sample size is determined (i.e. $k \leq g$), each interim analysis is carried out with $n_k = \left(\frac{\tau_k}{\tau_g}\right) n_g$ patients. After the sample size determination, (i.e. $g < k \leq K$, each interim analysis is carried out with $m_k = (M - n_g)\frac{\tau_k - \tau_g}{1-\tau_g} + n_g$ patients.

Let

$$S(q) = \frac{1}{\sqrt{q}} \sum_{i=1}^{q} (X_{A_i} - X_{B_i}) \tag{3.17}$$

denote the Z-statistic based on the data with q patients per group and $U_k^{(g)}$ denote the test statistic at the k^{th} interim analysis, with the sample size determination at the g^{th} interim analysis. Before sample size determination, the test statistic can be defined as:

$$U_k^{(g)} = S(n_k), k \leq g \tag{3.18}$$

and after the sample size determination, as:

$$U_k^{(g)} = S\left(n_g\right)\sqrt{\frac{\tau_g}{\tau_k}} + \frac{\sum_{i=n_g+1}^{m_k}\left(X_{A_i} - X_{B_i}\right)}{\sqrt{m_k - n_g}}\sqrt{\frac{\tau_k - \tau_g}{\tau_k}}, g < k \leq K \tag{3.19}$$

$U_k^{(g)}$ is repeatedly tested against C_k to test the hypothesis. If H_0 is not rejected prior to the g^th interim analysis, the sample size determination procedure goes ahead, and it is projected that M subjects will be recruited per group, unless the trial terminates before the final analysis (Cui 2017). M can be calculated using the following equation:

$$M = n_g + \left(\frac{\sqrt{\frac{\tau_K}{\tau_K - \tau_g}}\left(C_K - S\left(n_g\right)\sqrt{\frac{\tau_g}{\tau_K}}\right) + z_{1-\beta}}{\hat{d}_{obs}}\right)^2 \tag{3.20}$$

Alternative methods are also presented where sample size is to be determined in advance if required (Cui 2017).

Pocock *et al.* (2015) recommend the use of an AdGSD, provided realistic assumptions about the design parameters are used (Pocock 2015). However, they also advise that a simple approach may be advantageous, particularly when the trial team lacks an experienced statistician (Pocock 2015).

For the purposes of this thesis, the term AdGSD will refer to a design with 3 or more stages, and only a SSR in the penultimate analysis (i.e. final interim analysis planned) will be considered.

3.15 Delayed Response Group Sequential Designs

As described in Section 3.7.1, pipeline patients can be very problematic for standard GSDs and therefore often cannot be used when the time to the primary outcome measure is long (Hampson 2012). In 2012, Hampson and Jennison presented new methodology that can be used in order to overcome this (Hampson 2012). Their methodology for a two-arm clinical trial with normal data is presented below (Hampson 2012).

Consider a K stage GSD with K-1 interim analyses. Suppose that primary outcome responses are available after some time, λ post-randomisation. The DRGSD can halt recruitment at an interim analysis, but waits λ time before the analysis is carried out, to allow pipeline patients to have available data. After this time, a 'decision analysis' to accept or reject H_0 is conducted. For $k = 1, ..., K-1$, denote the number of responses at interim k by n_k. If recruitment stops, the number of subjects at the subsequent decision analysis is denoted by $\tilde{n}_k$. If recruitment continues beyond interim $K-1$, the final decision analysis is only conducted once all responses from the total sample, $\tilde{n}_k$, have bee collected. Once stopped, recruitment is no longer allowed to be restarted.

Let $\hat{d}_k$ denote the MLE of d based on n_k responses at interim k, $\mathscr{I}_k = \text{var}\left(\hat{d}_k\right)^{-1}$ as the Fisher information for d, and $S_k = \mathscr{I}_k \hat{d}_k$ as the score statistic. Similarly, Let $\tilde{d}_k$ denote the MLE of d based on $\tilde{n}_k$ responses at decision analysis k, $\tilde{\mathscr{I}}_k = \text{var}\left(\tilde{d}_k\right)^{-1}$, and $\tilde{S}_k = \tilde{\mathscr{I}}_k \tilde{d}_k$. The following algorithm is then followed:

At interim analysis $k = 1, ..., K-2$:

- **If** $S_k \leq c_{fut_k}$ **or** $S_k \geq c_{eff_k}$: Stop recruitment and proceed to decision analysis k
- **Otherwise**: Continue recruitment and proceed to interim analysis $k+1$

At interim analysis $K-1$:

- **If** $S_{K-1} \leq c_{fut_{K-1}}$ **or** $S_{K-1} \geq c_{eff_{K-1}}$: Stop accrual and proceed to decision analysis $K-1$
- **Otherwise**: Complete recruitment and proceed to decision analysis K

At decision analysis $k = 1, ..., K$:

- **If** $\tilde{S}_k \geq c_k$: reject H_0
- **If** $\tilde{S}_k < c_k$: accept H_0

DRGSDs offer one solution to the criticism standard GSDs have faced when dealing with long term primary outcome measures and methods have been developed to optimise DRGSDs for either expected sample size alone, or a combination of objectives (Hampson 2012). While some literature makes the comparison between AdGSDs and promising zone methodology, there is very little with respect to incorporating DRGSDs into the comparisons.

3.16 Analysis considerations

Following an adaptation to a trial, there is a risk of introducing bias to the estimated treatment effect. The EMA require trialists to stratify their final analysis to take into account the trial before and after the interim analysis (Magirr 2016), as before and after are 'inherently different cohorts' (Liu 2017b). This is in line with the most recent guidance from the Adaptive Designs CONSORT Extension (ACE) statement, which recommends presenting the treatment effect by recruitment stage separately (Dimairo 2020).

Bowden and Mander (2014) state that the MLE of d will generally be biased, as the trial's sequential nature is ignored following the LSW method (Bowden 2014). To overcome this, Wang *et al.* (2010) introduce a median unbiased estimator, shown to reduce the bias and mean squared error, but only when d is small and positive (Wang 2010).

Repeated Confidence Intervals (RCIs) methodology has been developed for GSDs, for repeated looks at the data during the trial progression. The probability of all intervals containing the true parameter of interest, d is maintained at level α. The methodology is described below.

3.16.1 Repeated confidence intervals with no adaptation

In 1984, Jennison and Turnbull proposed RCIs following GSDs, for normally distributed outcomes (Jennison 1984). As stopping a trial early depends on many factors, including practical aspects such as side effects, finance, quality of life etc., they state that there is no guarantee a stopping rule will be strictly adhered to. This means that the correct decision is therefore not guaranteed. In 1989, they extended this methodology to binary and time-to-event outcomes.

Suppose a trial has up to K analyses, and let ρ_k form a sequence of RCIs with two-sided level 1-2α for treatment difference $\hat{d}_{obs}$.

$$P_d(d \in \rho_k \; \forall \; 1 \leq k \leq K) = 1 - 2\alpha \tag{3.21}$$

At the k^{th} analysis, interval ρ_k is the set of all values d_0 which would be accepted by group sequential test $H_0 : d = d_0$ such that

$$P_d(d \in \rho_k \ \forall \ 1 \leq k \leq K) = P_d(d \text{ accepted at all } 1 \leq k \leq K) = 1 - 2\alpha \tag{3.22}$$

The group sequential test at stage k is written as reject H_0 at stage k if $|z_k| \geq c_k$, where z_k is the standardised test statistic, and c_k are critical values constructed to ensure the test has size 2α. So the intervals ρ_k can be written as $\rho_k = \{\theta : |z_k| < c_k\}$, or $\rho_k = (d_L, d_U)$, (lower, upper), so that $P_d(d_L < d < d_U \ \forall \ 1 \leq k \leq K) = 1 - 2\alpha$. Finally, intervals are roughly symmetric, and so

$$P_d(d_L < d \ \forall \ 1 \leq k \leq K) \approx P_d(d_U > \theta \ \forall \ 1 \leq k \leq K) \approx 1 - \alpha \tag{3.23}$$

It is worth noting, that if the trial is stopped prior to the K^{th} analysis, the RCIs are conservative, as not all RCIs have been seen yet.

3.16.2 Repeated confidence intervals with at least one adaptation

Müller and Schäfer (2001), and Mehta *et al.* (2007) extend the framework of RCIs to a GSD where one or more adaptive changes are made to the trial during the trials progression by constructing one sided confidence limits to a sequence of dual tests, derived from RCIs (Müller 2001; Mehta 2007; Jennison 1984; Jennison 1989; Gao 2013b). This again preserves the desired coverage probability at a pre-specified level, even if a stopping rule were to be ignored. Their methodology for a normally distributed outcome is described below.

Consider a two-arm trial for a treatment (t) group and control (c) group, with means μ_A, μ_B, treatment difference $d = \mu_A - \mu_B$, and one-sided hypothesis $H_0 : d \leq 0$, $H_1 : d > 0$. Similarly to above, the data is monitored up to K times, after observing $n_1 + n_2 + ... + n_K$ subjects.

At interim analysis k, $Z_k = \frac{\hat{d}_{obs}}{\sqrt{\mathscr{I}_k}}$, where $\hat{d_k}$ is the MLE of d and $\mathscr{I}_k$ is the Fisher Information $\mathscr{I}_k \approx [se(\hat{d_j}]^{-2} = n_k/4\sigma^2$. Then $Z1, ..., Z_K$ are multivariate normal with $E(Z_k) = d\sqrt{\mathscr{I}_k}$ and $Cov(Z_j, Z_k) = \sqrt{\mathscr{I}_j/\mathscr{I}_k} \ \forall \ j \leq k = 1, 2, ..., K$.

Suppose Lan DeMets stopping boundaries are used, with spending function $\zeta_\alpha(\tau)$ which is monotone increasing with $\zeta_\alpha(0) = 0$ and $\zeta_\alpha(1) = \alpha$. Then value $\zeta_\alpha(\tau_j)$ represents the cumulative Type I error used up to and including look j, where τ_j is the information fraction. The corresponding stopping boundaries $b_1, b_2, ..., b_K$ are found by

$$\zeta_\alpha(\tau_{j-1}) + P_0(Z_1 < b_1, ..., Z_{j-1} < b_{j-1}, Z_j \geq b_j) = \zeta_\alpha(\tau_j) \tag{3.24}$$

In order to preserve Type I error level α when one or more modifications are made to a trial, Müller and Schäfer introduced a principle to preserve the conditional rejection probability each time an adaptation is made. Suppose the first adaptation occurs at look $L < K$. The conditional rejection probability is defined as

$$\varepsilon = P\left(\cup_{j=L+1}^{K} \{Z_j \geq b_j\} | Z_L = z_L\right). \tag{3.25}$$

Provided this condition holds, Type I error will be preserved at level α no matter the adaptation type (e.g. SSRs, additional interim analyses, changing the timing of the interim analysis etc.). One can think of the remainder of the trial following the adaptation as a separate "secondary" trial with Type I error ε, with the previous trial up to L looks as the primary trial. The secondary trial is then monitored at information fraction $\tau_j^{(2)} = n_j^{(2)} / n_{k^{(2)}}^{(2)}$ and terminated at look $L^{(2)} \leq K^{(2)}$ where $j = 1, 2, ...k^{(2)}$. The observed test statistic at time of termination is therefore $Z_{L^{(2)}}^{(2)} = z_{L^{(2)}}^{(2)}$.

H_0 ($d \leq 0$) is rejected if and only if $z_{L^{(2)}}^{(2)} \geq b_{L^{(2)}}^{(2)}$ where the boundaries $b_j^{(2)}$ are determined from the error spending function, $\zeta_\varepsilon^{(2)}\left(\tau_j^{(2)}\right)$ where $j = 1, 2, ...k^{(2)}$, $\zeta_\varepsilon^{(2)}(0) = 0$ and $\zeta_\varepsilon^{(2)}(1) = \varepsilon$ such that

$$P(z_1^{(2)} \geq b_1^{(2)}) = \zeta_\varepsilon^{(2)}\left(\tau_j^{2)}\right), \tag{3.26}$$

and for $j = 2, 3, ..., K^{(2)}$:

$$\zeta_\varepsilon^{(2)}\left(\tau_{j-1}^{(2)}\right) + P\left(z_1^2 < b_1^2, ..., z_{j-1}^{(2)} < b_{j-1}, z_j^{(2)} \geq b_j^{(2)}\right) = \zeta_\varepsilon^{(2)}\left(\tau_j^{(2)}\right) \tag{3.27}$$

For the combined trial (primary and secondary, consisting of up to $L + K^{(2)}$ analyses, the value of the test statistic at look $L+j$ is

$$z^c_{L+j} = \left(Z_L\sqrt{n_L} + z_j^{(2)}\right)\sqrt{n_j^2}/\sqrt{n_L + n_j^{(2)}} \tag{3.28}$$

and the corresponding stopping boundary at look $L+j$ is

$$b^c_{L+j} = \left(Z_L\sqrt{n_L} + b_j^{(2)}\right)\sqrt{n_j^2}/\sqrt{n_L + n_j^{(2)}} \tag{3.29}$$

This can be extended to have an additional modification further on in the trial (Müller 2001; Mehta 2007). This can also be extended to binary and time-to-event responses (Müller 2001; Jennison 1989).

To construct an RCI for d at look $L^{(z)}$, it is necessary to first construct the overall level α test of Hypothesis H_η. This is done by shifting observed test statistics from the primary and secondary trials, to become $z_j(\eta) = z_j - \eta\sqrt{\mathscr{I}_j}, j = 1,2,...,L$ and $z_j^{(2)}(\eta) = z_j^{(2)} - \eta\sqrt{\mathscr{I}_j^{(2)}}, j = 1,2,...,L$ respectively.

Under $d = \eta$, the primary trial is multivariate distributed with $E[Z_j(\eta)] = 0$ and $Cov\left(Z_{j1}(\eta), Z_{j2}(\eta)\right) = \sqrt{\mathscr{I}_{j1}/\mathscr{I}_{j2}}$, and the secondary trial is multivariate distributed with $E[Z_j^{(2)}(\eta)] = 0$ and $Cov\left(Z_{j1}^{(2)}(\eta), Z_{j2}^{(2)}(\eta)\right) = \sqrt{\mathscr{I}_{j1}^{(2)}/\mathscr{I}_{j2}^{(2)}}$.

The conditional probability of rejecting H_η given $Z_l(\eta)$

$$\begin{aligned}\varepsilon(\eta) &= P_\eta\left(\cup_{j=L+1}^{k}\{Z_j(\eta) \geq b_j\} | Z_L(\eta) = z_L(\eta)\right)\\ &= P\left(\cup_{j=L+1}^{k}\{Z_j(\eta) \geq b_j\} | Z_L(\eta) = z_L(\eta)\right)\end{aligned} \tag{3.30}$$

$\varepsilon(\eta)$ decreases as h increases since conditional probability of crossing boundaries decreases with decreasing $Z_L(\eta)$. To apply the principle proposed by Müller and Schäfer to test H_η, the secondary trial must preserve level of $\varepsilon(\eta)$ instead of ε.

$$\begin{aligned}\zeta^{(2)}_{\varepsilon(\eta)}\left(\tau^{(2)}_{j-1}\right) + P\left(z_1^{(2)}(\eta) < b_1^{(2)}(\eta), ..., z_{j-1}^{(2)}(\eta) < b_{j-1}^{(2)}(\eta), z_j^{(2)}(\eta) \geq b_j^{(2)}(\eta)\right)\\ = \zeta^{(2)}_{\varepsilon(\eta)}\left(\tau^{(2)}_j\right)\end{aligned} \tag{3.31}$$

Then H_η is rejected at look L^2 if and only if

$$z^{(2)}_{2^{(2)}}(\eta) \geq b^{(2)}_{2^{(2)}}(\eta). \tag{3.32}$$

Since the secondary trial preserves conditional rejection probability, $\varepsilon(\eta)$, the test for H_η remains an α level test. To construct $100(1-\alpha)\%$ CI for $\hat{d}_{obs}$, identify all values of η at which corresponding level α dual tests H_η cannot be rejected. By Equation 3.32, values of η must satisfy $z^{(2)}_{L^{(2)}}(\eta) < b^{(2)}_{L^{(2)}}(\eta)$

Provided $z^{(2)}_{L^{(2)}}(\eta) - b^{(2)}_{L^{(2)}}(\eta)$ decreases monotonically with increasing η, then interval $(\hat{d}^{(2)}_L, \infty)$ is a $100(1-\alpha)\%$ RCI for $\hat{d}_{obs}$ where $\underline{d}_{L^{(2)}}$ is the unique value $\eta = \underline{d}_{L^{(2)}}$ at which $z^{(2)}_{L^{(2)}}(\underline{d}_{L^{(2)}}) = b^{(2)}_{L^{(2)}}(d_{L^{(2)}})$.

3.17 Trials Review

A table of characteristics of the 21 trials included in the systematic review are summarised in Tables 3.1-3.2, in order to assess how promising zone methodology is currently being used (NCT03243422, NCT02401412, NCT02915744, NCT01156571, NCT01513369, NCT00968708, NCT00887328, NCT01492725, NCT02388061, NCT01485185, NCT01482962, NCT02376985,NCT01816776, NCT02733991, NCT01556490, NCT01702636, NCT02979899, NCT03287076, NCT02692482, NCT01191801, UMIN000021431) (Deal 2018; Colwell 2018; Tripathy 2018; Miller 2017; Schindler 2018; Hirakawa 2018; White 2013; Churilov 2018; Campbell 2014; Campbell 2018; Boyle 2014; O'Connor 2015; Niikura 2016; Costanzo 2015; De Valk 2018; Chauhan 2018; Meretoja 2014; Mehta 2019; Muller 2018; Forni 2018; Ravandi 2015). Eight trials (38.1%) were still in progress, but included enough details to still be included in this review. Almost all trials only included/plan to include one interim analysis (19/21; 90.5%). No trial planned to use more than two interim looks. The median initially planned sample size was 156 patients (IQR (126,419)), which ranged from 100 to 10,900 patients for binary outcomes, and 24 to 850 for continuous outcomes. Time to event outcomes ranged from 120 to 417 initially planned events (median=273).

The median timing of the first interim analysis was over halfway through the originally sized study (median=62.2%, IQR(48.1%,73.8%)), calculated by dividing the sample size at the interim analysis by the initial sample size and multiplying by 100. One study was excluded due to insufficient information being reported for this calculation. All 21 trials (100.0%) planned for an interim analysis in the planning stage prior to the start of the trial.

Of the 11 trials that have been completed, 10 have reported their interim decisions: two trials continued with the original sample size, two trials terminated early, and 6 trials increased their sample size according to CP calculations and the appropriate increase as indicated by promising zone methodology. Twelve trials (70.6%) used stopping boundaries in addition to the SSR; four used efficacy only, three used futility only, three used both futility and efficacy, and the final two did not report any further details.

Similarly to the timing of the interim analysis calculation described above, both the maximum sample size and the actual sample size reported at the end of the trial (if applicable) were converted into a percentage of the initial sample size calculation. Only 14 of the 21 studies (66.7%) reported the maximum sample size they would be willing to increase to if the CP calculation fell in the promising zone. The median increase was by 63.1% IQR (42.6,100.0%) for sample sizes involving patients. Of the six studies who carried out a SSR at the interim analysis, the median actual increase in sample size was 42.3%, IQR(25.0%,68.3%)

Three trials had more than one primary outcome (details in the footnote of Table 3.2). Half of the outcomes (12/24, 50.0%) were binary, 6/24 (25.0%) were continuous, and 6/24 (25.0%) were time-to-event outcomes. Data for four primary outcomes (19.0%) could be collected in less than 3 days, five outcomes (23.8%) took 1-6 weeks, five outcomes (23.8%) took 8-12 weeks, two outcomes (9.5%) took 5-6 months and five outcomes (23.8%) followed patients up for >3 years.

One study had a crossover design with four treatment options. This study incorporated a group sequential approach, recruiting the initial sample size as cohort 1 and only recruiting a second cohort based on the findings at the interim analysis.

Six studies did not specify a phase. Of the studies that did, only one trial was Phase I, four studies reported as Phase II, and ten studies reported as Phase III.

Interim Analysis Details:	**Trials included in the review (N=21)**	
Number of interim analyses planned:	n/N	%
1	19/21	90.5%
2	2/21	9.5%
Timing of first interim analysis (% of initially planned sample size)		
Median (IQR)	57.3	(48.1, 72.5)
Status of information at first interim analysis		
Completed data collection (primary outcome)	11/20	55.0%
Recruited	5/20	25.0%
Events occurred	4/20	20.0%
Result at first interim analysis		
Continue	2/10	20.0%
Stop	2/10	20.0%
SSR	6/10	60.0%
Maximum sample size/initial sample size (%)		
Median (IQR)	63.1	(42.6, 100.0)
Actual sample size/initial sample size (%)		
Median (IQR)	42.3	(25.0, 68.3)
Stopping boundaries used?*		
Yes	12/17	70.6%
If Yes, Boundary type used		
Efficacy Only	4/12	33.3%
Futility only	3/12	25.0%
Efficacy and Futility	3/12	25.0%
Not specified	2/12	16.7%

Notes:*No information on useage of stopping boundaries for 4 trials

Table 3.1: A table of interim analysis characteristics of twenty-one trials included in the systematic review.

There were a range of disease areas investigated in the trials, including neurology (n=6), oncology (n=5), cardiovascular (n=3), diabetes (n=3), dental care (n=1), orthopaedic (n=1), ostomy (n=1) and pain(n=1). Over half of the trials were funded by Industry (12/21, 57.1%).

The EXAMINE Trial is the only study of the thirteen that uses an adaptive switch design (described in Section 3.12.3). This trial planned for a potential switch to a promising zone design, but terminated early as non-inferiority could be shown.

Specific CP values used to identify the promising zone used were reported for 15 trials and are reported in Table 3.3. Five trials were designed by the same institution and have used the same conditional power range for each trial (0.385 - 0.8).

Five trials also reported a “four-zone” trial based on observed conditional power values; one trial stopped and claimed non-inferiority if CP<0.2, three trials stopped for futility for values CP<0.03, CP<0.1 and CP<0.2, and one trial included an enrichment design based on a subgroup of patients. Primary reasons for using sample size re-estimation methodology are presented in Table 3.4. Specific design features were not well reported, and simplic-

General Trial Details:		
Primary outcome type**		
Binary	12/24	50.0%
Continuous	6/24	25.0%
Time-to-event	6/24	25.0%
Time to primary outcome		
0-3 days	4/21	19.0%
1-6 weeks	5/21	23.8%
8-12 weeks	5/21	23.8%
5-6 months	2/21	9.5%
3-5 years	5/21	23.8%
Number of groups		
2	20/21	95.2%
4	1/21	4.8%
Study design		
Cluster	1/21	4.8%
Crossover	1/21	4.8%
Parallel Group	19/21	90.5%
Phase[†]		
I	1/15	6.7%
II	4/15	26.7%
III	10/15	66.7%
Disease area		
Cardiovascular	3/21	14.3%
Diabetes	2/21	9.5%
Neurology	6/21	28.6%
Oncology	5/21	23.8%
Other[††]	5/21	23.8%
Progress		
In progress	8/21	38.1%
In analysis	2/21	9.5%
Completed	11/21	52.4%
Funder type		
Industry	12/21	57.1%
Non-Industry	9/21	42.9%

Notes: **Three trials had > 1 primary outcome: continuous and binary (n=1), binary and time-to-event (n=1), two time-to-event outcomes (n=1); [†]No information on phase for 3 trials [††]Other disease areas include dental care (n=1), Orthopaedic (n=1), Ostomy (n=1) and pain (n=1)

Table 3.2: A table of general trial characteristics of twenty-one trials included in the systematic review.

CP lower boundary	CP upper boundary
0.3	0.95
0.3	0.9
0.3	0.8
0.36	0.8
0.385	0.8
0.385	0.8
0.385	0.8
0.385	0.8
0.385	0.8
0.41	0.8
0.43	0.85
0.5	0.9
0.5	0.8
0.708	0.8
(Missing)	0.8

Table 3.3: Conditional power lower and upper boundaries of the promising zone for the 15 trials that provided conditional power ranges.

Reason	n
Uncertainty in treatment estimates	7
"Simplicity"/"convenience"	4
Efficiency	3
Ensure power is maintained	4
Unspecified	3

Table 3.4: Reasons for using sample size re-estimation methodology

ity/convenience was reported for 4 trials with no further explanation. It is assumed that this was a reason for choosing promising zone over an alternative uSSR design for these trials.

One limitation of the trials systematic review is that analysis methods used for the treatment estimate and related quantities estimated at the end of the trial were not captured in the review and summarised, as identified in Section 3.16. Additionally, the review could not capture the future treatment effect assumption used in the CP calculation as specific design details were rarely reported.

3.17.1 Case study

One additional case study was found through pearl growing, but did not meet the requirements for inclusion in the trials review, as a SSR was implemented after the trial had finished, and an investigative exercise. The CAPRISA study team considered implementing a promising zone design with stopping boundaries for their time-to event trial, before deciding to opt for the fixed sample size design with 92 events due to the high likelihood that the interim result would suggest an increase in sample size, and the logistical implications meaning that a larger study was not feasible (Taylor 2012).

Retrospective analysis of the CAPRISA study however revealed that an interim analysis after 66 events yielded similar efficacy results to those in the final analysis of the 92 event trial, and the increase in sample size would not have been triggered after all. Nevertheless, this result is highly speculative (Taylor 2012) and finding comparable evidence at 66 events as to at 92 events may have been due to chance. Further retrospective case studies and simulations may be able to validate these findings.

3.18 Areas of Research Deficit

This comprehensive literature review on Promising Zone and related methodology has highlighted key areas in which research is already well established, limitations to the methodology that should be addressed in order to more widely using Promising Zone methodology in the literature, and areas that are under researched and need more work. Assumptions of the future treatment effect is one key grey area in the literature. Despite some potential solutions, further research is required to fully understand the true effect this assumption plays on CP and the scenarios in which a different assumption should be made.

The work by Jennison and Turnbull (Jennison 2015) opens up the full range of CP values in which a sample size increase can be considered and introduces the possibility of a small sample size reduction. Alternative rules using the inverse normal approach could be considered for a number of scenarios including pharmaceutical vs publicly funded trials, different types of data and timings of interim analyses.

The effect of the magnitude of which the treatment effect is misspecified in the planning stage is not yet fully understood, and the implications of deliberately misspecifying the treatment effect to lure in funders with promising results should be investigated. Optimality criterion for Promising Zone designs should be established, with the ultimate aim of comparable criteria to compare GSDs, AdGSDs, DRGSDs and adaptive SSR. The restricted sample size level, n_{max} and the effect on power at different time points of interim analyses could be researched further, and simulations run on real life clinical data could be beneficial to the critique of promising zone (Menon 2013).

The research question of this thesis has been designed to answer as much of the research deficit as possible in this area, but also considers time constraints. The thesis therefore only focuses on uSSR designs using CP to determine the second stage sample size required; including promising zone design, the inverse normal combination test design, and the stepwise design methodologies.

The thesis asks “What sample size rule framework and associated design features need to be considered when using an uSSR in clinical trials with either a continuous or binary primary endpoint?", which will be specifically answered by investigating the following four key aspects:

- Compare existing methodologies for uSSR using CP calculations, with a focus on promising zone and combination test designs
- Incorporate stopping boundaries in each methodological framework and compare interim decision making
- Investigate the future treatment effect assumption used in the CP equation
- Explore CP values when observed effect sizes are equal to, or different by some amount to the target effect size
- Make recommendations for the planning of a future trial using SSR including operational considerations such as when an interim analysis should be carried out, and the maximum sample size increase to consider

The existing methodology comparison will predominantly focus on the promising zone design, and the combination test framework. The stepwise design will also be used to illustrate this SSR rule compared to the other two frameworks. However, as this design loses power compared to the other two, the fundamental reason for choosing this design is the inability to back-calculate the treatment effect. If this is however not an issue in the trial, then an alternative approach would be made, unless some fully optimised design for some prior δ, such as in the work of Wan (Wan 2013).

3.19 Update since December 2018

Since the full systematic review was completed in early January 2019, there have been further advancements in the field of promising zone and uSSR. The search was carried out using the same methods as in the original search, including databases, pearl growing techniques and grey literature for both methodological publications and real-world clinical trials reported by August 2020 inclusive. The updated systematic review was published in Trials in December 2020 (Edwards 2020). The findings and the implications for the thesis work carried out in later chapters is discussed.

Pilz *et al.* commented on both promising zone and combination test designs, before introducing a new optimal two-stage design approach (Pilz 2019). They state that even the designs optimised for some value of δ (such as the MCID) considered to be optimal by Jennison and Turnbull (2015) are not in fact necessarily optimal. As most trials analysed in this thesis have the MCID as the effect size in the sample size calculation, or do not state a value of δ which would be smaller than the target size but still clinically relevant, the designs in this thesis have not been optimised. Instead, a value of γ is chosen as an appropriate "trade-off" between additional sample size and conditional power gain, and therefore this would not impact the research carried out in this thesis (Jennison 2015). Next, Pilz *et al.* introduce Lagrangian multipliers to use in an optimisation procedure to find the optimal two stage design, which have been said to be "more transparent" than the γ parameter in the combination test design (Pilz 2020a; Pilz 2020b).

Additionally, promising zone has been extended to the case where data has a correlation structure such as Multi-reader multi-case (MRMC) studies (Huang 2020)), and when the design also incorporates a response adaptive randomisation aspect in the trial (Wang 2019; Gao 2020). An even more simplified approach to promising zone has been introduced, where CP only depends on z_1 and therefore the addition for more complicated endpoint such as recurrent hospitalisation data, and claims to simplify the design for even "non-sophisticated practitioners" (Wang 2020). None of these designs have any impact on the research of this PhD.

Niewcas *et al.* present a design for a binary trial where a shorter endpoint could be used in the interim analysis stage if it is not possible to base a SSR on a long term outcome, which could be useful in reducing the number of pipeline patients at an interim analysis (Niewczas 2019).

Additional comparsions between types of design have also been completed in recent years. Cytel provide slides from a presentation by Mehta which includes a comparison of a “constrained combination test design”, which builds on the work of Hsiao *et al.* (Hsiao 2018). Another AD vs GSD comparison has been published, with the conclusion that GSDs are more efficient than SSR methods (Casula 2020). As discussed previously, SSR designs do have their place in trial design and therefore this does not change the benefits of carrying out this PhD research. Additionally a thesis has been recently published that has compared promising zone and combination tests in an AdGSD setting (i.e. ≥ 2 stages). Results are similar to this thesis and will be discussed in Chapter 10 following presentation of thesis results (Jimenez 2020).

Asakura *et al.* have shown a contour plot of CP to investigate more than one primary endpoint (Asakura 2020). They also consider stopping a trial for futility, discussing potentially conservative stopping boundaries using O’Brien Fleming method (e.g.) in GSDs, and the dependency of requiring a future treatment effect assumption for CP or predictive power.

An alternative performance measure in the comparison of multiple ADs has been suggested (Herrmann 2020). Current performance measurements may compare the Average Sample Number (ASN), but may not be comparable when designs can stop for futility/efficacy, whereas others may not. Instead, it is recommended to only average sample size in the allowed SSR range (e.g. the “promising” zone in Mehta and Pocock’s design). This would lead to a more comparable estimate between designs. Due to time restraints, it may not be possible to incorporate this performance measure into the thesis, and will therefore be discussed in Chapter 10, or suggested for further work.

Guidance on the reporting of ADs has been issued, including a section on uSSR, with an example of the promising zone design (Dimairo 2020). A repeated search in *clinicaltrials.gov* yielded more results than the initial search (an additional 15 trials) which suggests researchers are becoming more aware of SSR and reporting as part of the basic study design. However, still no methods were shared except in two cases (one promising zone, one CHW), and so could be using any method, and may be blinded or unblinded.

An additional eight trials have been identified that have published that they have specified that they used promising zone design (n=4), or are currently implementing this design (n=4) (NCT03360396, NCT03088033, NCT03645603, NCT02562443, NCT02497469, NCT03340493, NCT03388762, UMIN000031136). Four report CP boundaries, of which two use $0.5 \leq \text{CP} < 1 - \beta$. One trial actually use stepwise methodology with a futility boundary incorporated. One further study reports a futility boundary, and one trial reports both efficacy and futility boundaries are to be used. There is a roughly even split between data types, times to outcome range from 0 days to 30 months, and equal numbers of trials funded by industry and non-industry sources. Therefore, there are similarities to results found in the initial systematic review, and no implications for the thesis have been identified.

3.20 Summary

Chapters 2 and 3 have provided an in detail background and review of uSSR methodology, with a focus on those designs using CP for decision rules. Promising Zone methodology has faced much criticism in recent years, with many statisticians and researchers highlighting various shortcomings of the design such as its inefficiency. However, some more positive publications have come forward, with the ultimate aim of improving the methodology for more general use. Both benefits and limitations of the promising zone design have been presented, as well as further extensions of the broad framework such as the combination test design and stepwise methodology. GSDs, AdGSDs and DRGSDs have been introduced, and some comparisons exist in the literature between standard GSDs and ADs. The current uptake of the promising zone design has also been presented, with 21 trials having planned to use or having already used promising zone in a trial (29 trials with the updated search). Characteristics have been summarised to see scenarios where this methodology is currently used, which will be used when requesting data for the PhD research to ensure similar trial/outcome characteristics.

This literature review has highlighted the areas that require more research, and the importance of the research questions investigated in this thesis in order to further understand the practical implementation for future trialists, and how promising zone compares to alternative designs.

Chapter 4 provides details of the data request process for the re-analysis of real-world trial data with continuous or binary endpoints, an analysis plan, and a summary of trial characteristics, and recruitment details.

4 | Data description

4.1 Introduction

Chapter 3 gave a comprehensive review of the current knowledge and research of uSSR designs. Results have been presented for the systematic review of clinical trials that have used, or have planned to use, the promising zone design for uSSR. Using the information obtained from the review, the next section of the thesis aims to obtain real trial datasets, and re-analyse them as ADs to compare three uSSR designs.

This chapter presents a background description of the 21 outcomes from 14 trials across industry and publicly funded sectors that have been obtained for data re-analysis. Methods for requesting the data, reasons for the choices of trials, and a comprehensive analysis plan are provided. Some notable examples are given for a discussion of key trial features such as patient and site recruitment, missing data, and details of original sample size calculations and results. A summary of trial characteristics follows, offering an complete overview of the data obtained for data re-analysis in this thesis.

4.2 Aims and objectives

The key aims of this chapter include:

- Describe how data was obtained
- Present the plan for analyses to be carried out
- Introduce examples of trials analysed in this thesis
- Summarise recruitment for both patients and sites for all trials

4.3 Obtaining the data

4.3.1 Industry trials

Suitable industry trials were identified through the Clinical Study Data Request (CSDR) system (`https://www.clinicalstudydatarequest.com`), a data sharing platform for data from a number of companies. A researcher may browse through basic details of all available studies, and add relevant required studies to a research proposal. This means multiple datasets can be requested through just one research proposal, even if they come from a number of different companies. The research proposal consists of a summary of the research in plain English, background information as to why this research is necessary, and details of key objectives and outcomes. This then goes through an independent review panel, who either approve or reject the application, or request further details.

For all available trials on the system, CSDR provided the sponsor name, study title, and a link to more study information, either through the sponsor website, clinicaltrials.gov, or EudraCT. A large proportion of these studies were reviewed for data request, as many were missing key information such as sample size or type of outcome data (i.e. binary, continuous, or time-to-event), or were not relevant, due to a time to event primary outcome. A total of 12 studies were chosen for data request. Two studies were removed before the data request was fully submitted, and so a further two studies were chosen in replacement. The full data request was submitted in May 2019. Two companies withdrew their datasets and moved data sharing platform in June/July 2019. This meant four studies were no longer available, and only eight studies were able to be requested. The application was approved in September 2019. Due to the multi-sponsor nature of the application, it took a number of weeks to create a data sharing agreement, which was signed by all relevant parties in December 2019. A further company moved data sharing platform following the signing of the agreement, resulting in one further study no longer being available. Whilst the company offered to transfer all documents to the new system, a further data sharing agreement would have been required. It was decided that the seven datasets still on the CSDR platform were sufficient, as time frames could be extended, or subsets of the data could be analysed in order to replace the missing trials. More details are shown in Table 4.1 and Figure 4.1. Access to the CSDR system was provided in January 2020, with limited data being ready at this time. A cut-off date of 20th April 2020 was chosen and only data that had been added to the system by this date would be used for the data re-analysis in this thesis.

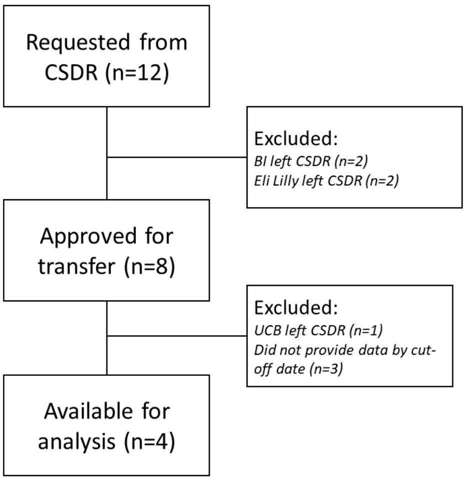

Requested from CSDR (n=12)
Excluded:
BI left CSDR (n=2)
Eli Lilly left CSDR (n=2)
Approved for transfer (n=8)
Excluded:
UCB left CSDR (n=1)
Did not provide data by cut-off date (n=3)
Available for analysis (n=4)

Figure 4.1: Flowchart of the data request process for industry data through the CSDR system.

4.3.2 Publicly funded trials

The National Institute for Health Research (NIHR) strongly encourage making data from trials available, in order to make maximum use of publicly funded data (National Institute for Health Research 2019). Suitable publicly funded trials were identified from two NIHR journals: Health Technology Assessment (HTA) and Efficacy and Mechanism Evaluation (EME). The corresponding author and statistician were contacted for each trial to request an anonymised dataset. A brief outline of the project was provided, with more detailed information being provided if necessary. Initially, 13 datasets were requested in August 2019. Due to the time constraint of completing data analysis, a cut-off of March 2020 for receiving data and signing relevant data sharing agreements was applied. Five datasets were successfully obtained following an application approval and/or data sharing agreement being signed. Four further studies agreed in principle but then had issues obtaining the data: two applications did not get approved in time, one data sharing agreement was unable to be signed, and one further study was not received in time. The other studies did not reply despite follow up emails. In December 2019, 5 further datasets were requested from the

University of Sheffield due to the lengthy process of obtaining data. All five requests were approved and datasets were provided shortly after. Figure 4.2 and Table 4.1 give further details.

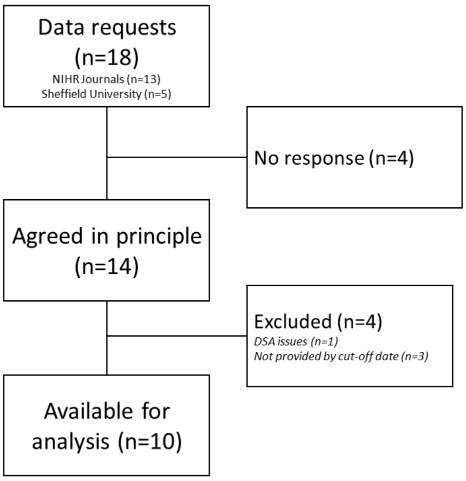


Figure 4.2: Flowchart of the data request process for publicly funded data.

4.3.3 Overview of data obtained

Across publicly funded and industry sectors, a total of 14 trials were obtained for data re-analysis. Due to the slow process of obtaining data, co-primary outcomes and/or secondary outcome data were used where available, resulting in 21 distinct outcomes. Additionally, two primary outcomes were 'reframed' at an additional time point. Table 4.1 provides basic trial details, highlighting the reframed trials, either by using an additional outcome, or re-imagining a time point. All trials randomised patients to two treatment arms except for one (Trial 11 - the 3MG study), which randomised to three.

Ethics approval from the University of Sheffield for the secondary use of patient data was obtained on 16th August 2019, specifying that only anonymised data was involved, and was obtained from already existing research (reference number 030485).

Data summary			
Continuous outcomes; publicly funded			
	Short	**Medium**	**Long**
Small	**Trial 1** FAST INdiCATE n=288; 6 weeks	**Trial 2** SELF n=86; 3 months	**Trial 3** Acupuncture n=241; 12 months
Large	**Trial 4B** CASPER (Minus) n=705; 4 weeks	**Trial 5** CASPER Plus n=485; 4 months	**Trial 4A** CASPER n=705; 12 months
Continuous outcomes; industry			
Small	**Trial 6B** (Epilepsy) n=133; 1 day	**Trial 6A** Epilepsy n=133; 19 weeks	**Trial 6C** (Epilepsy) n=133; 1 year
Large	**Trial 7A** Flu Vaccine (A/H1N1) n=2249; 28 days	**Trial 7B** Flu Vaccine (A/H3N2) n=2249; 3 months	**Trial 7C** Flu Vaccine (B1) n=2249; 1 year
Binary outcomes; publicly funded			
Small	**Trial 8** IMPROVE n=613; 30 days	**Trial 9** Corn plasters n=202; 3 months	**Trial 10** AMAZE n=352; 2 years
Large	**Trial 11A** 3MG n=1109 ; 7 days	**Trial 12** RATPAC n=2243; 3 months	**Trial 11B** (3MG) n=1109; 12 months
Binary outcomes; industry			
Small	**Trial 13** Nasal sprays n=300; <1 hour	**Trial 14A** Mencevax (A) vaccine n=296; 1 month	**Trial 14B** Mencevax (C) vaccine n=296; 1 year
Large	**Trial 7D** Flu Vaccine (A/H1N1) n=2249; 28 days	**Trial 7E** Flu Vaccine (A/H3N2) n=2249; 3 months	**Trial 7F** Flu Vaccine (B1) n=2249; 1 year

Table 4.1: A summary of the twenty four trials that data has been obtained for with sample size and time to primary outcome data becoming available.

4.4 Plan for analysis

4.4.1 Introduction

This analysis plan outlines the details of the proposed analyses to be undertaken on each trial that data is obtained for.

4.4.2 Objectives and endpoints

The aims of the data re-analysis for this thesis are as follows:

1. Consider key logistical factors including timing of the interim analysis and maximum increase of sample size
2. Investigate future treatment effect assumptions used in the CP equation
3. Compare SSR designs based on CP and their impact on n^*
4. Assess the addition of a futility bound at the interim analysis
5. Evaluate stability of the estimate and factors that could influence instability
6. Inform the further work within this thesis

Each objective is explained in more detail in Sections 4.4.2.2 - 4.4.2.7.

4.4.2.1 Logistical features

Whilst CP and new sample size n^* were calculated after every patient, three interim analysis time points were chosen to look at in further detail. Following the systematic review of trials implementing promising zone carried out, and reported in Section 3.17, specific interim analyses at 25, 50 and 75% of patients with data available seemed appropriate points to investigate. CP plots show the lines that correspond to these three analyses: solid lines for number with data available, and dashed for the number recruited at this time point. This is to show the difference between the CP that decisions are based on, and the CP if all currently recruited patients had data available. CP_{min} values are reported for promising zone and stepwise designs, n^* for all trials, and actual CP value are reported for each interim time point and for each future treatment effect assumption (Section 4.4.2.2).

Two maximum sample size increases were considered in the data re-analysis. Again, using the information available from the completed systematic review, 50% and 100% increases seemed the most feasible for a trial to consider, with higher increases often not being financially or logistically viable.

For the purposes of the data re-analysis, the treatment effect and variance from the original trial analysis are assumed to be the true population parameters. This is discussed in more detail in Section 5.5.

4.4.2.2 Conditional power assumptions

Conditional power has been previously described in Section 2.6, and the future treatment effect assumption discussed in Section 3.5.2. Herson *et al.* have previously recommended the use of an 80% optimistic confidence limit in conjunction with a futility boundary, which will be one of the investigated assumptions of this analysis. A 90% optimistic confidence limit has also been chosen to consider as a more conservative assumption (Pepe 1992).

CP will be calculated using

$$CP_\theta(n, z_\alpha | z_1) = 1 - \Phi\left\{ \frac{z_\alpha\sqrt{n} - z_1\sqrt{n_1}}{\sqrt{(n-n_1)}} - \frac{\tilde{d}\sqrt{n-n_1}}{\sqrt{2\hat{\sigma}^2_{obs}}} \right\}, \tag{4.1}$$

where θ represents one of the four treatment effect assumptions for 'future' patients:

- Current trend, $\tilde{d} = \hat{d}_{obs}$ (assuming the data observed so far is likely to continue for the duration of the trial)
- Hypothesised treatment effect, $\tilde{d} = d_{plan}$ (assuming the hypothesised treatment effect used in the original sample size calculation will be seen for the remainder of the trial)
- 80% optimistic limit, $\tilde{d} = \hat{d}_{obs} \pm Z_{1-\frac{0.2}{2}}\sqrt{\frac{2\hat{\sigma}^2_{obs}}{n_1}}$ (where ± depends on the direction considered optimistic)

- 90% optimistic limit, $\tilde{d} = \hat{d}_{obs} \pm Z_{1-\frac{0.1}{2}}\sqrt{\frac{2\hat{\sigma}^2_{obs}}{n_1}}$ (where ± depends on the direction considered optimistic)

Using the current trend ($\tilde{d} = \hat{d}_{obs}$), Equation 4.1 equates to Equation 2.10.

Conditional power using the four future treatment effect assumptions will be calculated after every patient and plotted. These same values will then be used to evaluate all three SSR designs and the impact of the four assumptions on interim decisions will be assessed.

4.4.2.3 Promising zone

The promising zone design is described in detail in Section 3.5. CP_{min} values will be calculated for every possible n_1 value (from 1 to n). If CP values lie between CP_{min} and $1-\beta$, the trial is said to be in the promising zone and sample size is re-estimated according to Equation 3.2, up to a maximum value of n_{max}. If CP values are greater than $1-\beta$, the trial lies in the favourable zone and no increase in sample size is considered. Similarly, if CP values fall below CP_{min}, no SSR occurs as the trial lies in the unfavourable zone.

In reality, only one time point would be chosen for the SSR to take place, and therefore only one value of CP_{min} would be calculated. This re-analysis calculates after every patient in order to illustrate CP stability throughout the trial and the impact of interim decisions.

CP_{min} values, corresponding zone, and new total sample size n^* will be reported for three interim analyses (see Section 4.4.2.1) for all trials, and n^* will be presented graphically after every patient.

4.4.2.4 Combination test design

The combination test design is reported in detail in Section 3.11.1. The design requires prior specification of the parameter γ, which can be thought of as a “tuning parameter”. The sample size is increased only up to the point where CP is increased by $< \gamma$ by adding one more observation (Jennison 2015). The objective function $CP_{\tilde{d}}(z_1, n^*) - \gamma(n^* - n)$ is then maximised, where the new sample size n^* can take any value between n_{rec} (the number recruited at the point of the interim analysis), and the maximum sample size n_{max}. The new total sample size will then be compared graphically with the other two SSR designs. Due to the in-built “stopping boundaries” in the design (where the trial stops at n_{rec} patients as CP is not increased by at least γ by recruiting one additional patient), no further boundaries are explored in combination with this design.

4.4.2.5 Stepwise design

The stepwise design methodology is described in detail in Section 3.12.1. The stepwise design considered is based on Rule 2 from Liu 2016. Three ‘step’ values r_j have been chosen and will use a symmetrical design, where sample size will be increased and decreased using the same r_j values. Provided the lower limit of the interval provided in Equation 3.16 is satisfied for the smallest r_j, all other interval values are free to be chosen by the user. The same values of n_{max} will be considered for all designs (1.5 and 2 times the original sample size). Stepped increments of sample size increase will be spread equally between 1 and n_{max} such that values of $r_j = 1.16, 1.33, 1.5$; and $r_j = 1.33, 1.67, 2$ for the two values of n_{max} investigated respectively.

CP_{min} values will again be calculated after every possible n_1 value, using a maximum increase of r_1 for each design (i.e. 1.16, or 1.33 times the original sample size for the two n_{max} values). Intervals for sample size increases have been standardised as much as possible across designs, and therefore only differ in terms of CP_{min}. The stepwise design is fully defined as:

$$\begin{aligned}(n^*/n) = {} & I(CP(z_1) \leq CP_{min}) \\ & + r_1 * I(CP(z_1) \in (CP_{min}, \frac{1}{2}(CP_{min} + 0.6)]) \\ & + r_2 * I(CP(z_1) \in (\frac{1}{2}(CP_{min} + 0.6), 0.6]) \\ & + r_3 * I(CP(z_1) \in (0.6, 0.7]) \\ & + r_2 * I(CP(z_1) \in (0.7, 0.75]) \\ & + r_1 * I(CP(z_1) \in (0.75, 0.8]) \\ & + I(CP(z_1) > 0.8)\end{aligned}$$

Similarly to the promising zone design, CP_{min} values and corresponding new total sample size n^* will be reported for the three interim analyses, and n^* will be presented graphically after every patient.

4.4.2.6 Futility bounds

Whilst the literature surrounding stopping boundaries based on CP values have suggested futility boundaries between 20-40%CP, using these definitions of futility would cause the trial to stop in some cases even at the lower end of the promising zone (Sully 2014). As these designs also incorporate a method to maintain power by increasing the samples size, the futility boundary may only be beneficial when researchers when there is overwhelming evidence that the trial is futile to continue. Whilst alternative futility boundaries may be considered, time constraints mean that only one boundary will be investigated, and a 10% boundary has been chosen for this reason.

For all designs, a 10% futility bound will be considered. If the CP value at the interim analysis falls below this limit, the trial would be stopped for futility and would not continue, even to the original sample size. Again, due to time constraints, no efficacy boundaries are considered for any design for the analyses in this thesis.

4.4.2.7 Stability of the estimate

The stability of the estimate was investigated in order to understand when might be the earliest time that an interim analysis could take place. For this, the end result of the original trial analysis was taken as the "true" treatment effect. First, patients were assumed to be recruited in the same order as the original trial, and the treatment effect estimated after every 10 patients. This is hereafter referred to the 'original sequential order'. Additionally, the treatment effect calculated after every 10 patients recruited in reverse order was investigated. In this scenario the last patient is now first, and first patient now last. Finally, the trial was randomly re-ordered 1000 different times, and again estimates were calculated after every 10 patients. The median estimate was taken and plotted with the original sequential order for comparison. Additionally, the 97.25/2.5 and 75/25% quantiles of the 1000 random orders were plotted. The original, reverse and the median of the random order estimates were compared graphically. The aim of this investigation was to see any bias in the order of patient recruitment, particularly if the “best” site opens first, or there is a rush for recruitment to finish on time at the end of the trial.

4.4.3 Methods

4.4.3.1 Data

Data has been specifically chosen to cover a range of scenarios in trial design: continuous and binary endpoints, publicly funded and industry settings, size of study (‘small’ and ‘large’) and time to primary outcome data availability (‘short’, ‘medium’ and ‘long’). No strict definitions have been used for classifying size of study or time to primary outcome, but are meant to be comparative to each other. With the exception of one study, all ‘small’ studies are ≤ 400 patients in total. ‘Short’ studies have outcome data available by 6 weeks at the latest, ‘medium’ outcomes range between 3-4 months (with one exception), and ‘long’ studies all have outcomes at $\geq$ 12 months.

Details of each individual trial are provided in Appendix B and Appendix C for background details and results respectively. Only the primary outcome will be used in the data re-analysis and no secondary outcomes will be assessed. Analyses are based on an Intention to treat (ITT) principle, as randomised group was more readily available than information on treatment received.

4.4.3.2 Reusing data

Due to the lengthy data collection process, not all data was able to be obtained in both publicly funded and industry settings (details in Sections 4.3.2 and 4.3.1). In order to still cover the original scenarios (data type, time to endpoint and size of trial), some data was used more than once. If data for another outcome was also provided (e.g. a co-primary outcome), then this data was used, with the time point being reframed (e.g. co-primary outcome 1 being used at the original time point, and co-primary outcome 2 imagined at a later date). However, this was not always possible, and some trials use the same data, imagined at a different time point. For full details of how datasets were used and re-used, see Table 4.1.

4.4.3.3 Modelling randomisation dates

Some studies were not willing to provide a randomisation date with the data. In order to work out pipeline patients (particularly affecting the combination design), dates were assigned a date within the recruitment period, while maintaining the same sequential order of patients provided. The number of months of the trial was first calculated, and month number was randomly generated *n* times. For instance, if a study started in September 2018 (month 1), and ended in January 2020 (month 17), then a random number between 1 and 17 would be generated *n* times (i.e. one for each patient in the trial). The day was then randomly generated based on the month - e.g. between 1 and 31 if the month was March, but only between 1 and 30 for April. Together these made a full date, which were ordered and then merged with the trial data set. There was one exception to this method, due to all recruitment taking place over a 12 day period. For this trial, a number between 1 and 12 was randomly generated *n* times, and this was converted to the corresponding date.

If a "day of primary outcome collected" was available, the primary data available date was based from that information. For instance, if "Day 28", 28 days were added to their modelled randomisation date. Otherwise, the same number was added to all randomisation dates (e.g. randomisation date + 365 days for a 1 year outcome).

4.4.3.4 Missing data

Where the derived primary outcome variable is not provided, this have been derived as closely as possible to the original analysis. However if only limited information on rules (e.g. for varying levels of missing data) then some small discrepancies may occur between the original analysis and the re-analysis in this thesis.

Some trials have had a problem with missing data, and may have used methods to overcome this such as multiple imputation in the original analysis. However, this will not be taken into account for the purposes of the data re-analysis. Missing data is reported for each trial in Appendix B.

If numbers available are substantially different to the planned sample size, (for example, if the trial terminated early), power has been adjusted to account for this discrepancy. Treatment estimates have been kept consistent with the original analysis in these cases.

4.4.3.5 Analysis models

For better comparisons between trials, the models were the same for each data type - ANCOVA for all continuous end points, and logistic regression for all binary endpoints. If initial treatment effect estimates were reported as an absolute difference in proportions, these were converted to the odds ratio scale using 2x2 tables, and the Standard Error (SE) calculated as $SE[log(OR)] \approx \sqrt{(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d})}$. If applicable, non-inferiority limits were also converted to the log scale. All models were adjusted for any covariates specified in the original analysis if data was available.

4.5 Case study: The IMPROVE trial

Due to the large number of datasets obtained, the IMPROVE trial is presented here as an example of the background information obtained, details of recruitment and original analysis and results, and missing data. Appendix B gives all trials summarised in the same manner.

4.5.1 Background

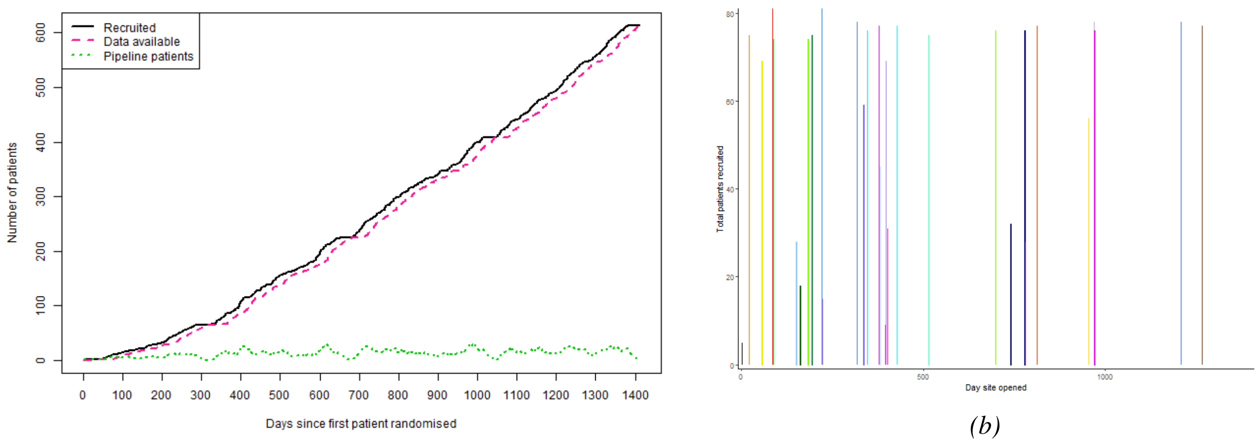
The aorta is the main blood vessel carrying oxygenated blood to the rest of the body. An Abdominal Aortic Aneurysm (AAA) is a swelling of the aorta in the abdomen, and, left untreated, can enlarge and eventually rupture (Ulug 2018). A ruptured AAA is a common cause of death, with many not reaching hospital in time, and even with surgery, only about half make it out of hospital alive. Surgical intervention is typically an open repair, and those that survive have a lengthy recuperation time. It is thought that a keyhole surgery intervention, endovascular repair, may shorten recovery time and result in a lower 30 day mortality rate compared to open repair.

4.5.2 Trial details

The IMPROVE trial (The Immediate Management of the Patient with Rupture: Open Versus Endovascular repair trial) aimed to compare these two surgical techniques. The primary outcome was mortality at 30 days. To detect a risk difference of 14.3%, assuming a mortality rate of 44.7% in the open repair group, and 30.4% in the endovascular repair group, with two sided 5% significance, 94% power, and 5% loss to follow up, required a total of 600 participants. The treatment effect was converted to an OR for data re-analysis, resulting in an odds ratio of mortality in the intervention group compared to the control group of 0.539, and SE of 0.171.

4.5.3 Results

A total of 613 patients (open repair: n=297, endovascular repair: n=316) were randomised to the IMPROVE study from 31 centres in the UK and Canada. Figure 4.3a shows recruitment rate, and number of patients with data available, or unavailable (pipeline patients). Figure 4.3b shows when sites were opened, and the total number recruited to each. Table 4.2 shows the time to 50% and 100% recruitment of both patients and sites.



(b)

(a)

Figure 4.3: Recruitment to the IMPROVE trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Figure 4.4 shows the missing data for the 30 day outcome and flow of participants

Total days of study	1410
Days to last patient recruited	1380
Days to 50% patient recruitment	811
Number of sites	31
Days to last site recruited	1266
Days to 50% site recruitment	381

Table 4.2: Recruitment summary for the IMPROVE study

through the study. No patients were missing outcome data, age or sex (baseline covariates). Missing baseline data came from the Hardman index covariate data.

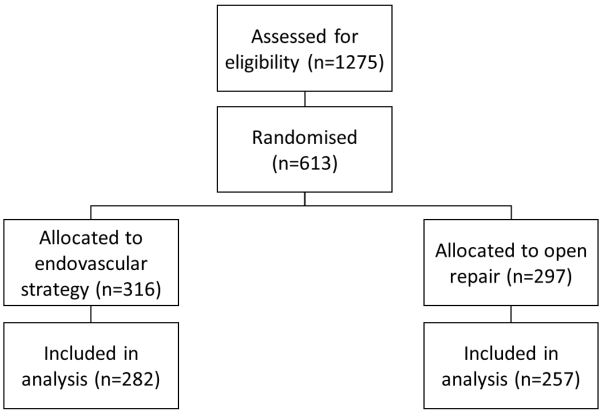
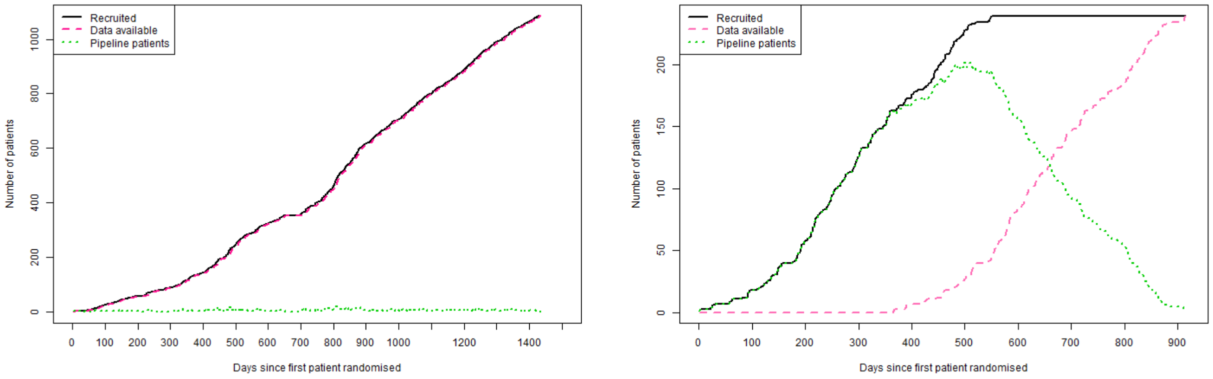

Assessed for eligibility (n=1275)
Randomised (n=613)
Allocated to endovascular strategy (n=316)
Allocated to open repair (n=297)
Included in analysis (n=282)
Included in analysis (n=257)

Figure 4.4: Flowchart of participants in the IMPROVE trial

Primary analysis used a Pearson Chi-squared test to asses proportions of patients surviving to 30 days in each group. A logistic regression was also used to provide an adjusted odds ratio, adjusting for baseline covariates age, sex and Hardman Index. The endovascular repair group resulted in slightly lower odds of death compared to the open repair group, but this was not statistically significant (OR=0.92, 95% CI (0.66, 1.28), p=0.62).

4.6 Additional case studies: Pipeline patients and site recruitment

The IMPROVE trial showed the pipeline patients in a trial with 30 days to outcome data availability. Figure 4.5 shows two further trials: 3MG and the Acupuncture trial. The 3MG trial recruited 1109 patients in 1427 days from 25 sites. The primary outcome was admission to hospital within 1 week of presentation at the emergency department. Due to



(a) *(b)*

Figure 4.5: Recruitment and time to data availability to the (a) 3MG trial and (b) Acupuncture trial. Green (dotted) lines show pipeline patients who have been randomised, but have no data available yet.

the short time to data becoming available, the recruited (black solid) and data availability (red dashed) lines are very similar to each other, and the pipeline patients (green dotted line) remains very low throughout the study.

The Acupuncture trial recruited 241 patients in 549 days across 16 sites. The primary outcome involved pain scores collected at 12 months post-randomisation. Due to the length of time between randomisation and data availability, the number of pipeline patients (green dotted line) is quite high until approximately day 500, from which point it steadily decreases. Therefore almost all patients have been randomised by the time any data has become available. Trials in this situation may not benefit from any design which allows a decrease in sample size, such as the combination test or any design with a stopping boundary. Additionally, if a sample size increase is found to be necessary through any of the designs, the trial would be facing a lengthy time delay, as patient recruitment would need to be restarted, and the full follow up time added to the total duration of the study. Another interesting trial in terms of recruitment is the Flu vaccine trial, shown in Figure 4.6, which recruited all 2249 patients over just a 12 day period across 38 sites. Even with a 28 day timeframe to outcome data availability, all patients have been recruited before any data becomes available, and this trial would be in a similar situation to the Acupuncture study mentioned above, had a SSR been involved in the original trial.

The IMPROVE trial showed an example of many sites being recruited throughout the

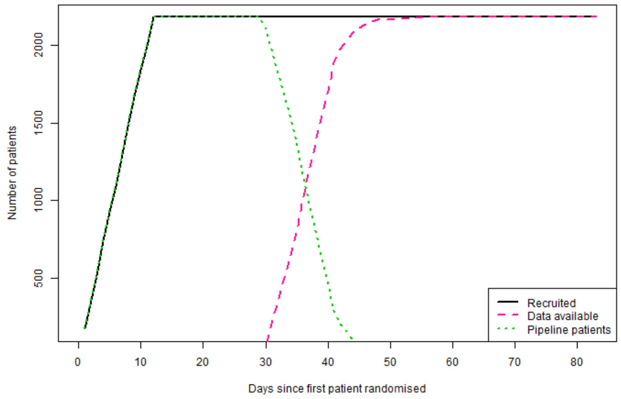
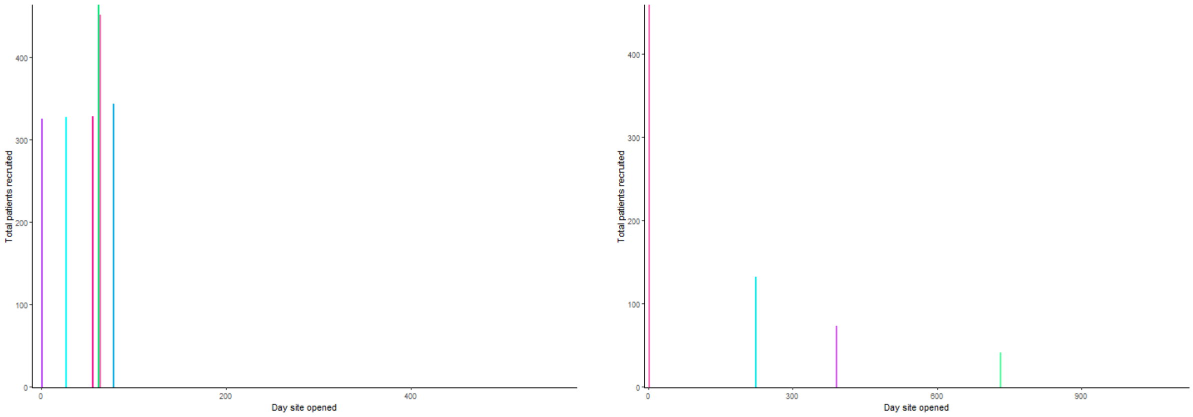


Figure 4.6: Time between patient recruitment and primary outcome data became available in the flu vaccine trial (all strains).

study duration and, with the exception of a few lower recruiting sites, recruited similar number of patients regardless of when they started recruiting (e.g. 2 sites that first started recruiting after day 1000, and yet were still two of the top recruiting sites). Figure 4.7 provides two additional examples of site recruitment. The RATPAC study (4.7a) recruited all their sites early on in the trial, and all sites recruited at least 320 patients (maximum=464). On the other hand, the CASPER trial had one predominant recruiting site that was the first to open and recruited 65% of the total patients. Three additional sites that started recruiting later into the trial recruited the remaining patients.



(a)

(b)

Figure 4.7: Site recruitment to the (a) RATPAC trial and (b) CASPER trial. The x-axis shows time to first patient recruited in each site, and y-axis represents the total number recruited in each site.

4.7 Overview of all trials

This section provides a summary of the 21 unique outcomes from 14 trial datasets reanalysed. Trials may be summarised at the trial, outcome, or recruitment level populations:

- Population 1: All original trials only. No reframed time points or additional outcomes analysed.
- Population 2: All trials (original or reframed) with distinct outcome data (e.g. CASPER MINUS uses the 4 month data from CASPER and so is included, but 3MG (1 year) uses exactly the same data as 3MG (1 week) and so is not included)
- Population 3: Trials that have distinct time to primary outcomes from recruitment data. The binary outcomes from the Flu vaccine trial have been excluded as these sub-trials all have the same recruitment date and time to primary outcome as their corresponding continuous sub-trials (i.e. one per strain)

This is further illustrated by Table 4.3.

			Population 1	**Population 2**	**Population 3**
			Original trial	Distinct outcome data	Distinct pipeline data
Continuous	Publicly Funded	FAST INdiCATE	✓	✓	✓
		SELF	✓	✓	✓
		Acupuncture	✓	✓	✓
		Casper MINUS	✗	✓	✓
		Casper PLUS	✓	✓	✓
		Casper Minus	✓	✓	✓
	Industry	Epilepsy (1 day)	✗	✗	✓
		Epilepsy (19 weeks)	✓	✓	✓
		Epilepsy (1 year)	✗	✗	✓
		Flu A/H1N1	✓	✓	✓
		Flu A/H3N2	✗	✓	✓
		Flu B1	✗	✓	✓
Binary	Publicly Funded	IMPROVE	✓	✓	✓
		Corn plasters	✓	✓	✓
		AMAZE	✓	✓	✓
		3MG (1 week)	✓	✓	✓
		RATPAC	✓	✓	✓
		3MG (1 year)	✗	✗	✓
	Industry	Nasal sprays	✓	✓	✓
		Mencevax A	✓	✓	✓
		Mencevax C	✗	✓	✓
		Flu A/H1N1	✗	✓	✗
		Flu A/H3N2	✗	✓	✗
		Flu B1	✗	✓	✗
		(TOTAL=24)	**(TOTAL=14)**	**(TOTAL=21)**	**(TOTAL=21)**

Table 4.3: A summary of information available for each of the 14 trials obtained for data re-analysis, reframed in 24 ways.

	Short (N=5)	Medium (N=6)	Long (N=3)	All (N=14)
	Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)
Number of patients (planned)	600 (300, 1200)	228 (130, 484)	400 (240, 704)	350 (240, 704)
Number of patients (recruited)	613 (300, 1109)	232 (133, 485)	352 (241, 705)	326 (241, 705)
Planned loss to follow up (%)	10.0 (5.0, 12.3)	15.0 (10.0, 20.0)	15.0 (15.0, 26.0)	15.0 (10.0, 15.0)
Observed loss to follow up (%)	8.0 (3.1, 12.1)	11.3 (7.5, 19.8)	18.8 (10.8, 27.4)	11.4 (7.5, 18.8)
Number of sites	25.0 (12.0, 31.0)	3.5 (3.0, 6.0)	11.0 (4.0, 16.0)	6.5 (3.0, 16.0)
Total study duration (days)	1241.0 (765.0, 1410.0)	696.0 (490.0, 869.0)	1126.0 (914.0, 2381.0)	891.5 (567.0, 1241.0)
Planned power	90.0 (90.0, 92.0)	80.0 (80.0, 80.0)	80.0 (80.0, 90.0)	80.0 (80.0, 90.0)
Study type				
Superiority	4/5 (80%)	5/6 (83%)	3/3 (100%)	12/14 (86%)
Non-inferiority	1/5 (20%)	1/6 (17%)	0/3 (0%)	2/14 (14%)
Non-inferiority limit	0.1 (0.1, 0.1)	0.1 (0.1, 0.1)		0.1 (0.1, 0.1)
Data type				
Continuous	2/5 (40%)	3/6 (50%)	2/3 (67%)	7/14 (50%)
Binary	3/5 (60%)	3/6 (50%)	1/3 (33%)	7/14 (50%)
Effect size (planned)	0.23 (0.14, 0.29)	0.37 (0.31, 0.55)	0.31 (0.30, 0.52)	0.31 (0.23, 0.42)
Funding type				
Publicly funded	3/5 (60%)	4/6 (67%)	3/3 (100%)	10/14 (71%)
Industry	2/5 (40%)	2/6 (33%)	0/3 (0%)	4/14 (29%)
Reached statistical significance?	2/5 (40%)	4/6 (67%)	2/3 (67%)	8/14 (57%)

Table 4.4: Trial characteristics of the 14 original trial datasets (Population 1)

	Short (N=7)	Medium (N=8)	Long (N=6)	All (N=21)
	Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)
Observed loss to follow up (%)	8.0 (3.1, 16.7)	9.0 (3.1, 16.0)	8.8 (2.7, 18.8)	8.0 (3.1, 16.7)
Effect size planned	0.29 (0.14, 0.40)	0.37 (0.23, 0.47)	0.31 (0.30, 0.40)	0.31 (0.23, 0.40)
Effect size observed	0.15 (0.07, 0.20)	0.24 (0.17, 0.39)	0.29 (0.26, 0.35)	0.22 (0.15, 0.32)
Reached statistical significance?	4/7 (57%)	6/8 (75%)	5/6 (83%)	15/21 (71%)

Table 4.5: Effect sizes, observed loss to follow up and statistical significance of the original trial from the 21 unique outcomes from the 14 trial datasets (Population 2)

Table 4.4 summarises characteristics of the 14 original trials, split by length to primary outcome data availability (“short”/“medium”/“long”). Most trials recruit to target, with a few exceeding their goals but only by a few patients. Three trials stopped recruiting early due to slow recruitment (one trial also showed efficacy). Short outcome trials had the highest median planned and recruited sample size figures, which explains why the longest median study duration in days is actually in the short outcome group. The median planned power was 80%, although of the 5 short outcome trials, the median was 90%. As time to outcome increased, the median planned and observed loss to follow up also increased, with one long outcome trial seeing 27.4% loss to follow up. All trials had at least 2 sites, and the maximum was 38. No original industry trials were in the long outcome category.

Almost all studies were superiority studies, with only the Flu and Mencevax studies having a non-inferiority endpoint.

At the outcome level (Population 2), observed loss to follow up was broadly similar across the three outcome length categories. Standardised observed effect sizes were calculated according to Table 1 in Rothwell *et al.*, which is able to summarise continuous and binary effect sizes in a comparable manner (Rothwell 2018), and are presented in Table 4.5. Whilst all outcome categories saw a decrease in median between planned and observed effect sizes, six trials actually saw a bigger effect than originally planned. Rothwell *et al.* also saw a decrease between target and effect sizes from an audit of 102 RCT reports.

	Short (N=7)	Medium (N=7)	Long (N=7)	All (N=21)
	Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)
Time (days) to:				
End of study	946.0 (765.0, 1410.0)	567.0 (234.0, 869.0)	1126.0 (444.0, 1792.0)	869.0 (490.0, 1241.0)
50% patient recruitment	482.0 (387.0, 811.0)	264.0 (79.0, 419.0)	419.0 (79.0, 847.0)	419.0 (238.0, 482.0)
100% patient recruitment	945.0 (761.0, 1380.0)	490.0 (199.0, 779.0)	761.0 (199.0, 1427.0)	761.0 (477.0, 945.0)
50% site recruitment	204.0 (7.0, 381.0)	22.0 (1.0, 56.0)	114.5 (1.0, 349.0)	54.0 (1.0, 222.0)
100% site recruitment	220.0 (176.0, 1262.0)	145.0 (15.0, 281.0)	472.0 (15.0, 1006.0)	212.0 (15.0, 732.0)
Percentage of pipeline patients at interim analysis				
25%	1.3 (0.0, 3.0)	21.6 (13.5, 37.2)	58.2 (30.0, 74.3)	24.1 (3.0, 58.2)
50%	2.1 (0.0, 4.9)	21.4 (12.0, 39.8)	48.7 (32.4, 49.8)	21.6 (4.9, 48.5)
75%	3.1 (0.6, 8.2)	21.5 (20.0, 23.9)	24.4 (22.6, 24.9)	21.5 (8.2, 24.2)
Percentage of patients left to recruit at interim analysis				
25%	71.9 (69.0, 72.8)	53.2 (37.5, 55.9)	16.7 (0.0, 43.3)	50.0 (16.7, 69.0)
50%	45.9 (45.1, 47.8)	28.1 (10.0, 33.8)	0.0 (0.0, 16.5)	27.8 (0.0, 45.1)
75%	21.9 (16.7, 22.6)	3.4 (1.0, 4.7)	0.0 (0.0, 0.0)	3.4 (0.0, 16.7)

Table 4.6: Summary of recruitment details for the distinct timeline information from the 14 trials (Population 3)

Table 4.6 gives details about the recruitment original trials and reframed timepoints (Population 3). As discussed previously from Population 1, the short trial outcome time category has the longest median recruitment time duration (945 vs 490 vs 761 days), although no longer has the longest median total study duration (946 vs 567 vs 1126 days). Unsurprisingly, short outcomes have much smaller percentages of pipeline patients, and larger percentages of patients still left to recruit. At the 75% interim analysis, medium and long outcomes both see very smaller pipeline patient percentages, which is explained by the next section of the table; having seen very small percentages of patients still left to recruit.

Overall, the selection of trials seem representative of trials in the general population. Despite being chosen specifically for length to data availability, sample size and outcome type, these trials will be also be able to provide information about SSR implementation with varying recruitment rates (recruitment periods ranging from 12 days (Flu vaccine trial) to 2016 days (AMAZE)), and trial outcomes (having seen decreased, approximately equal and increased effect sizes to that originally planned). Pipeline patients and percentage left to recruit at interim analyses will help provide information about SSR design impact, such as few/no decreases being seen in trials with a longer time to primary outcome.

4.8 Summary

This chapter has presented the full analysis plan for the retrospective data re-analysis, provided case studies of patient recruitment rates and site recruitment, and summarised all trials that will be used in the thesis. Chapter 5 presents the results from the data re-analysis, including CP plots, the impact of new sample size n^* for three SSR designs, and an investigation of estimate stability.

5 | Retrospective data analysis

5.1 Introduction

Chapter 4 describes in detail 21 distinct outcomes from 14 different trial datasets that have been obtained for re-analysis in this thesis. This chapter presents the results from the retrospective data re-analysis through case studies, and an overall summary of results. The re-analysis is split into two main sections. First, case studies will be presented to illustrate CP calculations, n^* comparisons from three designs, an in depth look at three chosen interim time points (25, 50 and 75% data available), with 2 values of n_{max} (1.5*n and 2*n), and the stability of the treatment estimate, which could help to inform the timing of an interim analysis. The second part of this chapter provides an overall summary of all trials re-analysed. Similarly to Chapter 4, all trial results are included in Appendix C, with only a few examples discussed in this chapter.

5.2 Aims and objectives

Specific aims of this chapter include:

- Present case studies to illustrate the choice of future treatment effect in CP calculations
- Compare three SSR designs and the impact on n^* values for each
- Investigate three interim time points and two values of n_{max}, and the decisions that would have been made in each trial
- Evaluate four boundaries for estimating stability and observe where these are met for each trial

5.3 Case studies

This section provides case studies of CP values plotted after every patient from patient 10 onwards (unless otherwise stated) using four treatment effect assumptions: current trend (black), 80% optimistic limit (blue), 90% optimistic limit (green) and the original hypothesised treatment effect (H_1) (red). Grey vertical lines represent the three interim analyses investigated: solid lines show the number of patients with data available, and dashed lines show the number recruited at each interim time point. The shaded areas represent the zones if a promising zone design were being considered, using CP_{min} calculated with 2 values of n_{max}. The upper grey block represents the favourable zone (any value above 1-β), and the two lower grey blocks represent the unfavourable zone for the two values of n_{max}: an increase of 50% and 100% respectively. The white horizontal line represents the futility boundary at 10% (if using).

New total sample size n^* plots compare the three SSR designs investigated: promising zone (pink), combination test (green), and stepwise (blue). If the same data is used at more than one time point (e.g. the Epilepsy trial), the combination test design yields different results for each as it depends on the number of patients recruited, and is also plotted (in orange and/or purple).

For all trials, the treatment estimate calculated after every 10 patients is plotted: pink for the original sequential order of patients, purple for the median estimate of 1000 random orders of trial patients, and blue for the reverse order of the original sequential order (see Section 4.4.2.7 for full details). Confidence intervals are provided for original order and reverse order, slightly offset to better distinguish between the two. Dashed purple lines represent 2.5, 25, 75 and 97.5 percentiles. The estimate at the end of the original trial, assumed to be the "true" estimate, is shown as a black dotted line. A black dashed line represents the hypothesised treatment effect prior to the start of the trial. Four boundaries are investigated for describing the estimate as stable in terms of ±Standard Deviation (SD): if it lies within ±0.05*SD (green) ,±0.1*SD (pale green), ±0.2*SD (yellow) or ±0.3*SD (pale pink) from the original trial estimate. Both the first time the estimate enters each boundary, and the instance where the estimate enters the boundary and remains there for the rest of the trial are reported in Section 5.4 for all trials combined.

5.3.1 Case study: The CASPER PLUS trial

The original CASPER PLUS trial the study showed a mean difference of 1.92 points (95% CI 0.85-2.99, p<0.001) in favour of the collaborative care group (Bosanquet 2017). Figure 5.1 shows CP values for the CASPER PLUS trial, re-imagined with an adaptive uSSR design. The hypothesised effect line remains consistently high (≈1) throughout the trial. The other three lines start at zero, and gradually increase to 1, fluctuating between zones before this point (∼430 patients). The current trend line remains below the futility boundary (0.1) until patient 125, unfavourable until 176, varying between all three zones until patient 350, where it remains in the favourable zone. The optimistic limits rapidly alter between all zones for the first 120 patients. After this point, the 90% limit remains in the favourable zone, and the 80% limit dips twice into the promising zone. Table 5.1 summarises the number of times each line falls into each zone, when CP is calculated after every patient's data becomes available.

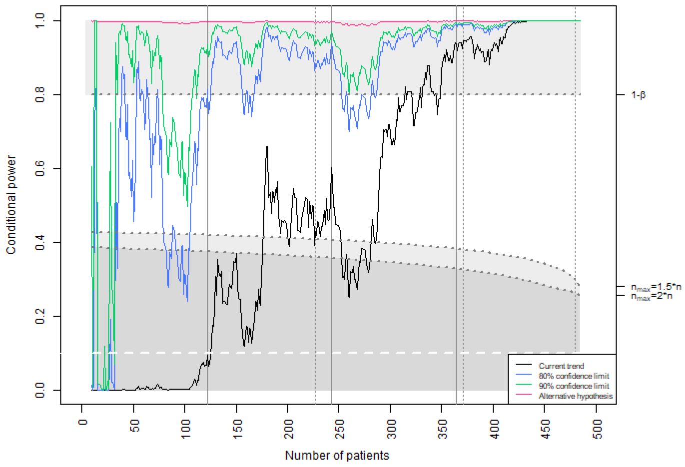


Figure 5.1: Conditional power calculated after every patient in the CASPER PLUS trial

Table 5.2 presents decisions for the three specified interim analyses. Promising zone and stepwise designs only increase sample size at 50% data available, using the current trend assumption. With n_{max}=1.5, both designs reach the maximum increase of 50%, and with

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	155	476	334	425
	Promising	119	0	97	34
	Unfavourable	86	0	25	4
	(Futility)	116	0	20	13
n_{max}=2	Favourable	155	476	334	425
	Promising	132	0	104	34
	Unfavourable	73	0	18	4
	(Futility)	116	0	20	13

Table 5.1: Number of times CP values fall in each zone for the promising zone design for the CASPER PLUS trial. For a design where no futility boundary is considered, these values become unfavourable instead.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 121	n_{max}=728	0.093	0.419	Unfavourable	485	227	0.466	485
	Recruited:227	n_{max}=970	0.093	0.374	Unfavourable	485	227	0.44	485
50%	Available: 243	n_{max}=728	0.602	0.406	Promising	728	485	0.46	728
	Recruited: 371	n_{max}=970	0.602	0.357	Promising	763	485	0.43	970
75%	Available: 364	n_{max}=728	0.938	0.382	Favourable	485	480	0.447	485
	Recruited: 480	n_{max}=970	0.938	0.328	Favourable	485	480	0.41	485
HYPOTHESISED EFFECT									
25%	Available: 121	n_{max}=728	0.997	0.419	Favourable	485	326	0.466	485
	Recruited:227	n_{max}=970	0.997	0.374	Favourable	485	326	0.44	485
50%	Available: 243	n_{max}=728	0.998	0.406	Favourable	485	371	0.46	485
	Recruited: 371	n_{max}=970	0.998	0.357	Favourable	485	371	0.43	485
75%	Available: 364	n_{max}=728	0.999	0.382	Favourable	485	480	0.447	485
	Recruited: 480	n_{max}=970	0.999	0.328	Favourable	485	480	0.41	485
80% OPTIMISTIC LIMIT									
25%	Available: 121	n_{max}=728	0.812	0.419	Favourable	485	496	0.466	485
	Recruited:227	n_{max}=970	0.812	0.374	Favourable	485	496	0.44	485
50%	Available: 243	n_{max}=728	0.938	0.406	Favourable	485	441	0.46	485
	Recruited: 371	n_{max}=970	0.938	0.357	Favourable	485	441	0.43	485
75%	Available: 364	n_{max}=728	0.989	0.382	Favourable	485	480	0.447	485
	Recruited: 480	n_{max}=970	0.989	0.328	Favourable	485	480	0.41	485
90% OPTIMISTIC LIMIT									
25%	Available: 121	n_{max}=728	0.935	0.419	Favourable	485	425	0.466	485
	Recruited:227	n_{max}=970	0.935	0.374	Favourable	485	425	0.44	485
50%	Available: 243	n_{max}=728	0.971	0.406	Favourable	485	415	0.46	485
	Recruited: 371	n_{max}=970	0.971	0.357	Favourable	485	415	0.43	485
75%	Available: 364	n_{max}=728	0.994	0.382	Favourable	485	480	0.447	485
	Recruited: 480	n_{max}=970	0.994	0.328	Favourable	485	480	0.41	485

Table 5.2: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

n_{max}=2, promising zone increases by 78%, compared to the maximum of 100% increase for the stepwise design. The combination test would have increased in just one scenario (80% optimistic limit, 25% data available) by just 11 patients (2%). The largest decrease occurs at 25% data available under the current trend, to 47% of the original sample size planned. Figure 5.2 compares n^* under each assumption after each patient has data available (i.e. after every value of n_1 from patient 12 onwards). No increase is seen at any point using the hypothesised effect assumption. The combination test design would decrease sample size at any interim point before 370 patients, after which the sample size would remain at the originally planned sample size. Large fluctuations in sample size can be seen in the other three assumptions in all three designs: predominantly early on in the trial for the optimistic limits, and mainly in the middle part of the trial for the current trend. This is likely due to the fluctuations in observed CP early on in the trial. An additional spike in sample size can be seen using the 80% limit between 250 and 300 patients, with the largest rise in sample size being seen for the stepwise design (blue).

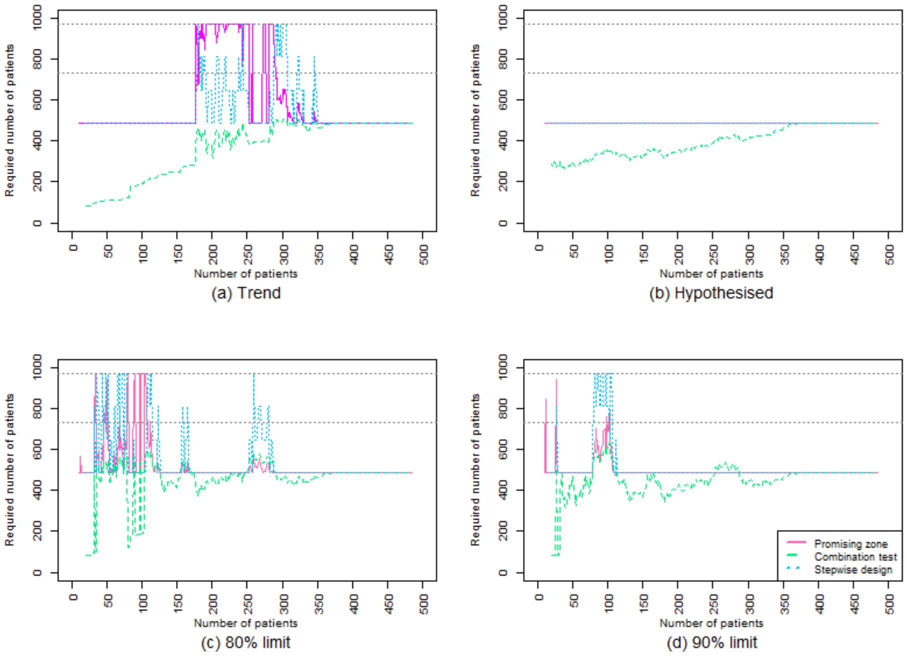


Figure 5.2: Comparison of three SSR designs for the CASPER PLUS trial data

Figure 5.3 shows the sequential and reverse order estimates calculated from patient 20 onwards, after every 10 patients. The original order estimate here starts far below the original analysis treatment effect. with even the upper boundary of the 95% Confidence Interval

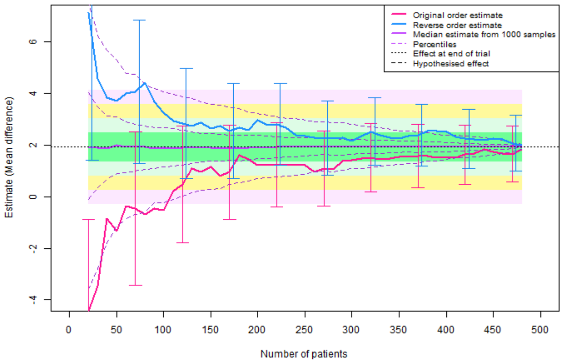


Figure 5.3: A comparison of sequential order, reverse order and the median of 1000 random orders in terms of stability of the estimate for the CASPER PLUS study

(CI) not reaching the largest boundary investigated, ±4*SE. However, from this point forward it increases, reaching the ±4*SE, ±3*SE ,±2*SE, and ±1*SE boundaries by 110, 120, 130 and 180 patients respectively. From 300 patients onwards (62% through the trial), the estimate is within ±1*SE of the end estimate, and remains there. The reverse order however, takes longer to first reach the ±1*SE boundary (240 patients), and to remain in this boundary (410 patients onwards, 85% through the trial).

5.3.2 Case study: The IMPROVE trial

Figure 5.4 shows CP calculated after every patient assuming four different treatment effects. The hypothesised treatment effect assumption line starts at 1 and gradually decreases, leaving the favourable zone by patient 167, the promising zone by 257 (n_{max}=1.5*n) or 269 (n_{max}=2*n). This line (red) has the highest CP values at any value of n_1 (x-axis). The current trend line starts very low, and remains always below the 10% futility bound, except for one instance at 20 patients, which reaches the promising zone if n_{max}=2 (otherwise is classed as the unfavourable zone). The optimistic limit assumptions result in large fluctuations early in the trial, but settles to almost zero by the second interim time point (50% data available). Both limits have a spike around 110-200 patients, but the 80% line stays within the unfavourable zone for wither value of n_{max}. The 90% optimistic limit line lies in the promising zone a total of 35 times between patient 125 and 177 for n_{max}=1.5, and 45 times in the same interval for n_{max}=2. Table 5.3 summarises the number of times each line falls into the four zones (with futility zone being included with unfavourable if no stopping boundary is being used).

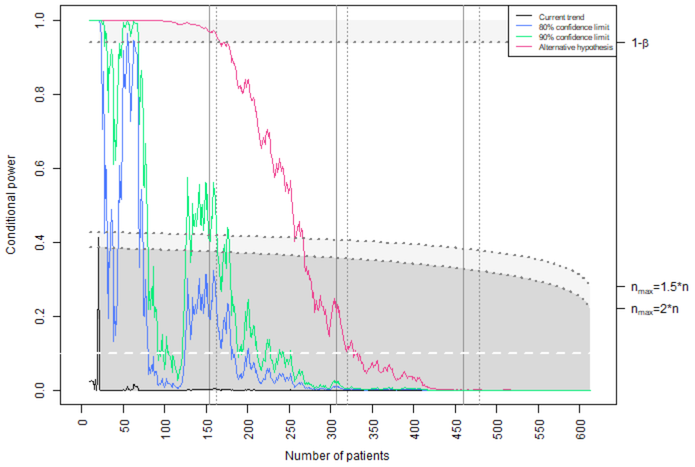


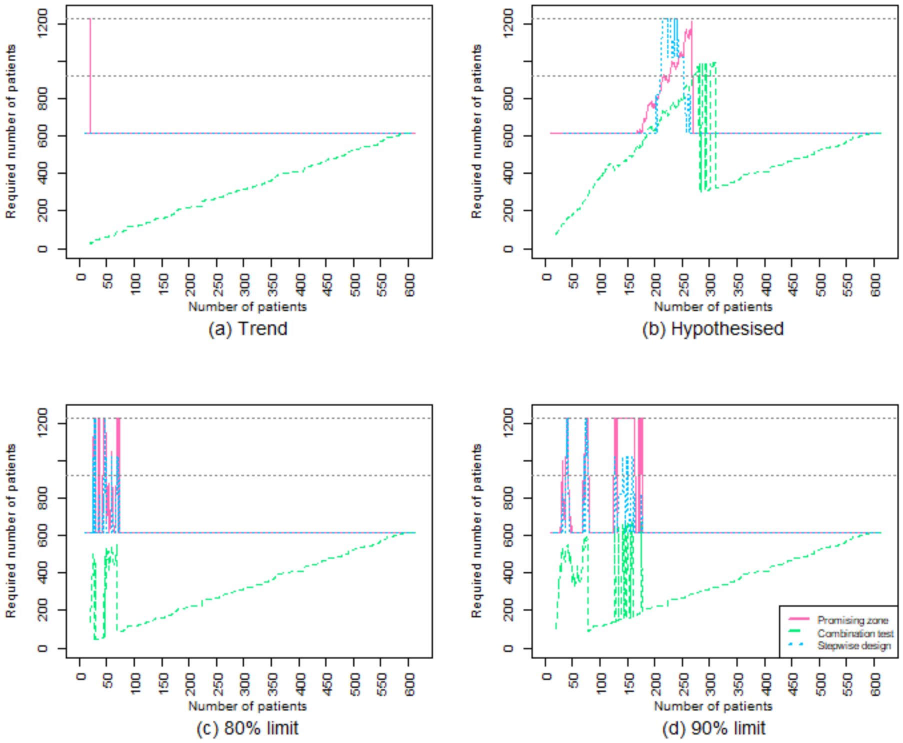
Figure 5.4: Conditional power calculated after every patient in the IMPROVE trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	0	159	18	40
	Promising	0	97	32	65
	Unfavourable	1	63	79	85
	(Futility)	603	285	475	414
n_{max}=2	Favourable	0	159	18	40
	Promising	1	100	35	74
	Unfavourable	0	60	76	76
	(Futility)	603	285	475	414

Table 5.3: Number of times CP values fall in each zone for the promising zone design for the IMPROVE trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table 5.4 presents three specified interim analysis time points (25, 50 and 75% patients with data available), comparing decisions made and new total sample size from the three SSR designs. No sample size increase would have been seen for any time point using the stepwise design, and all values of $n^* = n = 613$, the original planned sample size. One increase (n^*=1226) can be seen using the promising zone design, assuming a 90% optimistic limit and maximum increase of twice the original sample size and 25% data available interim time point. The new sample size from the combination test ranges from a decrease in sample size of 74% of the originally planned sample size, n, (seen at 25% data available assuming current trend or either optimistic limit), to an increase in sample size of 62%, seen at 50% data available under the hypothesised effect assumption and a n_{max}=2, following a CP value of 21%.

Figure 5.5 compares n^* values for all n_1 values, for three SSR designs. Other than one sharp peak of n^* for the promising zone at the 20 patient point discussed previously, no increase in sample size can be seen under the current trend assumption. The combination test would have decreased the sample size for every n_1 value using the current trend or optimistic 80% interval, and almost every n_1 value under the other two assumptions. The largest increase in sample size from the combination test design is from the hypothesised treatment effect assumption, from ≈ 200-320 patients. The promising zone and stepwise designs also see an increase in this range under the same assumption, but return to no increase in sample size sooner.



(a) Trend

(b) Hypothesised

(c) 80% limit

(d) 90% limit

Figure 5.5: Comparison of three SSR designs for the IMPROVE trial data

Figure 5.6 shows the estimate calculated after every 10 patients in the original sequential order and reverse order of the trial dataset from patient 40 onwards, with 4 investigated

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 154	n_{max}=920	0.001	0.419	Unfavourable	613	162	0.466	613
	Recruited:162	n_{max}=1226	0.001	0.374	Unfavourable	613	162	0.44	613
50%	Available: 307	n_{max}=920	0	0.406	Unfavourable	613	320	0.46	613
	Recruited: 320	n_{max}=1226	0	0.357	Unfavourable	613	320	0.43	613
75%	Available: 460	n_{max}=920	0	0.382	Unfavourable	613	479	0.446	613
	Recruited: 479	n_{max}=1226	0	0.328	Unfavourable	613	479	0.41	613
HYPOTHESISED EFFECT									
25%	Available: 154	n_{max}=920	0.97	0.419	Favourable	613	494	0.466	613
	Recruited:162	n_{max}=1226	0.97	0.374	Favourable	613	494	0.44	613
50%	Available: 307	n_{max}=920	0.211	0.406	Unfavourable	613	320	0.46	613
	Recruited: 320	n_{max}=1226	0.211	0.357	Unfavourable	613	991	0.43	613
75%	Available: 460	n_{max}=920	0	0.382	Unfavourable	613	479	0.446	613
	Recruited: 479	n_{max}=1226	0	0.328	Unfavourable	613	479	0.41	613
80% OPTIMISTIC LIMIT									
25%	Available: 154	n_{max}=920	0.195	0.419	Unfavourable	613	162	0.466	613
	Recruited:162	n_{max}=1226	0.195	0.374	Unfavourable	613	162	0.44	613
50%	Available: 307	n_{max}=920	0.007	0.406	Unfavourable	613	320	0.46	613
	Recruited: 320	n_{max}=1226	0.007	0.357	Unfavourable	613	320	0.43	613
75%	Available: 460	n_{max}=920	0	0.382	Unfavourable	613	479	0.446	613
	Recruited: 479	n_{max}=1226	0	0.328	Unfavourable	613	479	0.41	613
90% OPTIMISTIC LIMIT									
25%	Available: 154	n_{max}=920	0.409	0.419	Unfavourable	613	162	0.466	613
	Recruited:162	n_{max}=1226	0.409	0.374	Promising	1226	162	0.44	613
50%	Available: 307	n_{max}=920	0.017	0.406	Unfavourable	613	320	0.46	613
	Recruited: 320	n_{max}=1226	0.017	0.357	Unfavourable	613	320	0.43	613
75%	Available: 460	n_{max}=920	0	0.382	Unfavourable	613	479	0.446	613
	Recruited: 479	n_{max}=1226	0	0.328	Unfavourable	613	479	0.41	613

Table 5.4: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

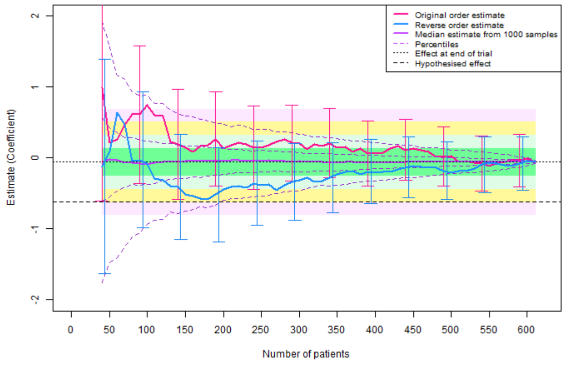
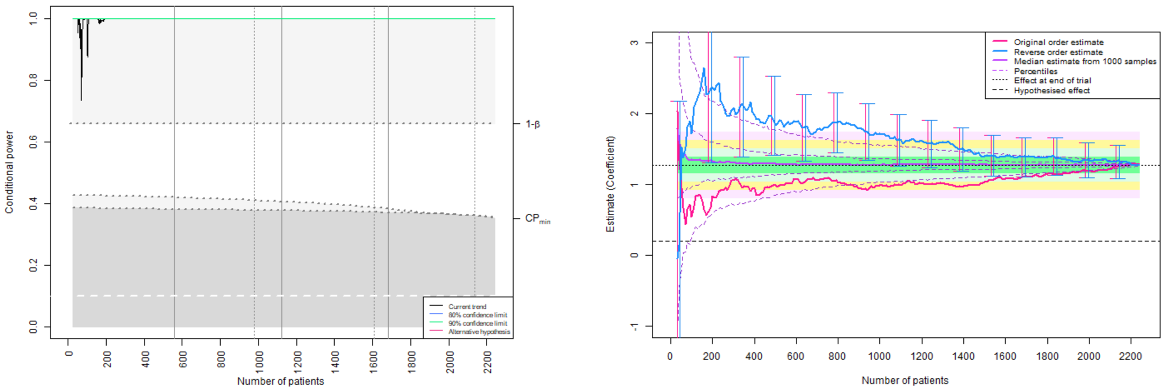


Figure 5.6: A comparison of sequential order, reverse order and the median of 1000 random orders in terms of stability of the estimate for the IMPROVE trial

boundaries for stability definition. The reverse order estimate always lies within the ±4*SEs of the assumed true treatment effect (black dotted line). However, the original order estimate does not, until 110 patients (18% through the trial). The original order estimate first reaches the ±1*SE boundaries later than the reverse order (160, 26% through the trial compared to 40, 7%). However, both then subsequently leave this boundary, and do not remain there until patient 350 (57%) for the reverse order estimate, compared with 460 (75%) for the original order estimate. Note that while the estimate remains above the hypothesised effect line, the model coefficients are on a log scale, and therefore a coefficient of 0 implies a OR of 1, meaning no treatment difference.

5.3.3 Additional case studies

CASPER PLUS presents an example where the original analysis was statistically significant in favour of the intervention (collaborative care) group. However, because the treatment effect was much lower at the start of the trial than at the end, the CP value using the current trend starts near 0, and only rises to 1 nearer the end of the trial. Conversely however, the hypothesised effect line (red) stays almost consistently at 1 throughout the duration of the trial. The IMPROVE trial however was not found to be statistically significant, with a slight over estimation of the treatment effect early on, corresponding to an early spike in CP under the current trend. The hypothesised line also reaches zero, and is much slower to decrease, but reaches the futility boundary by patient 350. There is some fluctuation using either optimistic limit, but both lines reach the futility bound before the hypothesised CP line. Figure 5.7a shows one more CP example using the data from the RATPAC trial. Despite under-estimating the final observed treatment effect during the study, the effect was still much higher than planned throughout the trial duration. Other than one small dip in CP in the very early stages under current trend, CP otherwise remains at 1 at all time points, for all four treatment effect assumptions. This trial was terminated early, with one of the reasons being due to a CP calculation by the DMSC showing efficacy for the primary outcome (>99.9%). It should be noted that as this trials was terminated early, power has been recalculated using the number of available patients for re-analysis.



(a)

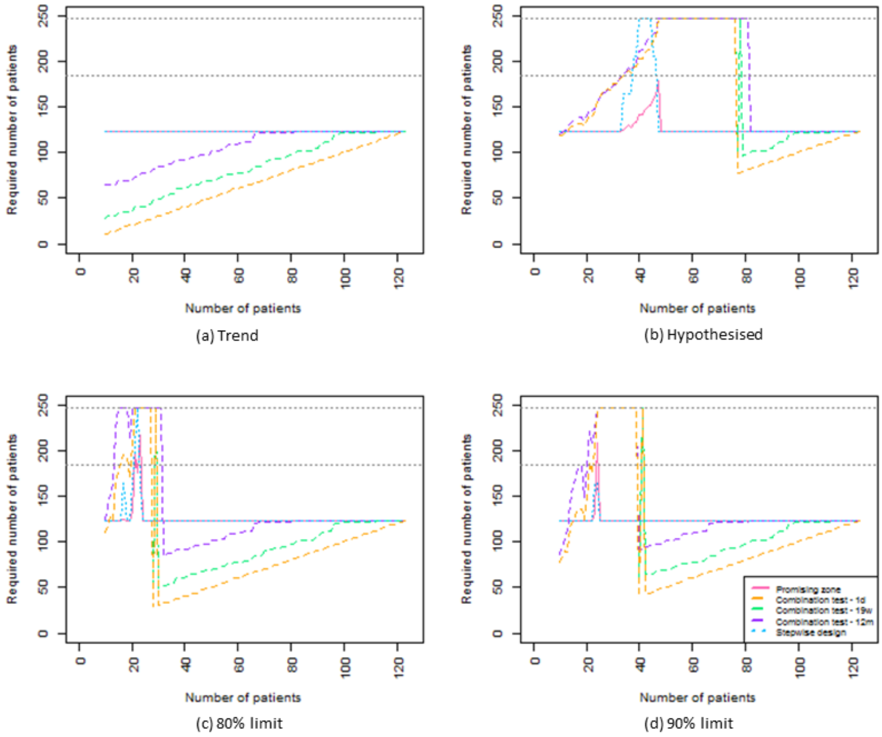
(b)

Figure 5.7: (a) Conditional power calculated after every patient in the RATPAC trial (b) A comparison of sequential order, reverse order and the median of 1000 random orders in terms of stability of the estimate for the RATPAC trial

One further example of n^* required for three designs is provided in Figure 5.8 (Epilepsy study). Due to the multiple time points of the reframed epilepsy trials and therefore a difference in pipeline patients, the combination test yields different n^* decisions for the three time points. Therefore, orange represents the 1 day time point, green for the 19 weeks (original trial analysis) and purple for the 12 month outcome. For this trial, the current trend assumption does not result in an increase at all, with a decrease being seen in all three combination test designs prior to around 80 patients (2/3 through the trial). More increases in sample size can be seen in the hypothesised effect assumption, predominantly in the combination test design case, before establishing a decrease instead later on in the trial. Optimistic values show a similar situation, but increases occur in a much smaller percentage of values of n_1 than in the hypothesised effect.

Figure 5.9 provides two further examples of estimate stability in the FAST INDiCATE and 3MG trials.

The FAST INDiCATE original sequential order starts by under-estimating the treatment effect, but lies within the ± 1*SE boundary by patient 50 (17% through the trial), and remains there for the remainder of the trial. The sequential estimate briefly falls into the ±3*SE boundary at 30 patients, but remains within the ±2*SE limit at all other values of n. Comparatively, the reverse order estimate spends a little longer in the ±3*SE boundary, over-estimating the treatment effect to start. However, it slowly decreases, entering



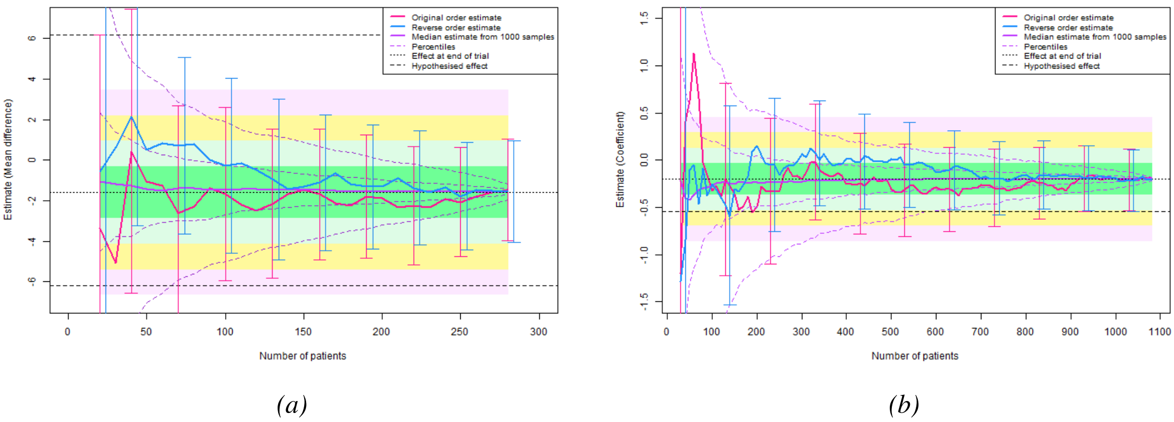
(a) Trend

(b) Hypothesised

(c) 80% limit

(d) 90% limit

Figure 5.8: Comparison of three SSR designs for the Epilepsy trial data



(a)

(b)

Figure 5.9: A comparison of sequential order, reverse order and the median of 1000 random orders in terms of stability of the estimate for two trials: (a) FAST INdiCATE and (b) 3MG

the ±1*SE boundary at 120 patients, and remaining there from that point forwards. Note that two hypothosised lines appear on this graph, as both groups combine an interventional and control group together (Functional Strength Training (FST) vs Movement Performance Therapy (MPT)).

The original estimate of the 3MG trial is highly variable at the start, and does not keep within even the ±4*SE boundaries until 80 patients onwards. After this point, it quickly reaches the ±2*SE boundaries, and other than straying into the next boundary at 190 patients, stays within these limits. The estimate lies entirely within ±1*SE boundary only from patient 660 onwards (61% through the trial). Whilst also starting outside any investigated boundary, the reverse order estimate is quicker to reach each boundary compared to the original order estimate. The reverse estimate reaches the inner ±1*SE boundaries at 540 patients and remains there from that point forward.

5.4 Overview of trials

This section provides an overall summary of all the trials re-analysed. Table 5.5 provides the median, lower and upper quartiles of conditional power at three interim time points under four future treatment effects. As expected, the median CP is higher for significant trials than non-significant trials. However, current trend assumption has smaller values at the 25% time point (median=73.1%, LQ=9.3%). Had an interim time point been carried out at 25% data available, sample size would have increased in some cases, or even stopped for futility (having dropped below the 10% limit), despite having a significant finding with the original n patients. For non-significant trials however, the current trend appears to have a very low median CP at all investigated time points. The hypothesised effect however seems slow to drop CP, with a median CP at the 50% time point being 63.8%. Arguably in this situation, a SSR may have been effective in this scenario, and trials may have had a significant result had sample size been increased. However, taking the original treatment effect as the assumed true effect, this would have lead to additional patients being recruited, when there is no true treatment effect in the population. Optimistic limits appear to have high CP at all time points for significant trials, as well as low CP values for non-significant trials from 50%

onwards. Table 5.6 shows n^* as a percentage of original n for a maximum increase of 1.5*n

Conditional Power		**Not significant (N=6)**	**Significant (N=15)**	**All (N=21)**
Median (LQ,UQ)				
Current trend	25%	0.4 (0.1, 5.0)	73.1 (9.3, 100.0)	45.4 (1.4, 98.0)
	50%	1.4 (0.0, 12.8)	99.9 (60.2, 100.0)	84.5 (14.2, 100.0)
	75%	5.9 (0.0, 20.8)	100.0 (95.3, 100.0)	99.0 (22.4, 100.0)
Hyp.	25%	98.2 (84.0, 99.5)	99.7 (95.4, 100.0)	99.5 (95.4, 100.0)
	50%	63.8 (21.1, 95.6)	100.0 (96.5, 100.0)	99.0 (77.2, 100.0)
	75%	21.1 (0.0, 87.7)	100.0 (99.9, 100.0)	100.0 (87.7, 100.0)
80% limit	25%	32.5 (19.5, 71.8)	100.0 (81.2, 100.0)	98.2 (61.0, 100.0)
	50%	14.8 (0.7, 55.8)	100.0 (98.9, 100.0)	100.0 (58.3, 100.0)
	75%	16.4 (0.0, 46.9)	100.0 (99.9, 100.0)	100.0 (49.3, 100.0)
90% limit	25%	56.5 (40.9, 88.6)	100.0 (93.5, 100.0)	99.7 (81.8, 100.0)
	50%	23.5 (1.7, 69.5)	100.0 (99.6, 100.0)	100.0 (71.7, 100.0)
	75%	20.4 (0.0, 55.2)	100.0 (99.9, 100.0)	100.0 (57.6, 100.0)

Table 5.5: Summary of median, lower quartile and upper quartile conditional power values from all trials at three interim time points for four treatment effect assumptions, for statistically significant and non significant trials

(corresponding to 150% in the table). A value of 100% indicates no sample size increase, and n=n^*. For non-significant trials under the current trend assumption, all median n^* values using the combination test design are lower than the original n, with the lower quartile at 25% time point as low as 26.4% of the original trial size. At the 75% time point, the median is 94%, and even see some increases in sample size (upper quartile 131.9%), which is also the only increase seen across all time points and all designs using the trend assumption. The maximum value of 150% can be seen in the upper quartile of all combination test designs for non-significant trials at all time points using either optimistic limit, with promising zone only seeing this level of increase using the 90% limit. As expected, higher increases are indicated for the non-significant trials, where all but two upper quartiles are larger than 100% - promising zone at 25% with the trend assumption (118%), and a very small increase (102%) using the combination test at 25% with the 80% limit assumption. Additionally, some decreases can be seen in the lower quartile values for the combination test (with lower quartile values of 49% for trend, 67% with hypothesised, and 62/52% with the 80/90% limits respectively, which all happen at the earliest time point, 25%). It should be noted however that a different result may or may not have been found had only half the patients been recruited to the trial. However, this design could be beneficial if the same result can be found with only half the patients.

$(n^*/n)*100$			**Not significant (N=6)**	**Significant (N=15)**	**All (N=21)**
			Median (LQ, UQ)	**Median (LQ, UQ)**	**Median (LQ, UQ)**
n_{max}**=1.5***n					
Trend	25%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 118.4)	100.0 (100.0, 100.0)
		CT	44.9 (26.4, 100.0)	83.3 (49.4, 100.0)	62.5 (46.8, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	59.1 (52.2, 69.8)	100.0 (71.8, 100.0)	90.0 (63.4, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	94.0 (78.1, 131.9)	100.0 (96.6, 100.0)	100.0 (95.2, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Hyp.	25%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	90.3 (72.9, 142.3)	100.0 (67.2, 100.0)	100.0 (67.3, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	PZ	100.0 (100.0, 107.4)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	103.1 (85.4, 150.0)	100.0 (76.5, 100.0)	100.0 (85.4, 100.7)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	101.8 (90.2, 112.1)	100.0 (96.6, 100.0)	100.0 (96.6, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
80% limit	25%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	65.9 (27.9, 150.0)	100.0 (62.4, 102.3)	100.0 (43.6, 102.3)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	PZ	100.0 (100.0, 142.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	113.9 (63.4, 150.0)	100.0 (90.0, 100.0)	100.0 (77.9, 100.0)
		SW	100.0 (100.0, 133.3)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	PZ	100.0 (100.0, 142.4)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	94.0 (83.1, 150.0)	100.0 (96.5, 100.0)	99.0 (95.2, 100.0)
		SW	100.0 (100.0, 116.7)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
90% limit	25%	PZ	100.0 (100.0, 150.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	114.1 (66.9, 150.0)	100.0 (52.0, 100.0)	100.0 (66.9, 109.2)
		SW	100.0 (100.0, 117.4)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	PZ	100.0 (100.0, 115.6)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	113.1 (63.4, 150.0)	100.0 (85.6, 100.0)	100.0 (85.1, 100.0)
		SW	100.0 (100.0, 116.7)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	PZ	112.7 (100.0, 150.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	100.8 (90.2, 150.0)	100.0 (96.5, 100.0)	100.0 (96.5, 100.0)
		SW	100.0 (100.0, 133.3)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)

Table 5.6: Percentage of n required for the new sample size at three interim time points, implementing three SSR designs using four future treatment effect assumptions. 100% indicates no change in sample size, greater than 100% indicates an increase, up to 150% of the original sample size

Table 5.7 shows the percentage of times the CP value under four assumptions would fall in each of the four zones used in the promising zone design (or three if not incorporating futility, as this line would be added to the unfavourable zone). For non-significant trials, the largest median percentage in futility (or futility and unfavourable combined), is seen using the current trend. Here, the median percentage in the favourable zone was 0.3%, with an upper quartile of 5%. Both optimistic limits also see the largest percentage in futility/(futility + unfavourable) zones for non-significant trials, whilst largely staying in the favourable zone for significant trials (median=98.3% and 98.4% respectively). The hypothesised effect sees similar median percentages to the optimistic limits, but has a slightly larger percentage in the favourable zone for non-significant trials. The current trend assumption has a median of 11.9% in the promising zone, indicating an increase in sample size, despite having a significant result with the original *n* patients.

Promising zone		**Not significant (N=6)**	**Significant (N=15)**	**All (N=21)**
n_{max}**=1.5*n**		Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)
Current trend	Futility	74.9 (40, 100)	4.0 (1, 21)	15.6 (2, 38)
	Unfavourable	9.9 (0, 55)	4.0 (0, 18)	4.0 (0, 20)
	Promising	2.3 (0, 5)	11.9 (0, 26)	9.6 (0, 20)
	Favourable	0.3 (0, 5)	70.6 (33, 97)	55.2 (10, 82)
Hypothesised	Futility	26.3 (9, 47)	0.0 (0, 0)	0.0 (0, 4)
	Unfavourable	6.7 (4, 10)	0.0 (0, 0)	0.0 (0, 1)
	Promising	17.9 (12, 30)	0.0 (0, 26)	7.9 (0, 26)
	Favourable	36.7 (22, 77)	100.0 (74, 100)	91.7 (54, 100)
80% limit	Futility	40.0 (12, 79)	0.0 (0, 2)	0.2 (0, 7)
	Unfavourable	11.3 (7, 30)	0.0 (0, 1)	0.4 (0, 5)
	Promising	22.4 (5, 50)	0.1 (0, 16)	4.0 (0, 20)
	Favourable	13.8 (3, 22)	98.3 (84, 100)	89.2 (32, 100)
90% limit	Futility	37.3 (12, 69)	0.0 (0, 2)	0.0 (0, 5)
	Unfavourable	12.3 (7, 16)	0.0 (0, 1)	0.1 (0, 5)
	Promising	28.3 (11, 44)	0.1 (0, 4)	2.1 (0, 19)
	Favourable	16.8 (7, 35)	98.4 (91, 100)	91.5 (44, 100)

Table 5.7: Percentage of trial duration spent in each zone at three interim time points and four treatment effect assumptions for the promising zone design

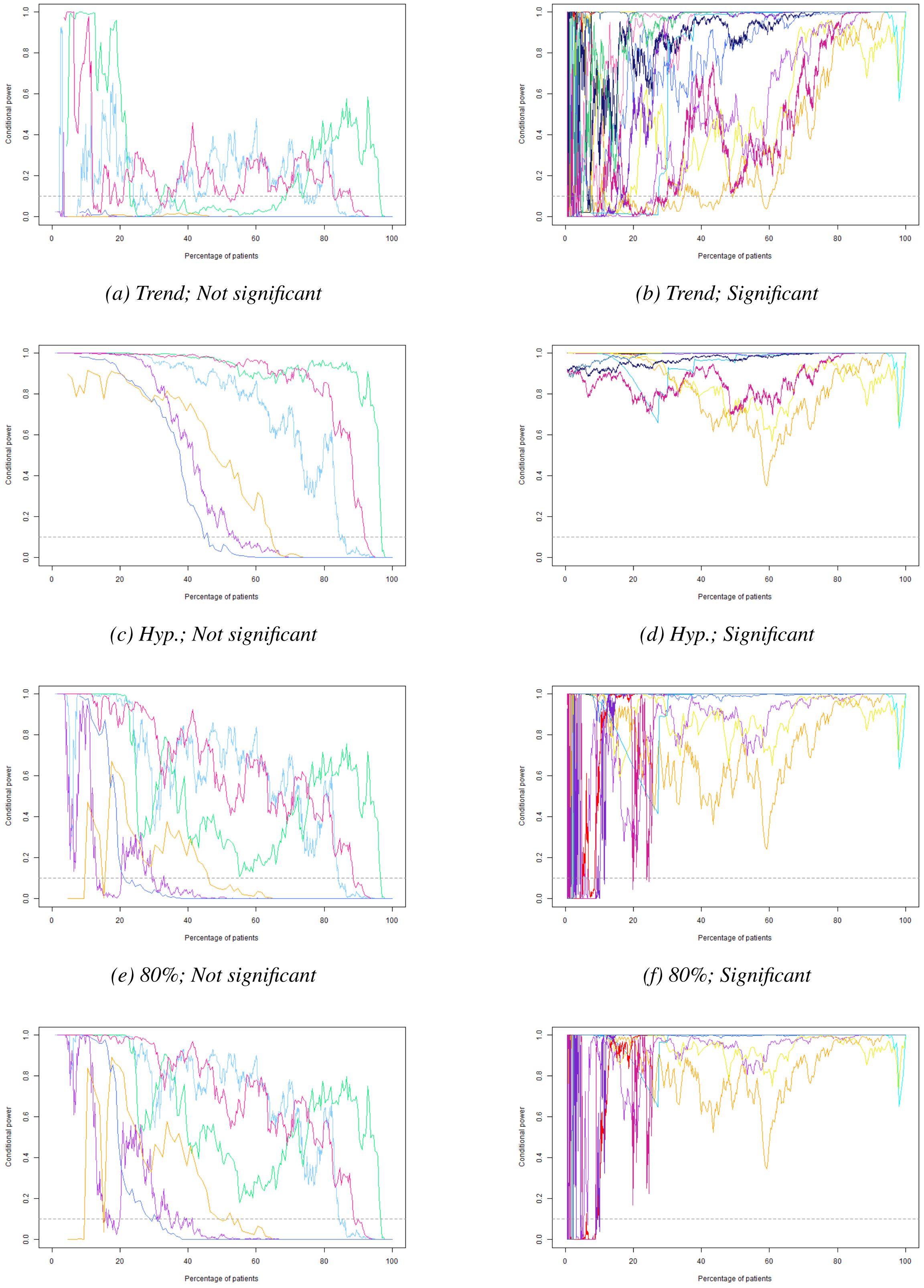
Similarly to Table 5.7, Table 5.8 shows the percentage in each sample size state (i.e. decrease, remain the same or increase in sample size) for the combination test design. The trend assumption largely stops the non-significant trials early, which could be similar to a futility rule. However, the trend assumption would also decrease sample size on average 14.8% of the time for the significant trials, which may not have necessarily resulted in a significant end result. The hypothesised assumption behaves in a similar manner, but with less decreases in sample size for non-significant trials. The confidence limit assumptions are very similar to the hypothesised assumption, but have slightly larger proportion of time spent decreasing sample size in non-significant cases and slightly less decreases in significant trials in the 80% assumption.

Combination test		**Not significant (N=6)**	**Significant (N=15)**	**All (N=21)**
n_{max}**=1.5*n**		Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)
Current trend	Decreased	87.8 (67, 98)	14.8 (0, 81)	37.4 (0, 83)
	Remained	8.1 (4, 21)	66.2 (9, 100)	20.7 (5, 100)
	Increased	0.0 (0, 28)	0.0 (0, 7)	0.0 (0, 7)
Hypothesised	Decreased	44.0 (14, 78)	26.1 (0, 81)	33.3 (0, 78)
	Remained	7.6 (4, 21)	66.2 (9, 100)	20.7 (4, 100)
	Increased	52.0 (18, 56)	0.0 (0, 5)	0.0 (0, 51)
80% limit	Decreased	56.6 (19, 96)	17.0 (0, 79)	30.5 (0, 81)
	Remained	5.9 (4, 12)	66.2 (9, 100)	19.3 (6, 100)
	Increased	32.1 (4, 76)	0.0 (0, 20)	0.0 (0, 26)
90% limit	Decreased	52.0 (26, 89)	26.1 (0, 81)	29.0 (0, 81)
	Remained	6.3 (4, 12)	66.2 (9, 100)	19.3 (6, 100)
	Increased	38.4 (10, 66)	0.0 (0, 12)	2.9 (0, 23)

Table 5.8: Percentage of trial duration spent in sample size state at three interim time points and four treatment effect assumptions for the combination test design

Table 5.9 shows the same percentage in each sample size state split between significant or not trials, with states shown as step values (i.e. r_0 implies no increase, r_3 implies the greatest increase allowed). Similarly, almost no increases take place in significant trials, with very slightly higher amount of time increasing under the current trend. Some increases occur for non-significant trials, with the most amount of time increasing at any step value occurring in the 90% limit assumption, closely followed by th 80% assumption.

The current trend has the lowest CP of the assumptions, for both significant and non-significant trials, and is slow to reach the favourable zone, or close to one for significant trials. Whilst the IMPROVE and SELF trials may have benefited from a futility boundary at any time point due to almost zero CP under the current trend, almost all significant trials would have stopped early with any interim analysis prior to around 30% through the trial. Even at $\approx$ 60% through the trial, AMAZE (orange) would have stopped for futility had an interim analysis been done at that point, despite finding a significant result at n patients. The graphs suggest that the current trend may not be a useful assumption to use if wanting to incorporate a futility boundary. The hypothesised effect starts high and is slow to decrease for non-significant trials, with only one reaching CP<0.1 (or 10%) by the halfway mark,



(a) Trend; Not significant

(b) Trend; Significant

(c) Hyp.; Not significant

(d) Hyp.; Significant

(e) 80%; Not significant

(f) 80%; Significant

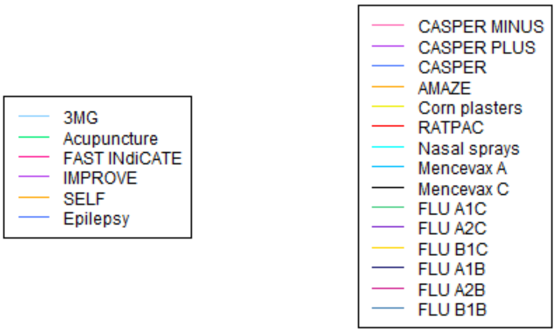
(g) 90%; Not significant

(h) 90%; Significant

Figure 5.10: Conditional power values during the study progression for all trials for four treatment effect assumptions, split by not significant (left) and significant (right)

Stepwise		**Not significant (N=6)**	**Significant (N=15)**	**All (N=21)**
n_{max}**=1.5*n**		Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)
Current trend	r_0	98.6 (97, 100)	93.1 (81, 100)	96.8 (87, 100)
	r_1	0.6 (0, 2)	2.8 (0, 6)	1.7 (0, 5)
	r_2	0.3 (0, 1)	2.1 (0, 5)	1.4 (0, 4)
	r_3	0.3 (0, 1)	0.9 (0, 4)	0.7 (0, 3)
Hypothesised	r_0	89.9 (79, 92)	100.0 (93, 100)	98.5 (89, 100)
	r_1	3.5 (2, 6)	0.0 (0, 4)	0.4 (0, 4)
	r_2	2.5 (2, 4)	0.0 (0, 2)	0.4 (0, 3)
	r_3	4.1 (4, 4)	0.0 (0, 1)	0.1 (0, 4)
80% limit	r_0	81.5 (59, 96)	99.8 (89, 100)	97.5 (84, 100)
	r_1	5.2 (1, 15)	0.1 (0, 4)	0.8 (0, 5)
	r_2	5.8 (1, 14)	0.1 (0, 3)	0.9 (0, 5)
	r_3	7.6 (1, 13)	0.1 (0, 3)	0.7 (0, 4)
90% limit	r_0	75.9 (62, 94)	99.9 (98, 100)	98.4 (93, 100)
	r_1	9.3 (3, 12)	0.0 (0, 1)	0.9 (0, 3)
	r_2	8.6 (2, 15)	0.0 (0, 1)	0.4 (0, 3)
	r_3	6.2 (1, 11)	0.0 (0, 1)	0.6 (0, 2)

Table 5.9: Percentage of trial duration spent in step at three interim time points and four treatment effect assumptions for the stepwise design



(a) Non-significant trials *(b) Significant trials*

Figure 5.11: Legend for Figure 5.10

and three of the six trials above this boundary until at least 85% through. Incorporation of a futility bound would be unlikely to “save” any additional patients as they will likely have been recruited by this point in the trial. Therefore a futility boundary may not be useful in this scenario. The hypothesised effect however, is the highest of all CP values for the significant trials, with most trials having very high CP, with a few exceptions, but none reach the futility boundary, and only one trial gets close to the unfavourable zone (if using), depending on the logistical factors that influence the CP_{min} value.

The 80 and 90% optimistic confidence limit assumptions see very large fluctuations of CP at the very start of the trial, particularly in the significant trials. This rapidly changing CP largely stops by 15% through the trial, with the exception of one trial (Flu vaccine strain A2, both continuous and binary endpoints). After the 25% time point, no significant trial would have been stopped early for futility, had a boundary been included. By 30%, most trials have very high CP values, with AMAZE again having a low peak, but not as severe as under the current trend. Additionally, the nasal spray trial has a low peak around 95%-100% through the trial. CP is still somewhat slow to reach low values for the non-significant values, but Epilepsy and IMPROVE fall relatively quickly, with SELF following shortly behind. The remaining three trials fall largely in the promising zone, and may or may not have benefited from a SSR and additional patients.

One limitation however, is that there are only six non-significant trials and fifteen significant trials. With such a small sample size, it is difficult to make convincing decisions based on the results so far. It would be more informative to see how each trials fairs, when observing exactly the planned effect, a smaller than planned effect, no effect at all, or even an opposite effect to that predicted (i.e. control/placebo group is performing better than the intervention group). Therefore, a method to transform the data observed in the trial to match each of these discussed scenarios will be presented in the next chapter, to more comprehensively investigate CP assumptions and their impact on interim decision making.

Table 5.10 gives details of when the treatment estimate first enters, and remains within, 7 boundaries of ±0.25 - 4 standard errors of the end treatment effect of the original trial. As expected, the first instance entering each boundary is much earlier than the ‘remains within’ time frame. For the original order, the median time to remaining within ±1*SE is not until 57% of the way through the trial (LQ=25%, UQ=64%), and even later for smaller boundaries (68%, 78% and 92% for 0.75, 0.5 and 0.25 respectively). The median time to remaining within 3*SE is 11%, and 21% for 2*SE. This means an early interim analysis may be based on a less stable estimate, and decisions based on this value may lead to variable results (i.e. between 24% and 26% of data available, decisions may be very different). This ties in with the plots of n^* of each trial, having most of the fluctuation in sample sizes occurring within the first half of the trial. The reverse order yields similar findings, indicating that patients are not necessarily inherently different whether they are recruited near the start or the end, but that the estimate is just less stable with fewer patients.

Percentage		Short (N=7)	Medium (N=8)	Long (N=6)	All (N=21)
	Bound	**Median (LQ, UQ)**	**Median (LQ, UQ)**	**Median (LQ, UQ)**	**Median (LQ, UQ)**
Original order					
First enters boundary	±0.25*SE	21.0 (8, 25)	38.0 (33, 86)	27.0 (12, 34)	31.0 (21, 38)
	±0.5*SE	8.0 (7, 25)	35.5 (30, 85)	25.5 (12, 34)	30.0 (12, 37)
	±0.75*SE	8.0 (4, 25)	25.5 (11, 32)	25.5 (12, 34)	20.0 (7, 33)
	±1*SE	8.0 (4, 25)	23.5 (11, 29)	14.5 (12, 21)	17.0 (7, 25)
	±2*SE	4.0 (4, 7)	18.0 (2, 25)	7.5 (4, 11)	7.0 (4, 16)
	±3*SE	4.0 (2, 7)	13.0 (2, 24)	7.5 (4, 11)	7.0 (4, 11)
	±4*SE	4.0 (1, 7)	13.0 (2, 23)	7.5 (4, 11)	4.0 (3, 11)
Remains within boundary	±0.25*SE	90.0 (88, 98)	94.0 (85, 99)	94.0 (89, 96)	92.0 (89, 97)
	±0.5*SE	83.0 (72, 88)	79.0 (74, 89)	77.5 (67, 78)	78.0 (73, 84)
	±0.75*SE	52.0 (28, 71)	76.0 (55, 84)	66.0 (63, 75)	63.0 (47, 76)
	±1*SE	25.0 (16, 61)	63.0 (47, 78)	52.0 (46, 61)	57.0 (25, 64)
	±2*SE	16.0 (13, 21)	30.0 (25, 46)	14.0 (9, 23)	21.0 (14, 29)
	±3*SE	7.0 (7, 14)	24.0 (17, 29)	11.0 (9, 11)	11.0 (9, 21)
	±4*SE	7.0 (4, 14)	15.5 (10, 23)	11.0 (8, 11)	11.0 (7, 16)
Reverse order:					
First enters boundary	±0.25*SE	14.0 (7, 18)	52.5 (28, 89)	40.0 (25, 65)	26.0 (14, 65)
	±0.5*SE	7.0 (7, 11)	47.0 (22, 68)	28.5 (24, 57)	25.0 (7, 54)
	±0.75*SE	7.0 (5, 11)	28.5 (13, 45)	18.0 (5, 57)	15.0 (7, 35)
	±1*SE	7.0 (5, 7)	19.5 (7, 35)	17.0 (5, 21)	9.0 (5, 21)
	±2*SE	7.0 (4, 7)	19.5 (6, 29)	13.5 (4, 17)	7.0 (4, 17)
	±3*SE	7.0 (4, 7)	18.0 (5, 22)	8.5 (4, 11)	7.0 (4, 11)
	±4*SE	7.0 (3, 7)	11.5 (4, 20)	7.0 (4, 11)	7.0 (4, 11)
Remains within boundary	±0.25*SE	85.0 (76, 96)	96.5 (92, 99)	94.5 (88, 95)	95.0 (87, 96)
	±0.5*SE	71.0 (56, 88)	94.5 (81, 97)	86.5 (82, 92)	88.0 (71, 94)
	±0.75*SE	47.0 (39, 62)	82.0 (72, 85)	60.0 (53, 69)	63.0 (47, 81)
	±1*SE	42.0 (32, 57)	71.0 (39, 80)	51.5 (50, 67)	53.0 (40, 69)
	±2*SE	19.0 (13, 46)	28.5 (17, 38)	24.5 (17, 50)	23.0 (17, 41)
	±3*SE	7.0 (5, 11)	22.0 (11, 32)	15.5 (10, 19)	13.0 (7, 21)
	±4*SE	7.0 (5, 7)	17.5 (10, 29)	9.5 (7, 11)	7.0 (5, 16)

Table 5.10: Table to show the first time the original order and reverse order treatment estimate enters each boundary and the last time (i.e. the estimate remains within this boundary for the remainder of the trial) for all trials, split by statistical significance in terms of percentage through the original n patients

Therefore, the later an interim analysis can be done, the more stable the estimate, and therefore the more reliable the decision of the SSR can be, regardless of the rule used. This will be further discussed in Chapter 6.

5.5 Summary

This chapter has presented the results from the data re-analysis section, through case studies demonstrating CP calculations, SSR design implementation and an investigation of the stability of the estimate.

The current trend and hypothesised assumptions have been shown to have little benefit from a futility boundary. Optimistic limits have similar sample size increases/decreases and time in zones/sample size states to the hypothesised effect at three investigated time points, yet could benefit from a futility boundary incorporation. Figure 5.10 has illustrated a faster decrease in CP than the hypothesised effect in non-significant trials, and faster CP increase than the current trend assumption in significant trials, possibly indicating a ‘middle-ground’ between the two. Stability of the estimate has also been summarised, concluding that observed estimates on average only remain within ± 1 standard error of the end treatment effect by $\approx 57\%$ through the trial on average, and an early interim analysis may base decisions on a less stable estimate. More investigation would be required to draw any firm conclusions however, from either the SSR investigation or stability of the estimate.

The next chapter will again re-analyse the data retrospectively, but will transform the treatment response of each trial to look at an observed effect size in relation to its planned effect size. This will allow a more comprehensive investigation of CP assumptions and their impact on interim decisions. CP characteristics can be directly compared between all trials, as the effect at the originally planned n patients is known (i.e. set to be exactly as planned, smaller than planned, zero or negative). CP, n^*, and zones/sample size states will again be compared when the observed effect is exactly as planned, a fraction of that planned, zero, and even negative.

6 | Investigation of planned versus observed effect sizes

6.1 Introduction

Chapter 5 presented results using data from 21 trial outcomes and the original sequential order, unchanged observed treatment effect. However, some trials observed effect was nearer the effect planned than others, leading to the variation in CP values. This chapter introduces a method of data transformation, such that trials can be directly compared in relation to their planned effect, whilst keeping the same pattern of accruing data (estimate stability graphs). CP lines using four assumptions will be compared for all trials when the observed effect is exactly as planned ($\hat{\delta}_{obs}$=δ_{plan}), a fraction of that planned ($\hat{\delta}_{obs}$=$\upsilon\delta_{plan}$) where $0 < \upsilon < 1$, zero ($\hat{\delta}_{obs}$=0), or even negative ($\hat{\delta}_{obs}$=$-\upsilon\delta_{plan}$).

This chapter starts by presenting methods to transform the data for either continuous or binary outcomes and chooses values of υ to investigate and design frameworks to carry forward based on the previous chapter. Finally, results will be reported in the form of percentage of time in sample size states, new sample size as a percentage of original sample size, and plots of conditional power lines.

6.2 Aims

This chapter aims to extend on the work of Chapter 5, again using retrospective trial data. The specific aims in this chapter is to:

1. Describe methods of data transformation such that the observed end result of each trial is comparable in terms of their originally planned effect size
2. Investigate conditional power assumptions at accruing time points when observed effect size is exactly as planned, a fraction of that planned, zero, or negative

3. Explore impacts on decision making (sample size modification/futility boundary) at interim time points for each situation of $\hat{\delta}_{plan}$
4. Inform simulation work to be carried out

6.3 Methods

In the previous chapter, trials were somewhat varying in terms of difference between observed and planned effect sizes. This was a great illustration of real-world trials, but led to very small sample sizes when trying to compare trials that were significant or not. This chapter aims to have a greater comparison between the trials, and to see what happens when the end observed effect at n patients is exactly as planned, smaller than planned by some margin, zero or even the opposite direction to that planned (negative trial), by transforming the data.

This section describes the trials to be used for this investigation, design frameworks to carry forward, and methods for transforming both continuous and binary endpoints to the desired observed effect at n patients.

6.3.1 Funding type

This section of the research has arisen as a direct result of the work done in Chapter 5, and was therefore not on the original plan of investigation when applying for data. Due to the strict nature of the data sharing agreement with industry and their lengthy amendment procedure to the initial contract, it was decided that the publicly funded trials would give enough case studies to sufficiently demonstrate the aims of this chapter. Therefore, 11 outcomes from 10 original trials are re-analysed in this chapter.

6.3.2 Study design

As mentioned in Section 3.18, the stepwise design was to be used as an illustration of how the framework compares to other designs, and the main comparison of the thesis is between promising zone and the combination test designs. The decision to use stepwise design largely stems from the concern of back-calculation of treatment effect at the interim analysis, but loses power in doing so. Where the back-calculation is not an issue, such as revealing only the maximum sample size and recruiting until told otherwise, this design would not offer any advantage over the other two. Therefore, from this point forward, only the two designs are to be compared.

6.3.3 Effect size values for investigation

Values of $\hat{\delta}_{obs}$ will be based in relative terms of δ_{plan}. Choices include:

$$\hat{\delta}_{obs} = \left\{1, \frac{2}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, 0, -\frac{1}{2}\right\} * \delta_{plan}$$

where a negative value indicates the control/placebo arm is better than the chosen intervention arm. Results for $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{4} * \delta_{plan}$ can be found in Appendix D.

6.3.4 Methods for adjusting the observed effect

This section describes the data transformation methods for continuous and binary endpoints, to obtain the chosen values of $\hat{\delta}_{obs}$ to be compared.

For continuous data, all observations were multiplied by some constant, chosen for each trial to obtain the planned SD for an analysis with the original n patients. Once the SD matched that planned, another constant value chosen for each trial was added to the intervention arm only. This ensured that both the SD and treatment difference were exactly as planned in the original trial. Different constants were added to see the effect of a smaller treatment difference, in terms of a fraction of the planned treatment effect, but standard deviation was kept as that planned in the original trial.

A similar method was implemented for binary data, but used the model coefficient, or $log(OR)$ calculated after every patient. All individual coefficients were multiplied by some constant such that the standard error at the end of the trial matched that planned. Another constant was then added to all transformed coefficients, to ensure that the $log(OR)$ was the same of the planned $log(OR)$. Again, different constants were added to observe a fraction of the planned effect size, but standard error was kept the same as that planned in the original trial.

6.4 Criteria for evaluating methodologies

In order to assess the chosen methodologies and their performance, criteria relating to operating characteristics such as required sample size and power are specified. Criteria have been chosen and have been informed by the current literature (Chapter 3) and have been refined following the work of the previous chapter (Chapter 5). Criteria have been split by observed treatment effect at the end of the trial as desired characteristics will be different across possible observed effect scenarios. In my opinion, these are the criteria that make the most logical sense, both statistically and logistically speaking.

As planned:
Does not increase sample size If sample size is increased, the increase is minimal Does not stop early for futility (if applicable) Does not excessively increase power beyond the pre-specified level Does not reduce power below the pre-specified level
Smaller than planned, but potentially clinically relevant:
Increases in sample size that also correspond to a higher probability of significance Does not stop for futility (if applicable)
Very small or no effect:
Does not increase sample size If sample size is increased, the increase is minimal Stops early for futility (if applicable) Does not inflate Type I error

Table 6.1: Criteria for the evaluation of the three methodology frameworks explored in terms of required sample size and power

The perfect method would be able to increase sample size only where necessary (when smaller than planned but not too small). It would have a high stopping for futility rate (if applicable) when there is no treatment effect, a very small effect, or even a negative effect. When the treatment effect is exactly as planned, the power would not fall below the prespecified level, but would also not excessively increase power.

In reality however, no method is perfect, and the closest method to the desired characteristics in each scenario will be the method recommended at the end of the thesis. Criteria starting with "Does not ..." will not be regarded as absolute, and a "low number" of instances will still be considered as acceptable for these situations. In these cases, the fewest number of instances will be the most favourable. More than one method may be recommended should a similar level of criteria be met.

6.5 Results

6.5.1 Observed effect = planned

Similarly to Chapter 5, Tables 6.2 and 6.3 show the median new sample size relative to the original n patients, and the proportion of times CP fall in each zone (corresponding to sample size decrease, increase or remain the same n) respectively. When $\hat{\delta}_{obs}$ is exactly as planned for the original n patients, promising zone is largely in the favourable zone, implying no sample size increase. Under the current trend assumption however, CP falls into the promising zone a median of 2.2% of the time (LQ,UQ)=(0,8), and even falls into the unfavourable zone 0.5% of the time (LQ,UQ)=(0,1). However this does not correspond to a change in median sample size increase, with promising zone yielding LQ and UQ's of 100% at all three interim analyses.

For binary data however, CP is much less stable (Figure 6.2) for all assumptions, but particularly current trend. Whilst the hypothesised, 80% and 90% limits do not seem to fall into the unfavourable or futility zones, there is between 5.6 and 16.4% (on average) falling into the promising zone, indicating a sample size increase. At the three specific interim analyses, this only results in very small sample size increase (UQ=100.2% for hypothesised, 100.5% for 80% limit and 100.0% for 90% assumptions). Current trend however only spends around 1/4 of its time in the favourable zone, and even sees a median of 14.1% time in the futility

$(n^*/n) * 100$		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising Zone	25%	100.0 (100.0, 100.0)	131.7 (100.0, 150.0)	100.0 (100.0, 131.7)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 107.9)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination Test	25%	77.5 (46.9, 100.0)	43.6 (34.9, 49.4)	49.4 (35.6, 100.0)
	50%	73.1 (54.9, 100.0)	71.8 (71.6, 83.5)	71.8 (61.4, 100.0)
	75%	98.3 (83.3, 100.0)	95.2 (91.8, 96.5)	96.5 (83.3, 100.0)
Hyp.				
Promising Zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.2)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination Test	25%	66.7 (45.8, 100.0)	60.2 (48.6, 63.9)	60.2 (45.8, 89.9)
	50%	73.1 (54.9, 100.0)	85.8 (71.6, 99.5)	76.5 (61.1, 100.0)
	75%	98.3 (83.3, 100.0)	95.2 (91.4, 96.5)	96.5 (83.3, 100.0)
80%				
Promising Zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.5)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination Test	25%	64.7 (37.2, 83.3)	52.5 (51.0, 68.5)	52.5 (43.6, 83.3)
	50%	73.1 (54.9, 100.0)	77.7 (71.6, 85.8)	76.5 (59.9, 100.0)
	75%	98.3 (83.3, 100.0)	95.2 (90.0, 96.5)	96.5 (83.3, 99.7)
90%				
Promising Zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination Test	25%	58.3 (35.8, 83.3)	47.7 (45.5, 61.3)	47.7 (43.6, 83.3)
	50%	73.1 (54.9, 100.0)	71.6 (71.3, 81.7)	71.6 (59.1, 100.0)
	75%	98.3 (83.3, 100.0)	95.2 (88.9, 96.5)	96.5 (83.3, 99.0)

Table 6.2: New total sample size (n^*) as a percentage of original planned sample size (n) comparison at three interim time points and four treatment effect assumptions when the observed treatment effect equals the planned effect

zone. At the 25% interim analysis, the median increase in sample size is 131.7%, with the UQ reaching the maximum allowed sample size of 150%. Even at the 50% interim analysis, the UQ is 107.9%, despite seeing the planned effect size at the original n patients.

Time spent in zone (%)		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	Futility	0.0 (0, 1)	14.1 (9, 22)	5.0 (0, 14)
	Unfav.	0.5 (0, 1)	12.8 (9, 15)	1.3 (0, 13)
	Prom.	2.2 (0, 8)	19.4 (12, 37)	8.4 (1, 19)
	Fav.	96.8 (95, 100)	27.7 (24, 45)	78.8 (28, 98)
Combination test	Decrease	67.3 (33, 82)	80.1 (66, 95)	73.5 (33, 92)
	Remain	23.3 (17, 65)	9.1 (4, 18)	17.6 (5, 24)
	Increase	0.0 (0, 0)	0.0 (0, 22)	0.0 (0, 22)
Hyp.				
Promising zone	Futility	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)
	Unfav.	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)
	Prom.	0.0 (0, 0)	12.0 (0, 17)	0.0 (0, 12)
	Fav.	100.0 (100, 100)	80.5 (77, 100)	100.0 (81, 100)
Combination test	Decrease	69.2 (33, 82)	80.1 (58, 95)	73.5 (38, 92)
	Remain	24.5 (17, 65)	9.4 (4, 18)	17.6 (5, 25)
	Increase	0.0 (0, 0)	0.0 (0, 19)	0.0 (0, 17)
80%				
Promising zone	Futility	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)
	Unfav.	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)
	Prom.	0.0 (0, 0)	16.4 (1, 25)	1.0 (0, 16)
	Fav.	99.8 (100, 100)	71.7 (68, 99)	99.0 (72, 100)
Combination test	Decrease	77.4 (33, 84)	80.1 (75, 94)	80.1 (33, 92)
	Remain	23.3 (17, 65)	8.8 (4, 18)	18.1 (5, 24)
	Increase	0.0 (0, 0)	0.3 (0, 2)	0.0 (0, 1)
90%				
Promising zone	Futility	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)
	Unfav.	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)
	Prom.	0.0 (0, 0)	5.6 (0, 11)	0.0 (0, 6)
	Fav.	99.8 (100, 100)	91.4 (89, 100)	99.7 (91, 100)
Combination test	Decrease	77.7 (33, 84)	80.1 (77, 95)	80.1 (40, 92)
	Remain	23.3 (17, 65)	10.0 (4, 18)	17.6 (5, 24)
	Increase	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)

Table 6.3: Percentage of trial duration spent in each zone at three interim time points and four treatment effect assumptions when the observed treatment effect is equal to the planned effect in terms

Figure 6.2 shows CP for continuous and binary trials separately. CP is almost 1 for all values of n_1 using the hypothesised effect, and all values past 10% of patients for either limit for continuous trials. Current trend is slower to reach 1 for most continuous trials, but does get there by around 40% through the trial. Binary trials however show much more variation, and four trials fall below 1 around 20-80% through the trial for hypothesised and optimistic limit assumptions. It should be noted that RATPAC (pink dashed line) actually showed a far greater effect than planned in the original trial, and also underestimated the observed effect throughout the trial (albeit still greater than planned) (Figure 5.7b). Because of this, shifting the observed data so that the end result matched that planned, moved the treatment effect for

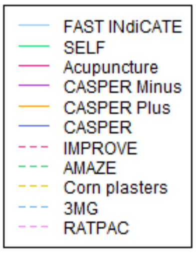


Figure 6.1: Legend for Figures 6.2 - 6.5

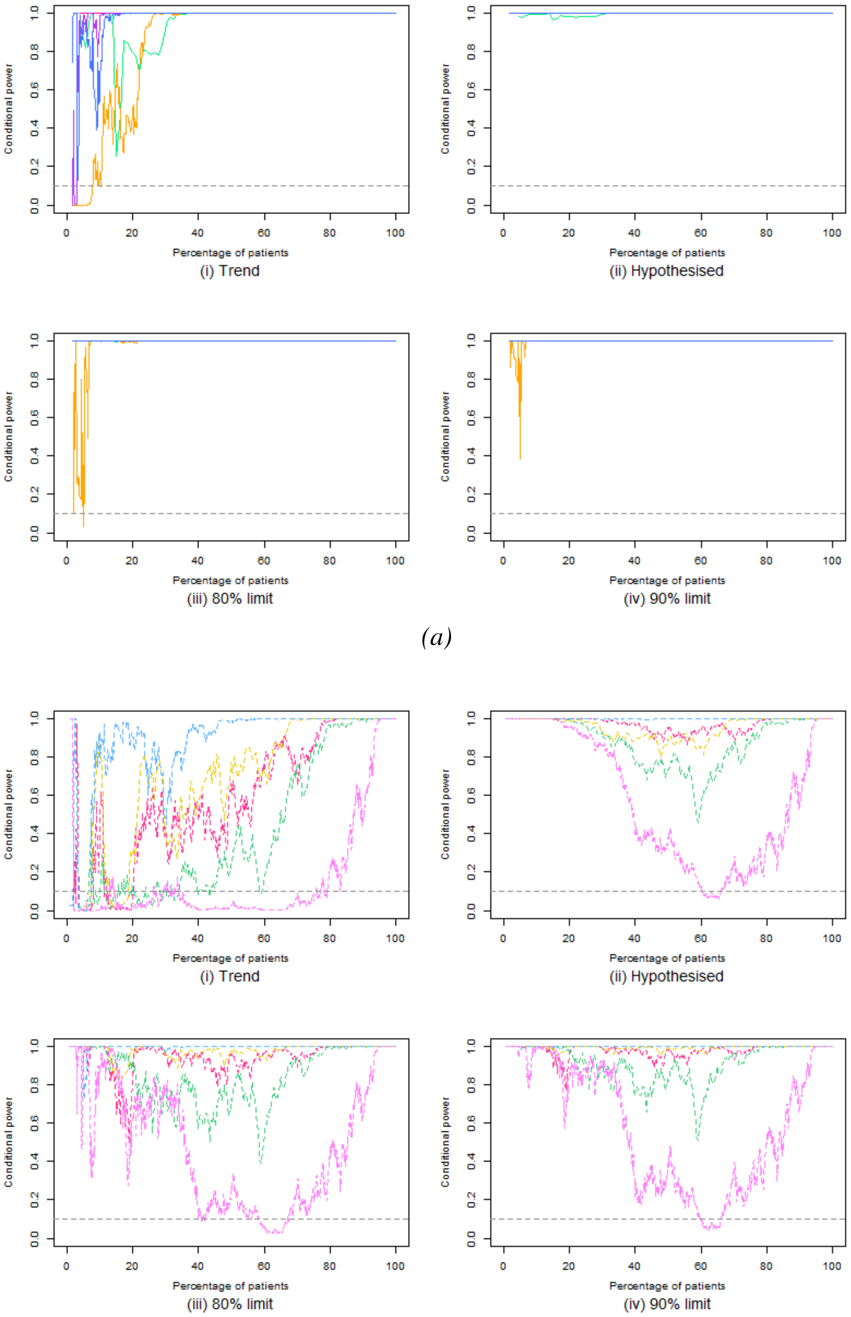
a large portion of this trial to below that planned, which is why the CP looks so low, despite seeing $\hat{\delta}_{obs} = \delta_{plan}$ at the end of the trial. The current trend assumption does not reach a CP of 1 until 45% through at the earliest for binary trials.

Combination test design mostly decreases sample size, with median sample size <100% in all cases. Again, the biggest decreases are seen earlier in the trial, which could be due to the lower number of patients recruited so far. A higher proportion of the trial is spent decreasing sample size using optimistic limits for continuous trials, and is the same across all assumptions for binary (median of 80.1% of the trial results in a decrease in sample size). No increases are seen using this design for continuous trials, but for binary outcomes the current trend assumption sees (LQ,UQ)=(0,22), hypothesised assumption sees (0,19), the 80% limit assumption sees (0,2) and the 90% limit assumption sees (0,0).

6.6 Observed effect = Two thirds planned

This section presents results when the treatment effect is not as high as was hoped, but still positive. Other proportions of δ_{plan} observed are presented in Appendix D.

Tables 6.4 and 6.5 show new sample size (n^*) and proportion in zones respectively. It should be noted that an observed effect size of two thirds of the planned size may not be clinically relevant in each trial, particularly if δ_{plan} for the original sample size has been



(i) Trend

(ii) Hypothesised

(iii) 80% limit

(iv) 90% limit

(a)

(i) Trend

(ii) Hypothesised

(iii) 80% limit

(iv) 90% limit

(b)

Figure 6.2: Conditional power when $\hat{\delta}_{obs} = \delta_{plan}$ for (a) Continuous trials (b) Binary trials

n*		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	79.1 (51.2, 145.6)	43.6 (26.4, 49.4)	51.2 (43.6, 145.6)
	50%	89.3 (69.4, 100.0)	71.6 (63.2, 71.8)	71.8 (63.2, 100.0)
	75%	99.5 (83.3, 100.0)	93.1 (82.2, 95.2)	95.2 (82.2, 100.0)
Hypothesised				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	112.4 (100.0, 138.8)	100.0 (100.0, 112.4)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 145.1)	100.0 (100.0, 100.0)
Combination test	25%	72.1 (55.6, 100.0)	67.5 (53.2, 68.3)	67.5 (53.2, 99.1)
	50%	82.4 (62.2, 100.0)	110.8 (71.6, 120.8)	88.4 (70.8, 110.8)
	75%	99.5 (83.3, 100.0)	105.4 (95.2, 139.3)	100.0 (83.6, 109.3)
80% limit				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 134.9)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	80.3 (49.0, 100.0)	65.3 (60.0, 90.9)	77.3 (49.0, 100.0)
	50%	79.5 (62.2, 100.0)	107.4 (71.6, 118.8)	82.6 (70.6, 107.4)
	75%	99.5 (83.3, 100.0)	96.5 (95.2, 138.0)	99.0 (83.5, 116.3)
90% limit				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 109.8)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	75.5 (45.1, 90.7)	56.4 (54.5, 80.3)	67.8 (45.1, 90.7)
	50%	77.2 (60.8, 100.0)	94.1 (71.6, 116.3)	77.9 (69.1, 100.0)
	75%	99.5 (83.3, 100.0)	96.5 (95.2, 139.5)	99.0 (83.3, 110.5)

Table 6.4: New total sample size (n^*) as a percentage of original planned sample size (n) comparison at three interim time points and four treatment effect assumptions when the observed treatment effect equals two thirds of the planned effect

Zones		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	Futility	2.3 (0, 5)	51.5 (14, 65)	5.3 (1, 52)
	Unfav.	3.5 (0, 4)	20.4 (11, 30)	4.0 (3, 20)
	Prom.	6.9 (5, 11)	15.1 (8, 32)	8.2 (5, 17)
	Fav.	87.3 (78, 94)	4.1 (4, 15)	64.5 (4, 90)
Combination test	Decrease	53.4 (16, 82)	80.1 (72, 84)	73.5 (16, 84)
	Remain	20.5 (16, 65)	2.0 (1, 2)	15.6 (2, 24)
	Increase	3.0 (0, 7)	12.1 (1, 22)	6.1 (0, 22)
Hypothesised				
Promising zone	Futility	0.0 (0, 0)	0.0 (0, 1)	0.0 (0, 0)
	Unfav.	0.0 (0, 0)	4.0 (2, 18)	0.0 (0, 4)
	Prom.	0.0 (0, 0)	42.5 (5, 43)	0.0 (0, 43)
	Fav.	100.0 (100, 100)	38.2 (36, 55)	99.7 (38, 100)
Combination test	Decrease	59.5 (33, 82)	43.4 (29, 80)	45.5 (29, 82)
	Remain	20.6 (17, 65)	1.7 (1, 2)	16.5 (2, 24)
	Increase	0.0 (0, 0)	53.3 (2, 66)	0.0 (0, 53)
80% limit				
Promising zone	Futility	0.0 (0, 0)	0.5 (0, 2)	0.0 (0, 2)
	Unfav.	0.0 (0, 0)	23.5 (2, 25)	0.0 (0, 24)
	Prom.	0.0 (0, 0)	35.4 (7, 44)	2.9 (0, 35)
	Fav.	99.8 (100, 100)	16.5 (13, 54)	97.1 (17, 100)
Combination test	Decrease	62.7 (33, 82)	39.4 (36, 80)	58.4 (33, 82)
	Remain	20.6 (17, 65)	1.8 (1, 2)	16.5 (2, 24)
	Increase	0.0 (0, 7)	54.4 (0, 61)	0.0 (0, 54)
90% limit				
Promising zone	Futility	0.0 (0, 0)	0.5 (0, 1)	0.0 (0, 1)
	Unfav.	0.0 (0, 0)	6.6 (2, 10)	0.0 (0, 7)
	Prom.	0.0 (0, 0)	16.1 (13, 49)	0.0 (0, 16)
	Fav.	99.8 (100, 100)	24.5 (23, 85)	99.7 (25, 100)
Combination test	Decrease	71.2 (33, 82)	58.0 (33, 80)	70.1 (33, 82)
	Remain	20.6 (17, 65)	2.2 (2, 3)	16.5 (2, 24)
	Increase	0.0 (0, 1)	34.2 (0, 63)	0.0 (0, 34)

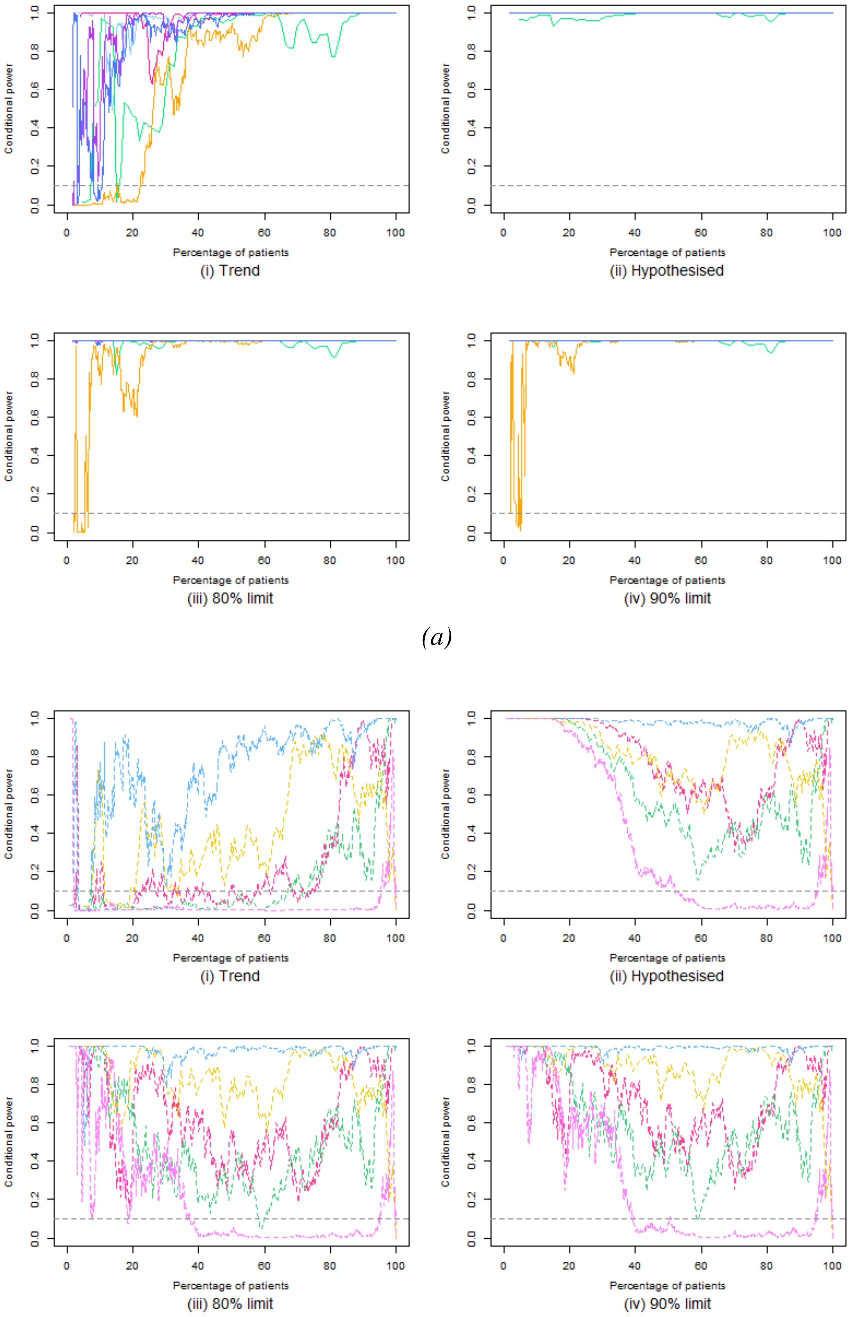
Table 6.5: Percentage of trial duration spent in each zone at three interim time points and four treatment effect assumptions when the observed treatment effect is equal to two thirds of the planned effect

chosen as the MCID. However, for the purposes of the comparisons of this thesis, it will be considered as clinically relevant, but any smaller will not be.

Continuous outcomes only see one increase in UQ (145.6% at 25% through the trial under the current trend assumption for the combination test design). Some increases can be seen for binary outcomes however for three assumptions (all except the current trend). The largest increase is seen under the hypothesised effect (UQ=145.1% at 75% through the trial using the promising zone design). For either optimistic limit, promising zone sees an increase at 25% through the trial in the upper quartile, whereas the combination test sees the largest increases later through the trial. Whilst the combination test still largely sees a decrease in sample size, now that the effect size is smaller than planned, the decrease is not as great as the previous example, particularly in continuous trials.

The median proportion in each zone is still predominantly in the favourable zone for all assumptions for continuous outcomes. The current trend assumption spends on average 6.9% of time increasing sample size using the promising zone design, with 3.5% and 2.3% in the unfavourable and futility zones respectively for continuous trials. The current trend assumption still sees the largest proportion of time in the futility zone of the four assumptions, particularly for binary trials, which spends a median time of 51.5% in the futility zone, compared to the lowest median value of 0% under the hypothesised effect.

This is further illustrated by Figure 6.3, which shows CP lines throughout the trial duration for continuous and binary endpoints. The current trend assumption fluctuates for the first half of the trial for almost all continuous trials, compared to high CP values for almost all continuous trials under the other three assumptions. Again, binary outcomes show much variation in CP throughout the trial duration. Even at 95% through the trial, there is some variation, and not all trials have reached their final decision of CP$\approx$0 or 1. This highlights the amount of noise in binary data compared to continuous, and it may not always be appropriate to carry out a SSR for binary outcomes. This will be further investigated through simulation work before any firm recommendation can be made.



(a)

(b)

Figure 6.3: Conditional power when $\hat{\delta}_{obs} = \frac{2}{3}\delta_{plan}$ for (a) Continuous trials (b) Binary trials

6.7 Observed effect = zero

Tables 6.6 and 6.7 again show sample size (n^*) and proportion in zones, but now showing when the end treatment effect at n patients is in fact 0 (i.e. no difference between treatments).

n*		Continuous (N=6)	Binary (N=5)	All (N=11)
Trend				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	65.0 (28.1, 100.0)	36.6 (26.4, 43.6)	43.6 (27.8, 83.3)
	50%	73.1 (54.9, 100.0)	55.4 (52.2, 71.6)	69.8 (54.3, 76.5)
	75%	98.3 (83.3, 100.0)	93.1 (78.1, 95.2)	95.2 (78.8, 99.0)
Hypothesised				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	106.1 (100.0, 118.6)	100.0 (100.0, 100.0)	100.0 (100.0, 118.6)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	86.5 (81.3, 100.0)	77.7 (64.9, 82.7)	82.7 (64.9, 100.0)
	50%	114.5 (88.5, 128.0)	71.8 (71.6, 96.4)	96.4 (71.6, 128.0)
	75%	150.1 (150.0, 150.1)	93.1 (78.1, 95.2)	150.0 (93.1, 150.1)
80% limit				
Promising zone	25%	104.0 (100.0, 130.8)	100.0 (100.0, 100.0)	100.0 (100.0, 124.3)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	86.4 (80.2, 100.0)	43.6 (26.4, 49.4)	80.2 (43.6, 100.0)
	50%	73.1 (54.9, 100.0)	71.6 (52.2, 71.8)	71.6 (54.3, 100.0)
	75%	98.3 (83.3, 100.0)	93.1 (78.1, 95.2)	95.2 (78.8, 100.0)
90% limit				
Promising zone	25%	100.0 (100.0, 120.9)	100.0 (100.0, 100.0)	100.0 (100.0, 120.9)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	91.8 (69.8, 106.6)	43.6 (26.4, 49.4)	69.8 (43.6, 96.3)
	50%	100.0 (76.5, 115.5)	71.6 (52.2, 71.8)	76.5 (69.8, 115.5)
	75%	98.3 (83.3, 100.0)	93.1 (78.1, 95.2)	95.2 (78.8, 100.0)

Table 6.6: New total sample size (n^*) as a percentage of original planned sample size (n) comparison at three interim time points and four treatment effect assumptions when the observed treatment effect is zero

Zones		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	Futility	98.5 (90, 100)	97.1 (94, 99)	98.1 (93, 100)
	Unfav.	0.9 (0, 5)	0.2 (0, 5)	0.9 (0, 5)
	Prom.	0.1 (0, 1)	0.0 (0, 0)	0.0 (0, 1)
	Fav.	0.3 (0, 3)	0.0 (0, 0)	0.0 (0, 1)
Combination test	Decrease	67.3 (33, 82)	88.2 (80, 95)	80.1 (61, 89)
	Remain	23.3 (17, 65)	8.8 (4, 18)	17.6 (5, 24)
	Increase	1.3 (0, 9)	0.0 (0, 0)	0.0 (0, 7)
Hypothesised				
Promising zone	Futility	24.5 (22, 34)	49.5 (39, 56)	34.2 (24, 50)
	Unfav.	10.7 (8, 14)	11.2 (10, 12)	11.1 (8, 14)
	Prom.	16.5 (11, 20)	14.9 (9, 16)	14.9 (10, 20)
	Fav.	46.7 (42, 53)	25.7 (25, 26)	37.4 (26, 49)
Combination test	Decrease	35.3 (29, 53)	79.3 (59, 80)	52.9 (29, 79)
	Remain	23.4 (15, 28)	8.3 (4, 9)	14.9 (5, 25)
	Increase	39.9 (31, 55)	15.4 (0, 29)	30.7 (15, 55)
80% limit				
Promising zone	Futility	57.1 (51, 75)	81.0 (54, 86)	61.9 (51, 86)
	Unfav.	18.6 (6, 20)	11.4 (7, 27)	17.6 (6, 23)
	Prom.	14.1 (3, 24)	5.1 (3, 8)	8.3 (3, 15)
	Fav.	6.2 (3, 15)	2.5 (2, 9)	5.3 (2, 15)
Combination test	Decrease	63.8 (33, 80)	83.2 (80, 95)	79.9 (33, 84)
	Remain	23.3 (17, 65)	8.8 (4, 15)	16.7 (5, 24)
	Increase	1.0 (0, 8)	0.0 (0, 5)	0.0 (0, 8)
90% limit				
Promising zone	Futility	47.9 (41, 67)	74.3 (42, 81)	50.3 (41, 81)
	Unfav.	18.0 (12, 22)	13.1 (7, 27)	17.0 (10, 22)
	Prom.	13.5 (3, 23)	6.3 (4, 10)	9.8 (4, 16)
	Fav.	17.0 (10, 21)	6.6 (6, 16)	15.2 (6, 21)
Combination test	Decrease	56.2 (33, 73)	80.1 (78, 95)	73.1 (33, 84)
	Remain	23.3 (17, 65)	8.8 (4, 14)	16.5 (5, 24)
	Increase	7.2 (0, 16)	0.0 (0, 10)	0.4 (0, 16)

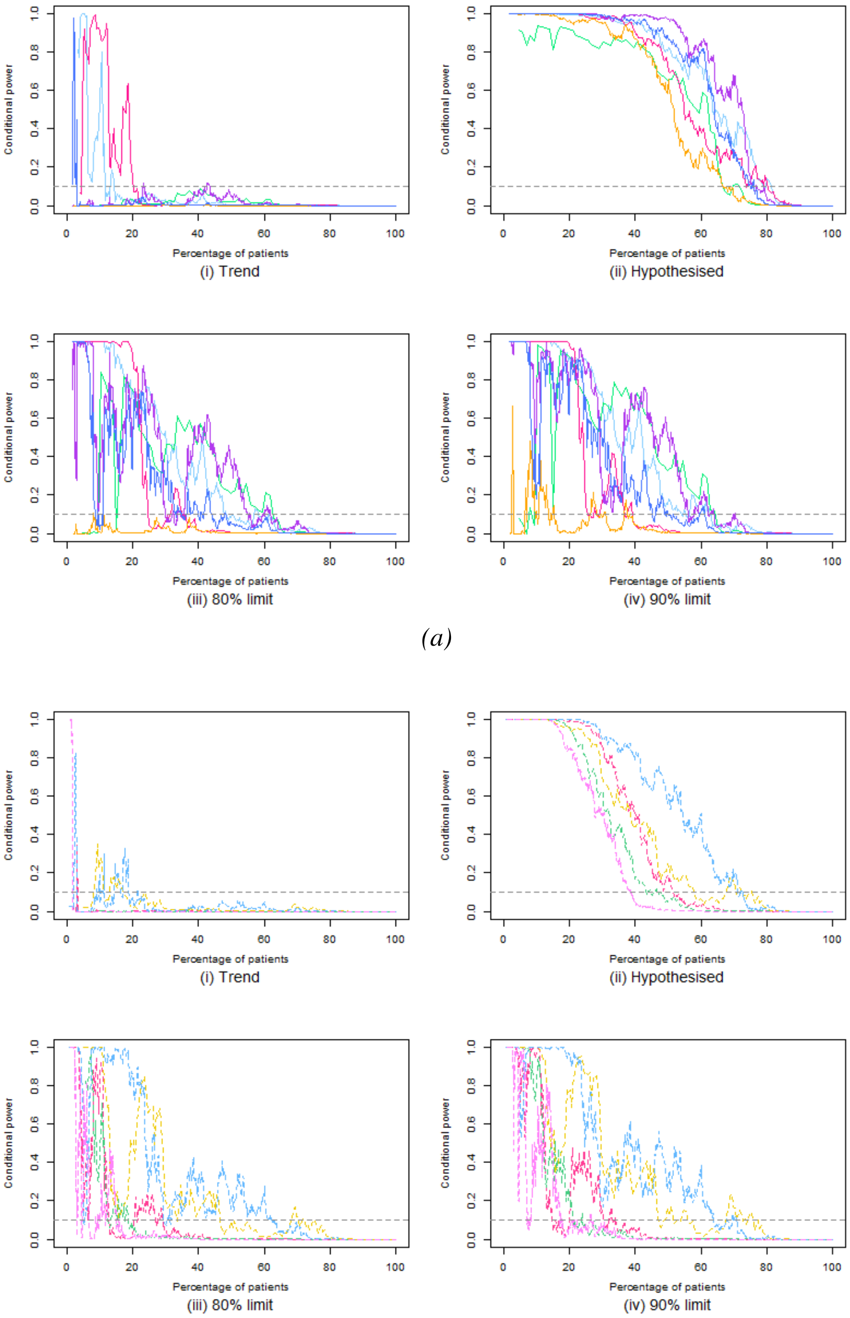
Table 6.7: Percentage of trial duration spent in each zone at three interim time points and four treatment effect assumptions when the observed treatment effect is zero

Again, the promising zone design sees some increases in sample size early on for continuous trials (UQ=130.8 and 120.9% for optimistic limits at 25% through the trial). Combination tests also see decreases again for continuous trials, but again the decreases are not as great as when the treatment effect is higher than zero. Some increases can be seen too, seeing the maximum increase allowed in the upper quartile of the hypothesised assumption at 75% interim look. Some increases are also seen using the 90% optimisitic limit. For binary, there appears no increase in sample size in the majority of studies, with the largest UQ equalling 100%. A large decrease in median sample size is seen using the current trend for binary data (median=36.6%) at 25% through the trial, corresponding to a median time of 97.1% of the trial in the futility boundary (98.5% for continuous trials). Despite no treatment difference at n patients, the hypothesised assumption still sees a median of 46.7% of the time in the favourable zone, and increases a median of 39.9% of the time with combination test design (continuous trials).

Looking at Figure 6.4, the current trend assumption is quick to fall below 10% CP, and remains there throughout. Here a futility boundary may have been able to decrease sample size had a boundary been implemented. Additionally, the hypothesised assumption takes longer to fall below this same boundary, and falls quickest for binary trials compared to continuous.

6.8 Observed effect = negative

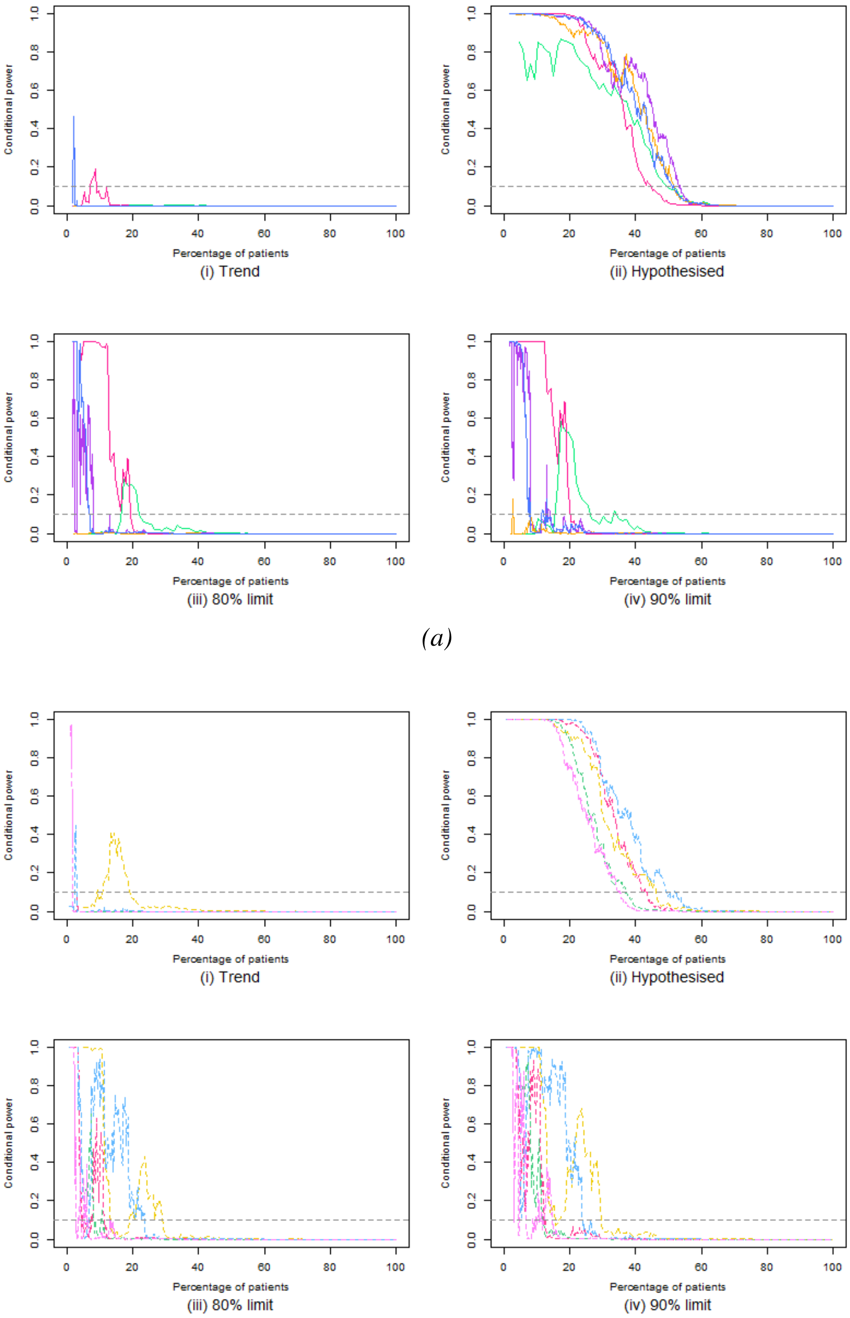
An investigation of negative treatment effect sees similar results to an observed effect size of zero, but has less fluctuation in CP (Figure 6.5). The trend assumption is again very quick to drop to zero CP for both binary and continuous endpoints. All trials have dropped below the 10% boundary by around 30% through the trial using either optimisitic limit, with the 80% falling very slightly faster. The hypothesised assumption is again much slower, dropping to zero between 40 and 60% through the trial for all endpoints.



(a)

(b)

Figure 6.4: Conditional power when $\hat{\delta}_{obs} = 0$ for (a) Continuous trials (b) Binary trials



(a)

(b)

Figure 6.5: Conditional power when $\hat{\delta}_{obs} = -\frac{1}{2}\delta_{plan}$ for (a) Continuous trials (b) Binary trials

6.9 Discussion

The transformation of the data is able to give a better understanding of how future treatment effect assumptions impact CP curves during the trial progression using real-world trial data, which we have seen is somewhat variable in the first half of the trial. Now, knowing what the trial will show at the end (whether as planned, or smaller), it can be seen where correct, or incorrect decisions would be made, and how quickly the CP reaches approximately zero or one for the remainder of the trial.

The current trend assumption was the quickest to go to zero when the observed effect was zero or negative, but was slower to get to one when the observed effect size was as planned, or smaller than planned but still positive. The hypothesised effect on the other hand was much slower to get to zero for no treatment difference or a negative effect. Additionally it was very variable throughout the trial duration for a positive effect, and was very slow to reach one for binary trials at the observed effect equal to that planned. Binary trials saw much more fluctuation in CP and this has been explained for the situation of the RATPAC trial, which saw a much larger observed effect than planned in the original trial.

The optimistic limit assumptions appear to be a good "middle ground" between the two recommended assumptions, reaching either close to zero or one at some time between the trend and hypothesised assumptions. In the negative treatment effect case, they reach zero by around 30% in the majority of trials, whether continuous or binary. Additionally, it appears as though the incorporation of a futility bound could be beneficial to this design and could help save sample size even when a promising zone design is used.

In terms of the criteria set out in Table 6.1 for evaluating the methodologies, the current trend assumption appears to be the least successful performing assumption. When the observed effect equals that planned, the promising zone design for binary outcomes sees a median increase of 32%, with an upper quartile of 50% increase in sample size, when the criteria indicates no increase should be seen. This does decrease over time however, and remains at the originally planned n patients by 75% through the trial. The 80% limit sees an upper quartile=0.5% increase at 50% through the trial in binary outcomes, which is minimal. For continuous outcomes, very few, if any, increases are seen. In terms of futility, no trial stops for futility for the hypothesised or either confidence limit assumption, but current trend stops for a median of 14.1% of trial duration in binary trials. Power is not able to be assessed through these investigations, and will be looked at in detail in the simulation work in Chapter 7.

For a smaller than planned effect ($\frac{2}{3}$ of the planned effect) that may still be clinically relevant, only the futility criteria from Table 6.1 can be assessed without simulations. The hypothesised effect has a median of 0% of trial duration in the futility zone (if using) for both binary and continuous outcomes. Both 80 and 90% limits see a median of 0.5% trial duration for binary outcomes, and 0% for continuous. The current trend assumption however sees a median of 2.3% and 51.5% of trial duration for continuous and binary outcomes respectively, indicating it is again the worst performing assumption of the four in terms of the criteria set out in this thesis.

Finally, when a zero treatment effect is observed, some sample size increases are observed for continuous outcomes for the hypothesised, 80% and 90% limit assumptions. The current trend assumption however, sees no increases in sample size, and additionally sees the highest percentage of trial duration spent in the futility boundary of the four assumptions. Therefore, when no treatment effect is observed, the current trend assumption is the best performing assumption in terms of the criteria in Table 6.1. The 80% limit has the next highest proportion of time in the futility zone (median=61.9% for any outcome).

Both 80 and 90% limits behave similarly and are comparable at any value of observed effect, and so only one of these limits will be taken forward to investigate in the simulation work. The 80% limit performs almost identically to the 90% limit assumption when $\hat{\delta}_{obs} = \delta_{plan}$ and $\hat{\delta}_{obs} = \frac{2}{3}\delta_{plan}$, but performs slightly better in terms of number of times an increase in sample size is observed, and more time in the futility zone when $\hat{\delta}_{obs}$ is zero or negative. For this reason, the 80% limit will be carried forward to the simulation work for further investigation.

Whilst there could be sufficient evidence to drop the current trend assumption for further investigation through simulations, it will be left in as it is the currently recommended assumption by Mehta and Pocock in order to show the difference of other assumptions comparatively (Mehta 2011). A futility bound could be useful in the case of the 80% limit assumption, which reaches zero by around the halfway mark when the end observed effect is zero or negative, but would stop less often when the observed effect is as planned or half that planned, than the current trend assumption currently recommended.

Understandably, the combination test results in lower sample sizes (n^*) than in the promising zone design, due to the ability to decrease, using the number recruited as the minimum sample size allowed. The later the interim analysis, the higher the sample size, as almost all patients have been recruited by this point (particularly with long outcomes). For the promising zone design however, most increases can be seen at the 25% interim analysis using current trend when the end effect is exactly as planned, showing it is too early to make a decision as it has an unnecessary sample size increase. When the observed effect is two thirds of that planned, some moderate increases are seen where a sample size increase may be appropriate, as it is still "promising" but not as high as was hoped in the planning stage. Ultimately, the biggest impact on sample size decisions is the CP assumption, and needs to be carefully considered together with the timing of the interim analysis.

This work has also showed the effect of the time to a stable estimate, particularly highlighting the underestimate of the RATPAC trial throughout the trial compared to the end result. Therefore, even when the end result was exactly as was planned in the sample size, the trial could have been stopped early for futility, had the same pattern of observed data been seen as in the original trial.

In light of these results, the simulation work will investigate promising zone and combination tests for current trend, hypothesised effect, and the 80% optimistic confidence limit. In addition, a promising zone design with the incorporation of a futility stopping boundary will be investigated more formally than in this chapter. Interim timings will only be investigated from the 50% data available point onwards, as any earlier could be making decisions on an unstable estimate. This time point has been chosen due to the work on the stability of the estimate in Chapter 5. Only a 50% increase in sample size has been considered so far, and it would be interesting to see the effect on a maximum cap in sample size would affect decisions. A small 10% maximum increase may behave very differently to a trial that is allowed to double its sample size and so the focus for the simulations will be n_{max}=1.1 and 2 times the original sample size n. Finally, only one value of γ has been evaluated for the combination test design, and simulation work will be able to evaluate multiple values of this tuning parameter, and its effect on power and ASN.

6.10 Summary

This chapter has presented methods used to transform data to obtain the end treatment effect to be as planned, a fraction of that planned, zero, or negative, for continuous and binary endpoints. This has increased the number of trials to compare at each scenario, while maintaining the pattern of the original data, rather than just observing the original trial. Two designs and four assumptions have been compared, with the combination test design requiring fewer patients than originally planned most of the time.

The current trend assumption appears to decrease to zero the earliest of the four assumptions when there is no treatment effect, or it is opposite to that originally planned. However, it is highly variable when the treatment effect is equal to that planned, or a fraction of the planned effect, where larger sample size increases can be seen compared to the other three assumptions. Optimistic confidence limits have been shown to perhaps be a good compromise between the two currently debated assumptions in the CP calculation and may benefit from a futility boundary as they are able to perform well both when the treatment effect is close to that planned, or when no treatment effect is observed.

A particular strength of this chapter, together with Chapter 5, is that real-world data has been used in a design comparison, with observed bias that may not be incorporated in any simulated data comparisons. As far as I am aware, this is the largest investigation of uSSR designs applied to real-world data. Other comparisons typically use one or two examples and look at the observed effect at one time point only, rather than seeing how this effect varies throughout the trial (Mehta 2011; Jennison 2015). A limitation of the work in this chapter is that power and Type I error cannot be observed due to the small number of trials. The simulation work in the next chapter will be able to provide this information, which is missing from the real-world trial investigations, and will be able to provide answers for the remaining criteria to evaluate the methodology from Table 6.1. Chapter 7 contains a more in-depth simulation protocol, and presents results from the simulation work for continuous outcomes.

7 | Simulations for continuous outcomes

7.1 Introduction

Chapter 5 used real trial data, assuming that the treatment effect observed in the original analysis was the true treatment effect and compared three designs and four treatment effect assumptions in the CP calculation. Chapter 6 extended this work to investigate the extent of misspecification of the target effect size in the planning stage. The implications for CP and corresponding sample sizes for two designs were compared when observing the planned effect, a smaller effect than planned, no difference in treatments, or a negative effect, focusing only on promising zone and combination test designs. The data re-analysis section however was not able to compare the power of the designs or Type I error, which is what the simulation work will largely focus on. This chapter extends the data re-analysis work, using simulations so that the true treatment effect is known, not assumed. Additionally, a much larger number of trials can be compared through simulations. Expected sample size and power will be investigated for three designs:

- Promising zone
- Combination test
- Promising zone with futility.

Logistical features will also be investigated, such as pipeline patients, timing of interim analysis, and maximum allowed sample size.

7.2 Aims

This chapter aims to introduce a detailed simulation plan, explaining how data will be simulated, designs implemented, and objectives will be assessed. The simulation plan in this chapter will only look at continuous outcomes, with binary data being approximately normal if the sample size is large enough, or expected event rates close to 50%. An additional example of binary simulations where event rates are low, and n is relatively small is provided in the following chapter.

Following the simulation plan, the results for the continuous simulations will be presented, split by observed treatment effect vs planned for 1, $\frac{2}{3}$, $\frac{1}{3}$ and 0 times the planned effect size. The results will be discussed in terms of the evaluation criteria seen previously in Table 6.1. Specific comparisons and objectives will be described in the simulation protocol in Section 7.3.1, and have been informed by previous results in Chapters 5 and 6.

7.3 Simulation plan

This section provides details on aims for the simulation study, focusing particularly on continuous outcomes. Simulation objectives and comparisons will be the same for binary outcomes, and so will not be repeated in Chapter 8. Methods for generating data, choices of logistical parameters, and details on how objectives will be evaluated and compared are presented in this simulation plan.

7.3.1 Specific aims

Specific aims of the simulation study for both continuous and binary outcomes include:

- Simulate data for two treatment groups
- Calculate CP values using three future assumptions: Current trend, hypothesised effect and 80% optimistic confidence limit of the current trend
- Compare ASN and power using SSR rules from three designs: Promising zone with/without futility bound, and combination test design
- Investigate interim analysis timing every 5% through the trial from 50% onwards
- Explore combination test specific details, including four values of the γ parameter

used in the objective function, and consider two recruitment rates in line with that observed in the data re-analysis section

- Consider four values of n_{max}: 1.1, 1.5, 2 and 3 times the original sample size.

- Modify the true effect, to investigate CP, ASN and power when the observed effect is as planned, smaller than planned, or no difference in treatment groups.

7.3.2 Simulation methods

7.3.2.1 Data generation

For continuous outcomes, data will be sampled from two normal distributions; one for each treatment group. Two values of δ will be explored δ=0.2 and 0.4, giving both a "small" and "large" trial for consideration in line with the choices made in obtaining data for retrospective data reanalysis. The specific values of the mean difference and standard deviations are not important, only the ratio between the two (i.e. the treatment effect δ). SD has been arbitrarily chosen as 20, with a corresponding mean difference of 4 and 8 for δ=0.2 and 0.4 respectively. The control group will be sampled from a normal distribution with a mean of zero, and SD of 20. The treatment group will be sampled from a normal distribution with a mean of either 4 or 8, and a SD of 20. Trials will have a planned 90% power and two-sided significance level of 5%, and the number of patients will be calculated based on this information and the planned effect size, resulting in initially planned target sample sizes of n=1052 and 264 respectively.

A true effect smaller than planned is also investigated, using $\hat{\delta}_{obs} = \frac{2}{3}\delta_{plan}$, $\hat{\delta}_{obs} = \frac{1}{3}\delta_{plan}$ and $\hat{\delta}_{obs} = 0$. Again, SD will be kept consistent, but the treatment group will be sampled from a normal distribution with the mean value now multiplied by $\frac{2}{3}$, $\frac{1}{3}$, or 0 accordingly. The planned effect size will be maintained at either 0.2 and 0.4 and so original sample size n will not be affected.

To investigate the treatment effect at a larger than planned n^* patients, the number of random numbers generated for each simulated trial group will be 3 times the original n_{group}(=$\frac{1}{2}n$), corresponding to the maximum allowed sample size investigated. Random numbers 1 to n_{group} from each normal distribution will correspond to the original planned fixed sample size trial.

To maintain reproducibility, the seed number will be pre-specified and kept consistent throughout trials. Seed numbers used are 123 for the control group, and 1234 for the intervention group.

The number of repetitions to conduct will be based on a performance measurement. The treatment estimate from the originally planned n patients was calculated using three different seed numbers for a number of values of repetitions, and when the three observed treatment estimates to two decimal places were equal, the number of repetitions was considered to be sufficient. This resulted in 50,000 repetitions to be carried out.

7.3.2.2 Pipeline patients

The number of pipeline patients at the interim analysis particularly affects the combination test design, and therefore rate of recruitment needs to be taken into account. Two rates are considered in this investigation, and have been informed by percentages of pipeline patients in the data re-analysis section (Table 4.6). A long time to endpoint is unlikely to provide any saving in sample size and it needs to be carefully considered whether a SSR is appropriate to use in these situations. Therefore, this investigation focuses on a “short” or “medium” time to primary outcome only. A short outcome is considered to have approximately 1-3% additional recruited patients, and a medium outcome is considered to have an additional 20-22% recruited patients to those that have data available (i.e. those included in the calculation of CP). These values have been informed from Table 4.6, summarising the percentage of pipeline patients for short and medium outcomes from the data obtained for re-analysis.

7.3.2.3 Parameter choices

Interim analyses are considered from 50% onwards only, due to the findings of estimate stability in Chapter 5. To assess the uncertainty of when the interim analysis should take place beyond this point, values of n_1 are explored every 5% of the original sample size, from 50% to 95%.

Four values of n_{max} are considered; 1.1, 1.5, 2 and 3 times the original sample size estimate, n. Two values will be explored in the main part of the thesis, 1.1 and 2, while additional results can be found in Appendix E. Whilst 1.5 and 2 were investigated in previous chapters, it is thought that a small increase such as 10% would be interesting to investigate. Furthermore, the systematic review in Chapter 3 showed the most frequently used maximum cap was double the original sample size. Therefore 1.1. and 2 times the original sample size will be the focus of the simulation work in the main chapters.

Finally, the γ parameter required in the combination test design will be explored, and the impact on sample size and power will be discussed. Specific values of 0.0001, 0.0002, 0.0005 and 0.001 will be investigated. These values represent a range of feasible values of the acceptable gain in CP per one additional patient.

7.3.3 Outcomes

CP values using the current trend, hypothesised effect and an 80% optimistic confidence limit will be calculated, and will be used to inform the new sample size required, according to three designs: promising zone, combination test, and promising zone with a futility boundary. Designs and conditional power assumptions will be compared using ASN and power. ASN will be the mean new sample size (n^*) across 50000 repetitions. Power is calculated as the number of simulated trials that are significant given n^* has been recruited, whether decreased, remained the same, or increased as directed by the SSR rule used. Futility rules will be considered 'binding' (i.e. the trial stops immediately should CP fall below the 10% boundary considered, with no exceptions). ASN and power will compare CP assumption, SSR design, value of the γ parameter for combination test design, and values of n_{max}.

7.4 Simulation results

This section presents the results for the simulation work of the thesis, carried out according to the simulation plan in Section 7.3. Results for δ=0.2 or δ=0.4 were found to be very similar, and therefore only results for δ=0.4 are presented in the main thesis. Results for δ=0.2 are also provided in Appendix E for completeness.

7.4.1 True effect = 0.4

Table 7.1 shows the mean treatment difference, SD, effect size and a difference between the true difference and estimated mean difference, calculated after every 5% of patients from 50% through the trial onwards. In line with the simulation plan, the true mean difference is 8, SD=20, and corresponding effect size δ_{plan}=0.4, and therefore values should be as close to these as possible.

	Information fraction										
$\hat{\delta}_{obs} = \delta_{plan} = 0.4$	**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**	**100%**
Mean $\hat{d}$	7.9974	7.9984	8.0013	7.9996	8.0014	7.9986	8.0033	8.0029	8.0022	8.0037	8.0009
Mean SD	19.9583	19.9618	19.9615	19.9624	19.9646	19.9668	19.9707	19.9726	19.973	19.9751	19.977
Mean $\hat{\delta}_{obs}$	0.4023	0.4021	0.4022	0.4019	0.4019	0.4017	0.4017	0.4016	0.4015	0.4015	0.4013
Mean $(d-\hat{d})$	-0.0026	-0.0016	0.0013	-0.0004	0.0014	-0.0014	0.0033	0.0029	0.0022	0.0037	0.0009
$\hat{\delta}_{obs} = \frac{2}{3}\delta_{plan}$											
Mean $\hat{d}$	5.3307	5.3317	5.3347	5.3330	5.3347	5.3320	5.3367	5.3362	5.3355	5.3370	5.3343
Mean SD	19.9583	19.9618	19.9615	19.9624	19.9646	19.9668	19.9707	19.9726	19.973	19.9751	19.977
Mean $\hat{\delta}_{obs}$	0.2681	0.2680	0.2681	0.2680	0.2680	0.2678	0.2679	0.2678	0.2677	0.2677	0.2676
Mean $(d-\hat{d})$	-2.6693	-2.6683	-2.6653	-2.667	-2.6653	-2.668	-2.6633	-2.6638	-2.6645	-2.663	-2.6657
$\hat{\delta}_{obs} = \frac{1}{3}\delta_{plan}$											
Mean $\hat{d}$	2.6640	2.6650	2.6680	2.6663	2.6680	2.6653	2.6700	2.6696	2.6688	2.6703	2.6676
Mean SD	19.9583	19.9618	19.9615	19.9624	19.9646	19.9668	19.9707	19.9726	19.973	19.9751	19.977
Mean $\hat{\delta}_{obs}$	0.1340	0.1340	0.1341	0.1340	0.1340	0.1339	0.1341	0.1340	0.1339	0.1340	0.1338
Mean $(d-\hat{d})$	-5.3360	-5.3350	-5.3320	-5.3337	-5.3320	-5.3347	-5.3300	-5.3304	-5.3312	-5.3297	-5.3324
$\hat{\delta}_{obs} = 0$											
Mean $\hat{d}$	-0.0026	-0.0016	0.0013	-0.0004	0.0014	-0.0014	0.0033	0.0029	0.0022	0.0037	0.0009
Mean SD	19.9583	19.9618	19.9615	19.9624	19.9646	19.9668	19.9707	19.9726	19.973	19.9751	19.977
Mean $\hat{\delta}_{obs}$	-0.0001	-0.0001	0.0001	0.0000	0.0001	0.0000	0.0002	0.0002	0.0001	0.0002	0.0001
Mean $(d-\hat{d})$	-8.0026	-8.0016	-7.9987	-8.0004	-7.9986	-8.0014	-7.9967	-7.9971	-7.9978	-7.9963	-7.9991

Table 7.1: Mean difference, SD, treatment effect and difference from the true population value for values between 50% and 100% of the originally planned n=264 when $\delta = 0.4$

The mean difference is always within 0.0026 of the true difference, and by 100% original planned sample size (n=264) is within 0.0009 of the true value. SD is very slightly underestimated at all time points, but this does not impact the mean treatment effect, which is correct to 2 decimal places (0.40). This same pattern follows when the true effect is adjusted to $\frac{2}{3}$, $\frac{1}{3}$ or 0 times the original 0.4 value. Therefore, the simulations are estimating these values well, and have sufficient repetitions. As stated in the simulation plan, the number of repetitions were chosen by calculating the mean difference at 264 patients using three different seed numbers for generating random numbers, but this table also confirms that this value is sensible.

Mean conditional power from 50000 simulations have been calculated using three future treatment effect assumptions at interim time points between 50 and 90% of the original study duration, and are shown in Table 7.2. At all interim time points, and values of δ values investigated, the current trend assumption has the lowest CP values of the three assumptions, and the hypothesised assumption the highest. As the true value of δ decreases compared to that planned, all CP values decrease. CP values are >92% when δ equals δ_{plan}, decreasing to values >17% only, when there is actually no difference between the two groups.

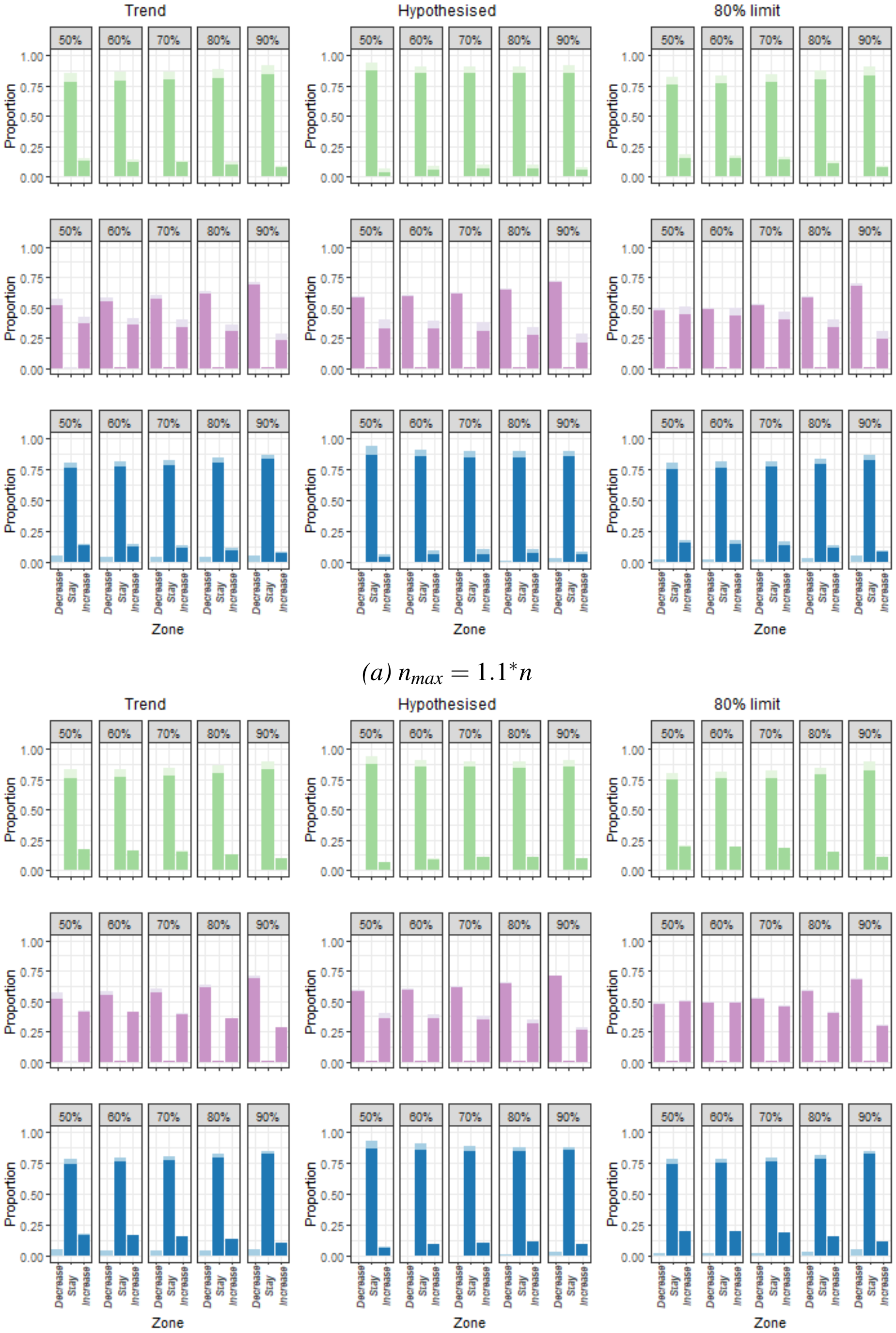
		Information fraction								
Conditional power		50%	55%	60%	65%	70%	75%	80%	85%	90%
δ=δ_{plan}=0.4	Trend	0.923	0.935	0.946	0.953	0.962	0.968	0.974	0.979	0.982
	Hypothesised	0.988	0.988	0.989	0.988	0.988	0.988	0.988	0.988	0.988
	80% limit	0.963	0.967	0.970	0.973	0.976	0.978	0.981	0.983	0.985
δ=$\frac{2}{3}\delta_{plan}$	Trend	0.756	0.771	0.786	0.799	0.813	0.824	0.839	0.853	0.863
	Hypothesised	0.944	0.936	0.929	0.923	0.918	0.912	0.906	0.900	0.895
	80% limit	0.854	0.855	0.857	0.859	0.863	0.864	0.869	0.873	0.877
δ=$\frac{1}{3}\delta_{plan}$	Trend	0.482	0.488	0.494	0.498	0.504	0.510	0.516	0.522	0.526
	Hypothesised	0.816	0.786	0.756	0.730	0.700	0.673	0.645	0.616	0.590
	80% limit	0.624	0.613	0.603	0.593	0.583	0.576	0.568	0.559	0.552
δ=0	Trend	0.218	0.210	0.203	0.198	0.192	0.187	0.182	0.177	0.173
	Hypothesised	0.584	0.523	0.465	0.418	0.367	0.327	0.285	0.247	0.217
	80% limit	0.337	0.312	0.289	0.271	0.251	0.236	0.219	0.202	0.189

Table 7.2: Mean conditional power values from 50000 repetitions for three treatment effect assumptions when δ=0.4 and n=264

All graphs of the combination test in the results section use a short outcome, with γ=0.0002. Alternative outcome timings and values of γ are compared in tables.

7.4.1.1 True effect = planned effect

Figure 7.1 shows the frequency of sample size states (i.e. decrease, remain, or increase) for three SSR designs: promising zone (green), combination test (pink), and promising zone with futility (blue). Darker shades indicate that the trial went on to be significant at the new recruited sample size n^* indicated by the SSR rule. A large number of trials were significant for all designs. It should be noted that when the observed effect is as planned, the proportion of significant trials should correspond to the pre-specified chosen power, 90%. This is also presented in a table later in this section.



(a) $n_{max} = 1.1^*n$

(b) $n_{max} = 2^*n$

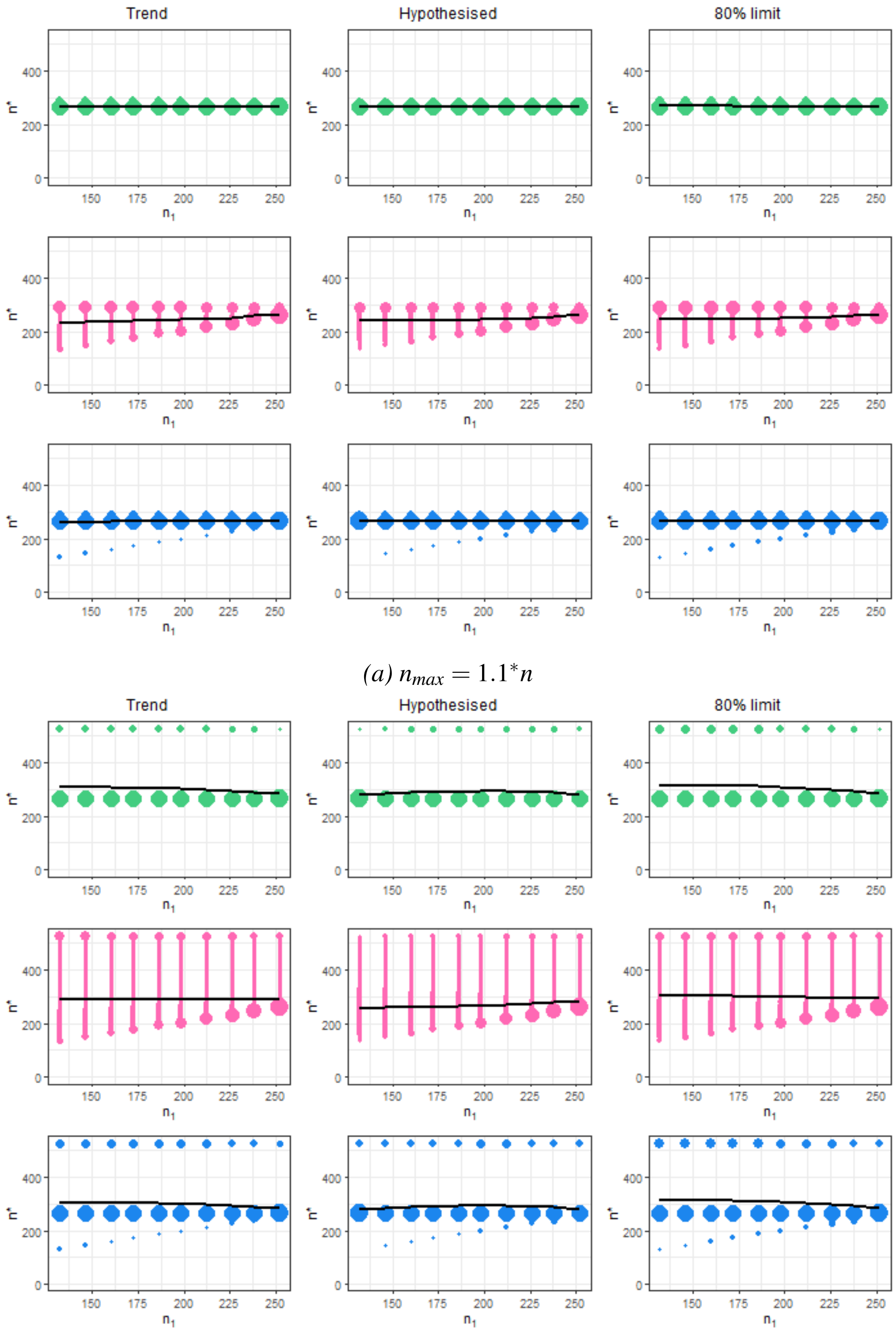
Figure 7.1: Sample size zones from 50000 simulations when $\delta = \delta_{plan}$ and n=264, for two values of n_{max}: comparing three designs and four observed treatment effects

The combination test design (pink) with γ=0.0002 and a short outcome shows more increases in sample size than either promising zone design, but also sees a large proportion of trials with $n^* < n$. Despite decreasing in sample size compared to the original n patients, trials are still predominantly found to be significant, at any interim timing. The 80% limit assumption (column 3) sees the highest proportion of increases in sample size for all designs, which decreases with a larger information fraction. A small number of trials stopped for futility for the final design (blue), of which were non-significant. This pattern is the same for n_{max}=1.1 and 2 times the original sample size.

Again comparing the three designs and three assumptions (using γ_2=0.0002 and a short outcome where appropriate), original sample size n_1 is plotted against new sample size n^* in Figure 7.2. Larger dots indicate a greater frequency of repetitions resulting at the n_1 and n^* combination. Colours have been kept consistent with designs (promising zone in green, combination test in pink, and promising zone with futility in blue), and the mean sample size from all 50000 repetitions is plotted with a black line.

Promising zone design remains fairly consistent at $n^* = n$ patients, with some increases up to n_{max}. Because of the higher n_{max} value, and therefore higher increases, mean sample size is slightly higher for n_{max}=2*n than 1.1*n in both promising zone designs. Mean sample size lines take the same shape between promising zone with and without a futility boundary. With a slight exception at $n_{max} = 2^*n$ using the 80% limit, required sample size increases with increasing n_1 for the combination test due to increasing n_{rec} values which this design depends on. Promising design with futility sees both increases and decreases (seen more clearly when $n_{max} = 2 * n$). The trend assumption has slightly more decreases earlier in the trial compared to hypothesised and 80% limit assumptions, which reverses as n_1 increases.

Table 7.3 shows ASN and power values for two primary outcome timings (short and medium) and four values of γ (γ_1=0.0001, γ_2=0.0002, γ_3=0.0005, γ_4=0.001). In general, the smaller the value of γ, the larger the ASN. This is particularly seen earlier in the trial (50% information fraction e.g.), meeting at around 269-301 patients by 90% through the trial with a short outcome. The smallest value of γ has the highest power for all assumptions and both



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure 7.2: Sample size from 50000 simulations when $\delta = \delta_{plan}$ and n=264, for two values of n_{max}: comparing three designs and three treatment effect assumptions

δ=0.4		SHORT					MEDIUM				
δ=δ_{plan}		Information fraction					Information fraction				
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	305	305	303	299	297	311	316	320	322	308
	Power	95.14	96.81	97.50	97.75	97.42	95.69	97.18	97.78	97.92	97.51
γ_2	ASN	288	290	290	290	291	296	303	309	314	303
	Power	93.82	95.83	96.70	97.33	97.16	94.58	96.38	97.26	97.62	97.31
γ_3	ASN	261	267	271	275	281	272	283	294	302	294
	Power	91.15	93.71	95.28	96.35	96.50	92.50	94.87	96.28	96.89	96.75
γ_4	ASN	232	243	252	262	273	246	263	278	291	287
	Power	87.43	90.79	93.14	94.87	95.70	87.43	90.79	93.14	94.87	95.7
HYPOTHESISED											
γ_1	ASN	276	276	276	278	285	278	282	290	300	296
	Power	95.85	96.71	97.15	97.49	97.28	95.88	96.81	97.34	97.66	97.4
γ_2	ASN	257	260	264	270	280	260	269	281	294	292
	Power	94.27	95.44	96.23	96.92	96.95	94.36	95.66	96.55	97.18	97.11
γ_3	ASN	233	240	248	259	274	238	253	269	286	287
	Power	91.09	93.17	94.46	95.67	96.23	91.44	93.68	95.21	96.21	96.46
γ_4	ASN	215	225	237	251	269	223	241	261	280	282
	Power	87.63	90.06	92.34	94.13	95.30	88.37	91.19	93.65	95.01	95.63
80% Limit											
γ_1	ASN	323	322	317	309	301	326	330	331	329	313
	Power	97.80	98.43	98.39	98.31	97.68	97.95	98.56	98.59	98.43	97.78
γ_2	ASN	304	305	302	298	295	308	314	318	320	306
	Power	96.98	97.87	98.12	98.07	97.48	97.17	98.11	98.36	98.27	97.62
γ_3	ASN	274	279	280	281	285	281	291	300	307	297
	Power	94.92	96.48	96.98	97.34	97.01	95.46	96.99	97.61	97.71	97.23
γ_4	ASN	246	254	260	267	276	256	270	284	295	289
	Power	91.91	94.14	95.09	96.08	96.3	92.93	95.17	96.31	96.78	96.59

*Table 7.3: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when δ=δ_{plan}, n=264, n_{max}=2*n*

endpoints, with values ranging between 95.14 (current trend, short outcome) and 98.59 (80% limit, medium outcome), which is well above the pre-specified power of 90%. Increasing the value of γ decreases the power, and γ_4=0.001 has power that has dropped below the nominal rate for the 50% time point under the current trend or hypothesised assumptions, but increases with increasing information fraction. A 60% interim time point using γ_4 would be acceptable to use with the current trend or hypothesised assumption (power = 90.79 and 90.06% respectively). Choice of γ would depend on the assumption used, and the timing of the interim analysis, but could be chosen such that power equals the specified level through simulations, for instance.

The hypothesised assumption results in the lowest ASN and 80% limit the highest, while the current trend assumption lies somewhere between the two, although by 90% there is at most 1 patient difference between the three assumptions. This pattern however does not always hold for power. Some slight sample size savings can be seen at one or more interim time points for γ_3 and γ_4 for all assumptions. A sample size saving, whilst maintaining power at 90% would be the ideal balance, also corresponding with Table 6.1 for evaluating methodology. The closest example to this is γ_4 using the hypothesised assumption at 60% data available of the original n=264 patients and a short outcome.

δ=δ_{plan}=0.4		**Information fraction**									
Promising Zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	310	309	308	306	306	304	300	296	292	284
	Power	93.05	93.24	93.41	93.44	93.56	93.55	93.46	93.37	93.12	92.44
Hypothesised	ASN	282	286	289	292	293	294	294	292	289	283
	Power	93.32	94.04	94.70	95.11	95.39	95.6	95.40	95.08	94.51	93.23
80% limit	ASN	317	317	316	315	313	310	305	300	294	285
	Power	94.45	94.49	94.59	94.49	94.42	94.25	94.01	93.64	93.29	92.49
Promising zone with Futility											
Trend	ASN	304	304	304	303	303	301	298	294	290	283
	Power	91.20	91.73	92.27	92.52	92.88	93.04	93.07	93.1	92.91	92.31
Hypothesised	ASN	282	286	289	292	293	294	293	292	289	283
	Power	93.32	94.04	94.70	95.11	95.38	95.59	95.39	95.07	94.48	93.20
80% limit	ASN	315	315	314	313	311	308	304	298	293	284
	Power	94.09	94.19	94.28	94.20	94.16	94.01	93.79	93.45	93.11	92.37

*Table 7.4: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions when δ=0.4, n=264 and and n_{max} = 2*n for promising zone with and without a futility boundary*

Table 7.4 shows a comparison of ASN and power across the promising zone designs considered designs for three CP assumptions, with and without a futility boundary. Both designs have power > 90%, and so power is greater than originally specified and maintained regardless of interim timing. However, it is also one of the criteria that a design should not have too high power, especially with an increase in the number of patients recruited, for both ethical and logistical reasons of recruiting more patients than necessary to a trial. The addition of a futility bound brings down the expected sample size, but only by 1-6 patients between 50 and 90% data available.

Depending on the choice of γ parameter and timing of the interim analysis, the combination test design can be chosen to have the desired power, whilst even seeing decreases in expected sample size, when the observed effect is exactly as planned. On the other hand, if a trial is expecting to see exactly as planned, a SSR may not be necessary at all, and an alternative fixed sample size would likely be considered in this scenario. Therefore the comparisons of alternative observed differences will be useful in evaluating the methodologies. Promising zone designs do increase ASN and power, but increases are small, and the addition of a futility boundary reduces this increase slightly.

Therefore, these simulations conclude that when the observed effect is as planned, the combination test design would be able to maintain power at the desired level and decrease expected sample size compared to a promising zone design. Therefore, in terms of the methodology criteria table (Table 6.1), this design is the best performing in this scenario. Whilst the incorporation of a futility boundary decreases ASN and power (whilst always remaining above the specified level) compared to the promising zone design, a small number of trials stop for futility, which is an undesirable characteristic when $\delta = \delta_{plan}$. However, this number is very small, particularly for the 80% limit and hypothesised assumptions. The risk of incorrectly stopping for futility should be taken into consideration when designing a trial, and weighed up against the smaller ASN and power.

7.4.1.2 True effect = $\frac{2}{3}$ planned effect

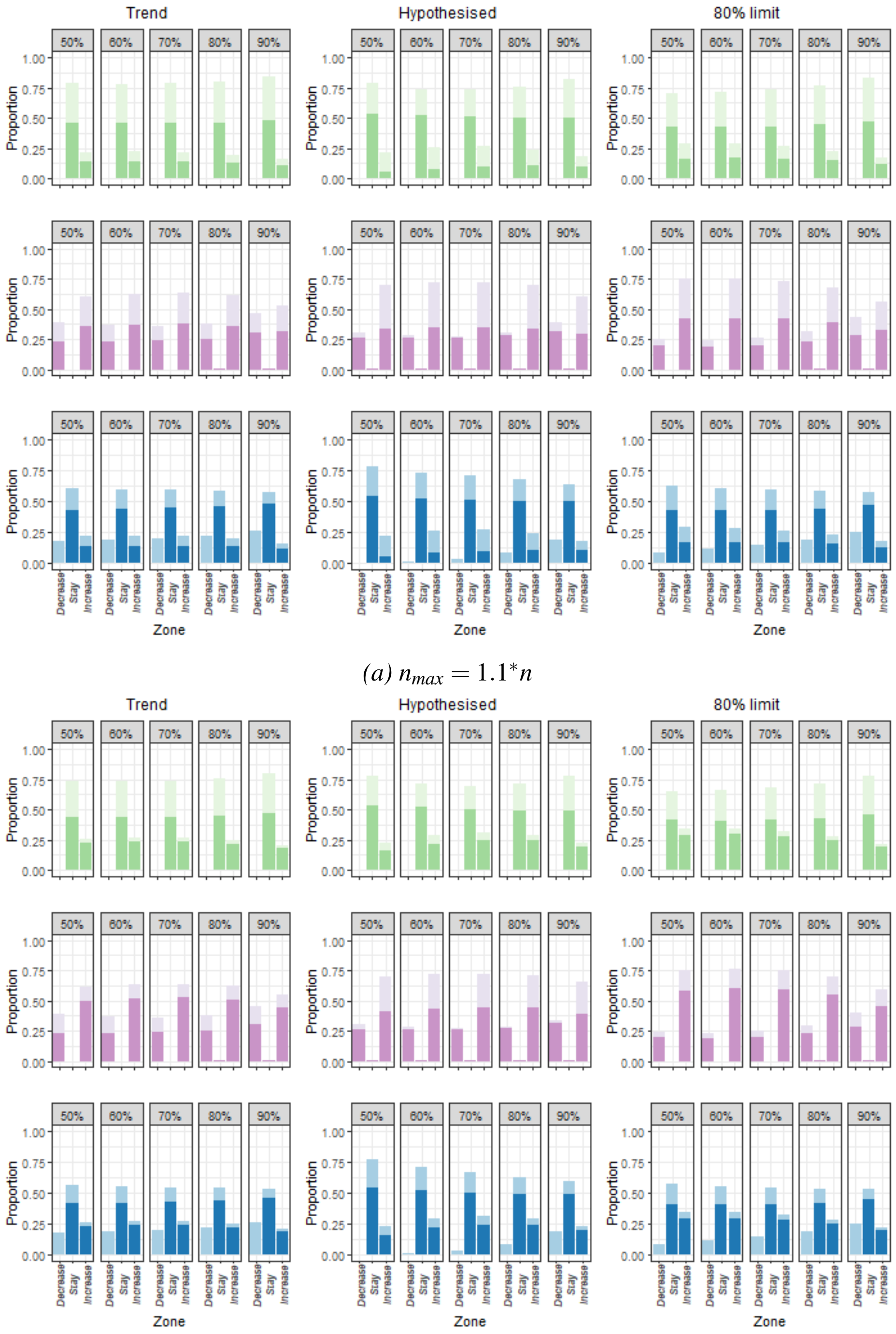
A value of $\frac{2}{3}$ may be considered smaller than hoped, but perhaps still clinically relevant, or of interest. Corresponding with Table 6.1, increases in sample size are fine, but need to also coincide with higher power. Additionally, if $\frac{2}{3}$ is to be considered clinically meaningful, there is no wish to stop the trial early for futility if applicable. This section presents the results for the comparison of the designs and CP assumptions in this scenario.

Figure 7.3 shows significance of trials that decreased, remained or increased for two values of n_{max} (again, 1.1 and 2 times the original n patients). As the timing of the interim analysis increases, the number of trials being increased in promising zone with/without futility are decreasing, with the exception of the hypothesised assumption. The largest frequency of increases occurs at 50% for the current trend and 80% limit assumptions, and at 70% of data available using the hypothesised assumption.

Incorporating a futility bound has decreased the proportion of non-significant trials when sample size remains the same, and all trials that are stopped early go on to be non-significant. The combination test design using γ_2 now has a larger proportion of increases compared to decreases, most prominent in the 80% limit and hypothesised assumptions. However, an increase in sample size does not always correspond with a significant result. As the interim timing increases from 50 to 80%, slightly more trials are found to be significant under the current trend or hypothesised assumptions, whereas the opposite effect is seen in the 80% limit assumption (albeit small).

Figure 7.4 shows the new required sample size at 10 interim timings between 50 and 95% data available, for three SSR designs, three CP assumptions and two values of n_{max}. The promising zone design (green) again sees very little change in expected sample size, with a maximum increase of 10% being too small to see much of a change in the ASN line. However, this line has increased in the $n_{max} = 2^{*}n$ patients compared to when $\delta = \delta_{plan}$, due to more increases being observed.

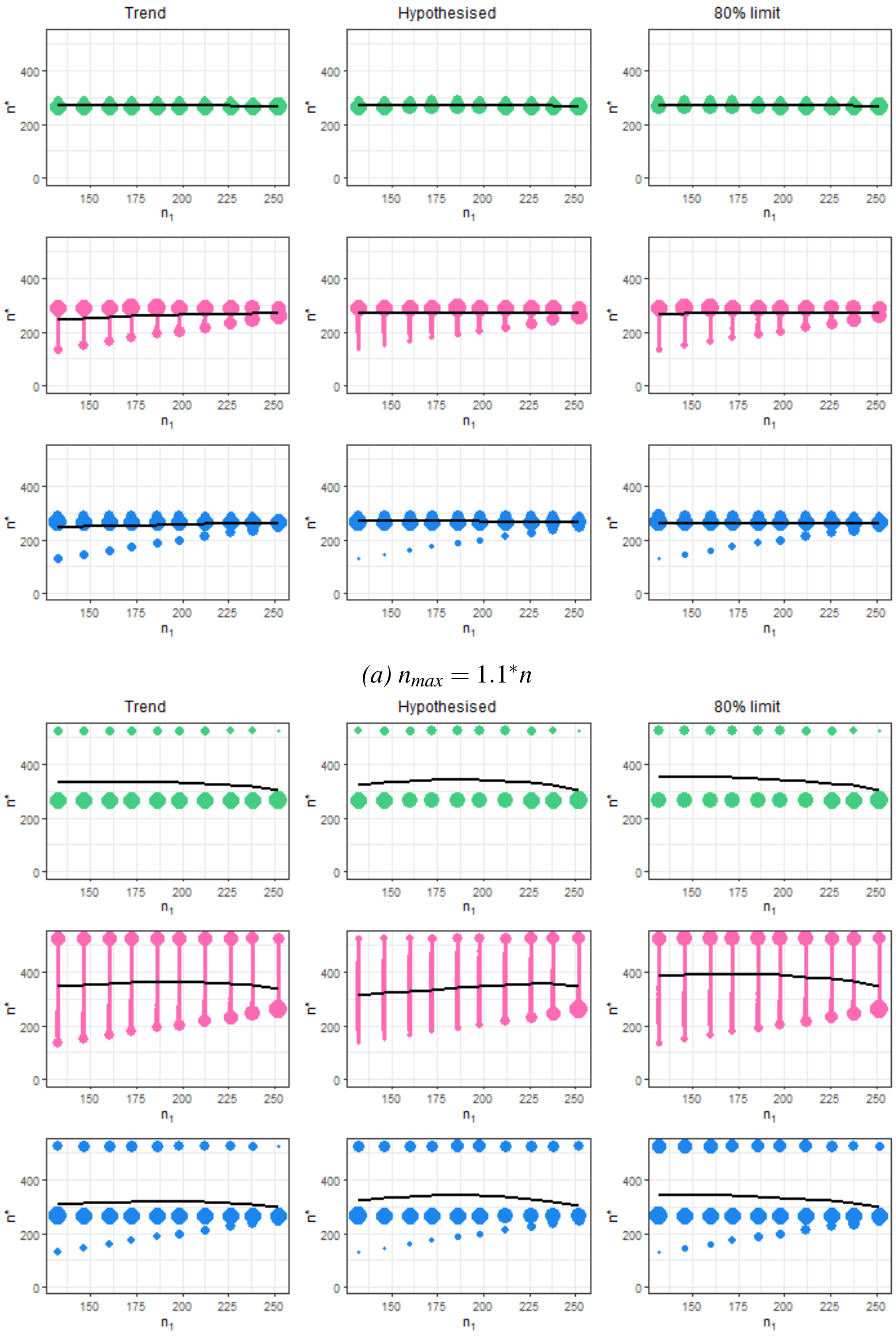
Both values of n_{max} see the combination test design (pink) see more instances of $n^{*} = n_{max}$, with some decreases also being observed. This results in higher expected sample size



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure 7.3: Sample size zones from 50000 simulations when $\delta = \frac{2}{3}\delta_{plan}$ and n=264, for two values of n_{max}: comparing three designs and four observed treatment effects



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure 7.4: Sample size from 50000 simulations when $\delta = \frac{2}{3}\delta_{plan}$ and n=264, for two values of n_{max}: comparing three designs and three treatment effect assumptions

lines than in the previous example. The current trend and 80% limit assumptions see a small increase in ASN to start, before decreasing as n_1 increases. The hypothesised effect however steadily increases in ASN between 50 and 90% data availability, before a small drop at 95% through.

Due to more trials stopping for futility (blue) under the current trend and 80% limit, $n_{max} = 1.1^{*}n$ sees a slight decrease in ASN compared to when the planned effect is observed. However, when $n_{max} = 2^{*}n$, the line has increased compared to when $\delta = \delta_{plan}$. Whilst the line takes the same shape as the promising zone design without futility for all three assumptions, but slightly lower in terms of n^{*}.

Table 7.5 investigates ASN and power between two lengths of primary outcome availability and the same four values of (γ=0.0.0001, 0.0002, 0.0005, 0.001).

Similarly to when observed δ is as planned, the larger values of γ decrease both ASN and power for all assumptions and percentage of pipeline patients (linked to timing of endpoint). In the scenario where δ is smaller than planned, power is a less important consideration, in the sense that it is not necessary for power to have reached a pre-specified limit. For comparative purposes however, both ASN and power are presented.

Again, the smallest values of ASN and power occur with the larger values of γ, with a short endpoint, a 50% data available interim analysis, using the hypothesised effect assumption. With the same values of γ, the trend assumption has comparatively higher values of ASN and power than hypothesised, and again the 80% limit assumptions sees the highest values for the three assumptions. However, if it's the case that an alternative value of γ would be chosen for each assumption, alternative rows could be compared. For instance, for hypothesised and trend assumptions, γ_3 saw values >90% for all time points, whereas this was seen for γ_4 for the 80% limit assumption. Thus, comparing 50% interim timing values of 67.77% power and 305 patients (80% limit) with 68.0% power with 311 patients (current trend) may be more appropriate. Additionally, the only instance that ASN is smaller than the original planned sample size is under the hypothesised effect assumption with γ_4=0.001 value.

δ=0.4		**SHORT**					**MEDIUM**				
δ=$\frac{2}{3}\delta_{plan}$		**Information fraction**					**Information fraction**				
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	369	377	381	378	366	375	385	390	390	373
	Power	75.27	77.2	78.06	77.61	75.04	75.20	77.29	78.13	77.69	75.04
γ_2	ASN	348	358	364	363	354	356	367	375	377	362
	Power	72.93	75.25	76.59	76.70	74.49	72.96	75.50	76.77	76.85	74.50
γ_3	ASN	311	324	333	337	334	322	336	348	354	343
	Power	68.00	71.15	73.27	74.34	72.99	68.34	71.61	73.73	74.73	73.08
γ_4	ASN	266	285	299	308	313	280	301	317	329	323
	Power	61.55	65.85	68.59	70.66	70.73	61.55	65.85	68.59	70.66	70.73
HYPOTHESISED											
γ_1	ASN	337	349	358	365	367	337	350	362	372	372
	Power	70.57	72.55	73.78	74.19	72.70	70.58	72.58	73.86	74.3	72.77
γ_2	ASN	315	329	341	351	357	315	331	346	360	363
	Power	67.33	69.53	71.26	72.25	71.44	67.37	69.63	71.45	72.45	71.53
γ_3	ASN	284	301	317	332	342	285	305	324	343	349
	Power	61.95	65.04	67.31	69.02	69.31	62.05	65.30	67.76	69.48	69.50
γ_4	ASN	259	278	297	315	328	261	284	306	328	335
	Power	57.30	60.74	63.78	66.13	67.12	57.65	61.28	64.65	66.98	67.43
80% Limit											
γ_1	ASN	408	413	410	400	380	410	416	416	409	386
	Power	79.65	80.28	80.08	78.78	75.41	79.63	80.30	80.10	78.84	75.42
γ_2	ASN	387	393	392	384	367	390	397	399	395	374
	Power	77.76	78.82	79.07	78.08	75.02	77.75	78.94	79.11	78.19	75.07
γ_3	ASN	349	357	360	356	345	354	364	370	370	354
	Power	73.74	75.47	76.43	76.19	73.79	73.86	75.74	76.72	76.47	73.87
γ_4	ASN	305	317	324	326	323	313	328	339	344	333
	Power	67.77	70.54	72.27	73.11	71.83	68.09	71.06	72.91	73.71	72.02

Table 7.5: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when $\delta=\frac{2}{3}\delta_{plan}$, n=264, $n_{max} = 2^{*}n$

Table 7.6 shows ASN and power for the promising zone design with or without a futility boundary. Using the hypothesised assumption, there is almost no difference between the two versions of the design. However, for the current trend and 80% limit assumptions, the futility design decreases both sample size and power. The greatest differences between the two designs is at 50% data availability, with smaller differences observed at the 95% point.

Whilst the 80% limit assumption sees the highest power at the 50% data available time point, this value decreases with increasing values of n_1. However, both the current trend and hypothesised assumptions see an increase in power with increasing n_1, before decreasing beyond 65% and 70% through respectively. The same pattern can be observed for ASN as for power.

δ=$\frac{2}{3}\delta_{plan}$		**Information fraction**									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	334	335	335	336	334	333	330	325	318	304
	Power	66.47	66.87	67.07	67.23	66.98	66.79	66.49	65.82	64.86	63.03
Hypothesised	ASN	324	333	340	344	346	345	341	334	324	307
	Power	69.45	71.38	73.05	73.94	74.62	74.50	73.63	71.69	69.12	65.21
80% limit	ASN	355	355	354	353	349	345	339	331	322	305
	Power	70.76	70.68	70.48	70.12	69.43	68.66	67.75	66.69	65.25	63.10
Promising zone with Futility											
Trend	ASN	310	313	316	319	320	319	318	316	312	300
	Power	64.04	64.83	65.43	65.80	65.86	65.82	65.72	65.26	64.44	62.77
Hypothesised	ASN	324	333	340	343	343	342	337	329	319	303
	Power	69.47	71.4	73.06	73.94	74.63	74.49	73.61	71.67	69.08	65.14
80% limit	ASN	345	345	343	342	338	334	329	323	316	302
	Power	70.31	70.2	69.96	69.58	68.92	68.19	67.29	66.25	64.92	62.85

Table 7.6: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions and three designs when $\delta=\frac{2}{3}\delta_{plan}$, n=264 and and $n_{max} = 2^*n$

If a value of $\delta = \frac{2}{3}\delta_{plan}$ is still considered clinically meaningful, all three scenarios have increased in sample size, in order to try and increase power. Whilst nominal levels are no longer obtained, more trials have been able to achieve a significant result despite seeing a smaller effect size than that planned. If this magnitude of effect is however not considered to be clinically relevant, it would be better to stop recruiting and use resources for an alternative trial. The hypothesised assumption would give the lowest expected sample size whilst maintaining high power in this scenario. Alternatively, the promising zone with futility design could be beneficial at any time point, as power is similar to the regular promising zone design for all three assumptions, but with a lower ASN.

The criteria for evaluating the methodology in Table 6.1 condones sample size increases, but only with a subsequent increases in power. It may vary on a cases by case basis between a trade-off for higher power but lower ASN, and each trial may have a different idea on which is the more favourable objective. The hypothesised assumption may see the highest power, but the current trend assumption sees the lowest ASN.

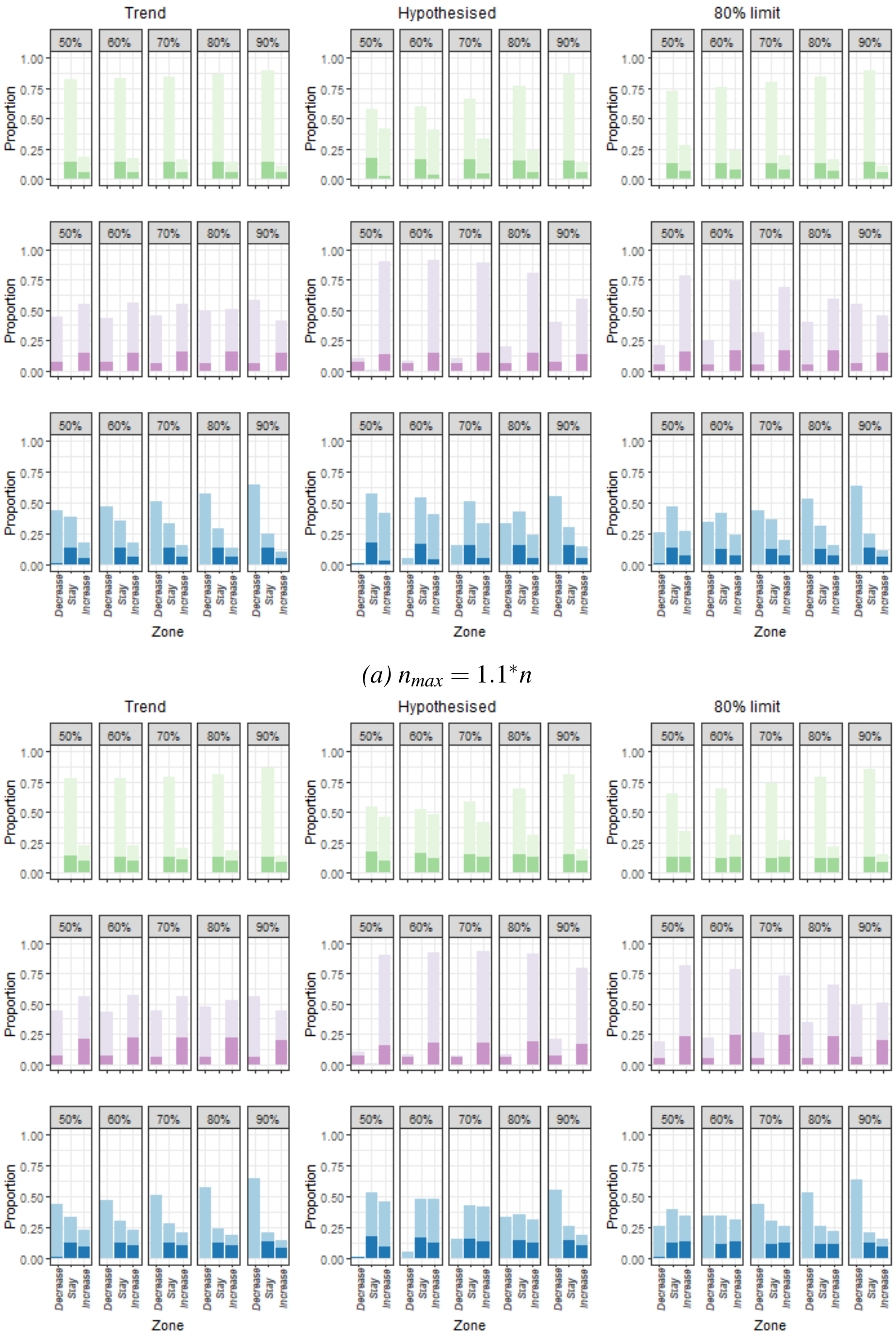
7.4.1.3 True effect = $\frac{1}{3}$ planned effect

A scenario considering $\delta = \frac{1}{3}\delta_{plan}$ represents an observed effect that is smaller than planned but too small to be clinically relevant. A sample size increase therefore is less constructive - even if able to push the trial to achieve statistical significance. Therefore, stopping earlier than the original n patients may be desirable, alongside fewer sample size increases. Researchers in this situation may wish only to focus on the lower ASN criteria identified in Table 6.1 to choose a design, as criteria based on power may not be relevant here.

Figure 7.5 shows significance of trials in the three possible decisions in terms of sample size for $\delta = \frac{1}{3}\delta_{plan}$, again for three designs, three assumptions and two values of n_{max}.

The promising zone design largely stays at the original n patients, which is the smallest value of n^* that this design allows. For all three assumptions, the number of trials that increase in sample size decreases with increasing values of n_1. The current trend assumptions sees the largest number of trials remaining at n patients, followed closely by the 80% limit. The hypothesised assumption however, has a much larger number of increases in sample size at 50-60% data availability. By 90% through the trial, similar proportions of remain/increases can be seen as the other two designs.

The addition of a futility boundary (blue) sees the same proportion of increases across the three investigated assumptions. However, a large proportion of those that previously remained at n patients in the regular promising zone design have now stopped early, at the corresponding value of n_1 patients with data available. At 50% through, a very small number of the trials that have stopped for futility actually have a significant result with



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure 7.5: Sample size zones from 50000 simulations when $\delta = \frac{1}{3}\delta_{plan}$ and n=264, for two values of n_{max}: comparing three designs and four observed treatment effects

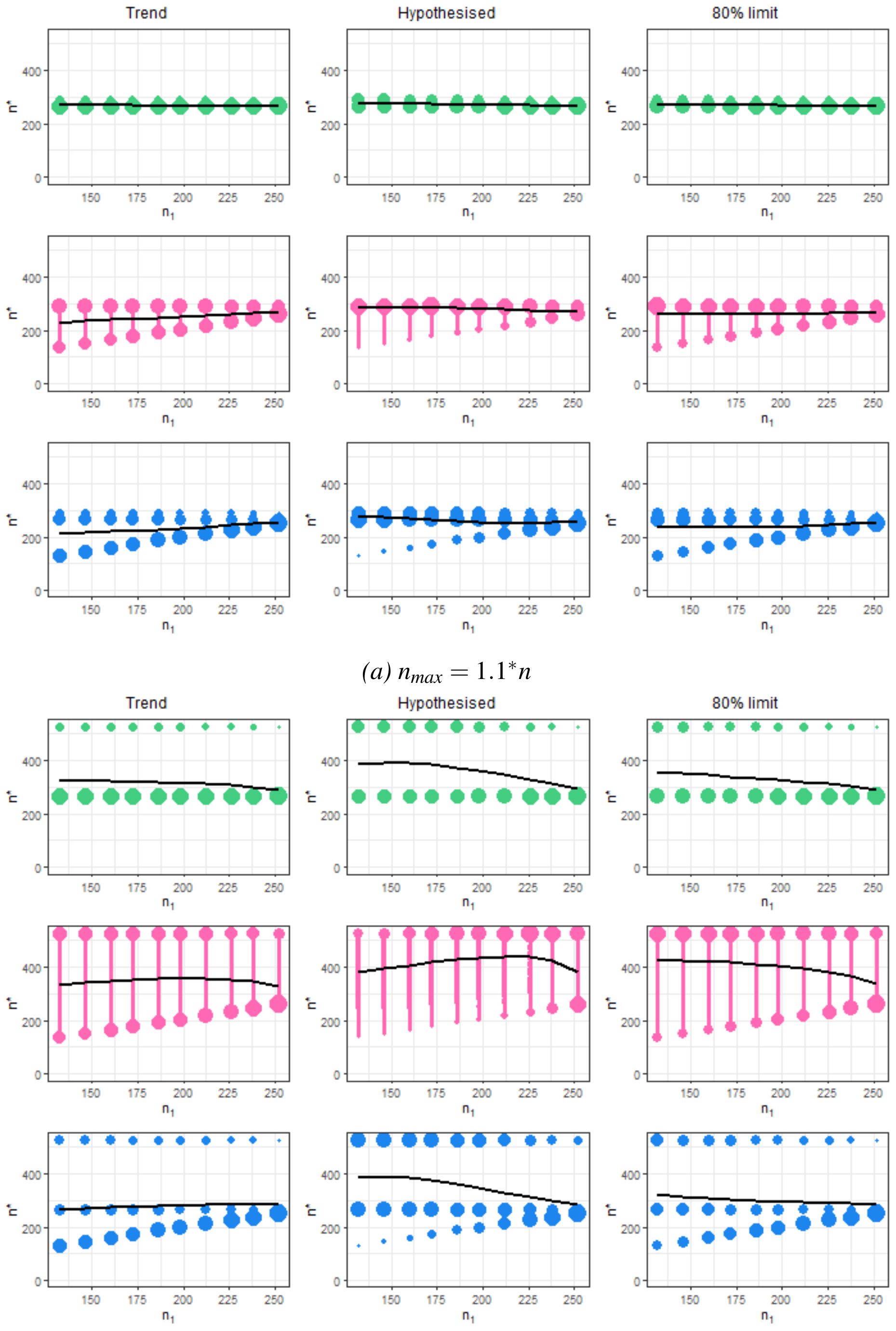
n_{rec} patients. The current trend assumption sees the largest proportion stopping for futility, again followed by 80% limit assumption. However, the hypothesised have very few trials stopping early for this design with the hypothesised assumption at 50% data availability, which greatly increases with increasing n_1 values.

Finally, the combination test (pink) sees a much larger proportion of increases in sample size, although the magnitude of such an increase cannot be seen with this graph (see Figure 7.6). Whilst all assumptions see a decrease in proportion of trials increasing sample size at the interim analysis with increasing data availability, the three assumptions start off at very different heights at the 50% time point. The hypothesised assumptions sees the highest proportion increasing at 50%, followed by the 80% limit. Additionally, the magnitude of the decrease in proportion increasing between 50% and 90% depends on the value of n_{max}. By 90% through the trial, the number of trials increasing is higher for $n_{max} = 2^*n$ than for 1.1^*n.

Figure 7.6 now looks at the magnitude of this sample size increase/decrease, and the impact on ASN with increasing information fractions.

When $n_{max} = 1.1^*n$, the promising zone sees very little difference in terms of the expected sample size line. The combination test design sees slightly more decreases using the current trend or 80% limit assumptions and correspond to a slight decreased expected line. The hypothesised effect assumption however is very slightly raised between 50 and 70% through the trial. This pattern is also observed in the promising zone design with futility: a decrease in the expected line for the trend and 80% limit assumptions due to increased stopping for futility, and a slightly raised line early in the hypothesised effect assumption graph.

Whilst the differences between $\delta = \frac{2}{3}\delta_{plan}$ and $\frac{1}{3}\delta_{plan}$ ASN lines are fairly subtle when $n_{max} = 1.1^*n$, differences are more visible when $n_{max} = 2^*n$, particularly in the case of the hypothesised assumption. In all three designs, the expected sample size has greatly increased. For either promising zone design, this increase is most prominent between 50 and 70% through the trial, and very few differences are observed by 90% through the trial.



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure 7.6: Sample size from 50000 simulations when $\delta = \frac{1}{3}\delta_{plan}$ *and n=264, for two values of* n_{max}*: comparing three designs and three treatment effect assumptions*

However, increases are observed at all interim time points for the combination test design, with many more instances of $n^* = n_{max}$ having been observed. For the current trend and 80% limit assumptions, both combination test and promising zone with futility see more cases of $n^* = n_{rec}$ and therefore have a smaller expected sample size compared to $\delta = \frac{2}{3}\delta_{plan}$.

δ=0.4		**SHORT**					**MEDIUM**				
δ=$\frac{1}{3}\delta_{plan}$		**Information fraction**					**Information fraction**				
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	356	370	375	374	359	372	386	393	393	369
	Power	28.98	29.54	29.43	28.32	26.56	27.72	28.49	28.52	28.03	26.39
γ_2	ASN	335	350	357	358	347	353	369	377	379	358
	Power	28.22	28.79	28.82	27.90	26.30	26.82	27.64	27.87	27.61	26.1
γ_3	ASN	297	315	325	331	327	318	337	349	355	339
	Power	26.33	27.14	27.62	26.81	25.56	24.77	25.92	26.54	26.55	25.35
γ_4	ASN	251	274	290	303	307	276	300	317	330	319
	Power	23.87	24.68	25.59	25.19	24.68	23.87	24.68	25.59	25.19	24.68
HYPOTHESISED											
γ_1	ASN	404	428	445	455	436	404	428	446	457	440
	Power	24.01	24.87	25.22	25.38	24.61	24.01	24.87	25.23	25.35	24.56
γ_2	ASN	381	408	430	442	424	381	408	430	445	428
	Power	22.74	23.48	24.19	24.58	24.10	22.73	23.49	24.19	24.56	24.07
γ_3	ASN	345	377	404	418	401	345	378	405	423	407
	Power	20.89	21.75	22.71	23.32	23.13	20.89	21.79	22.71	23.34	23.14
γ_4	ASN	314	349	378	393	377	315	350	381	401	384
	Power	19.39	20.31	21.55	22.17	22.30	19.40	20.46	21.64	22.31	22.34
80% Limit											
γ_1	ASN	447	443	433	415	382	452	451	444	429	390
	Power	29.71	29.85	29.66	28.63	26.62	29.18	29.28	29.11	28.43	26.51
γ_2	ASN	425	422	413	397	368	432	432	427	413	377
	Power	28.72	29.32	29.19	28.25	26.34	28.10	28.65	28.57	28.04	26.19
γ_3	ASN	381	381	376	365	344	392	396	394	386	355
	Power	27.29	27.94	28.09	27.25	25.70	26.49	27.15	27.36	27.04	25.55
γ_4	ASN	325	331	332	330	320	341	350	355	354	332
	Power	24.98	25.91	26.36	25.70	24.83	23.97	24.97	25.52	25.50	24.63

Table 7.7: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when δ=$\frac{1}{3}\delta_{plan}$, n=264, $n_{max} = 2^*n$

Table 7.7 shows ASN and power for four values of γ used in the combination test design for $\delta = \frac{1}{3}\delta_{plan}$. The current trend design sees the lowest value of ASN out of the three considered assumptions for all values of γ. However this is not the case for power, which is sometimes smaller and sometimes larger, depending on information fraction and comparative assumption. For instance, whilst the hypothesised effect sees the largest ASN values from 70% data availability point onwards, it is also associated with lower power values. This indicates that the design is increasing sample size where an increase has not perhaps been appropriate. For instance, using γ_4 with a short endpoint and 50% interim timing, the current trend assumption requires 251 patients to obtains 23.9% power, despite a decrease in sample size. The hypothesised effect however requires 314 patients, an increase of 50 patients, yet obtains 19.4% power; more patients for a lower power.

Looking at a trade-off between more patients and resulting power, the current trend assumption appears to be the most suitably behaved assumption, having similar but slightly less power than with the 80% limit assumption, but also with a lower ASN. This pattern applies regardless of being a short or medium end point. The hypothesised assumption however has not dealt with the situation of seeing a much smaller effect than planned well; resulting in high ASN and low power comparatively.

For all assumptions, increasing values of γ result in lower ASN, and lower power. A medium endpoint sees a higher ASN than a short endpoint, but largely sees lower power (albeit this difference is small).

Table 7.8 shows power and ASN for both versions of the promising zone design for the three assumptions. The current trend assumption again sees the lowest value of ASN in either design, whilst the hypothesised assumption consistently sees the greatest ASN value. The same pattern for ASN can also be seen for power values.

The design that allows for futility has lower values of ASN for all three assumptions, but power is not always lower when using the hypothesised assumption. In terms of Table 6.1 for evaluating the methodologies,the best design when observing a small (and therefore not clinically relevant) effect would have very few increases in sample size. If sample size increases do occur, the smaller the magnitude of increase, the better. In these terms, the

δ=$\frac{1}{3}\delta_{plan}$		Information fraction									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	324	324	324	322	319	317	313	308	302	292
	Power	23.31	23.46	23.25	23.31	23.06	22.97	22.67	22.13	21.56	20.63
Hypothesised	ASN	386	392	391	386	375	362	346	329	315	295
	Power	26.48	27.62	28.5	28.93	28.93	28.41	27.4	26.03	24.45	22.04
80% limit	ASN	356	351	346	340	334	328	321	313	304	293
	Power	25.85	25.61	25.2	24.86	24.35	23.87	23.24	22.41	21.71	20.62
Promising zone with Futility											
Trend	ASN	266	271	275	277	279	281	284	285	285	283
	Power	22.34	22.64	22.51	22.67	22.56	22.51	22.32	21.84	21.34	20.48
Hypothesised	ASN	385	390	386	378	363	347	329	312	300	287
	Power	26.8	27.81	28.61	29.02	29.00	28.45	27.44	26.05	24.43	22
80% limit	ASN	321	316	310	304	300	296	293	291	288	284
	Power	25.83	25.5	25.05	24.68	24.17	23.68	23.05	22.23	21.53	20.47

Table 7.8: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions and three designs when δ=$\frac{1}{3}\delta_{plan}$, n=264 and and $n_{max} = 2^{*}n$

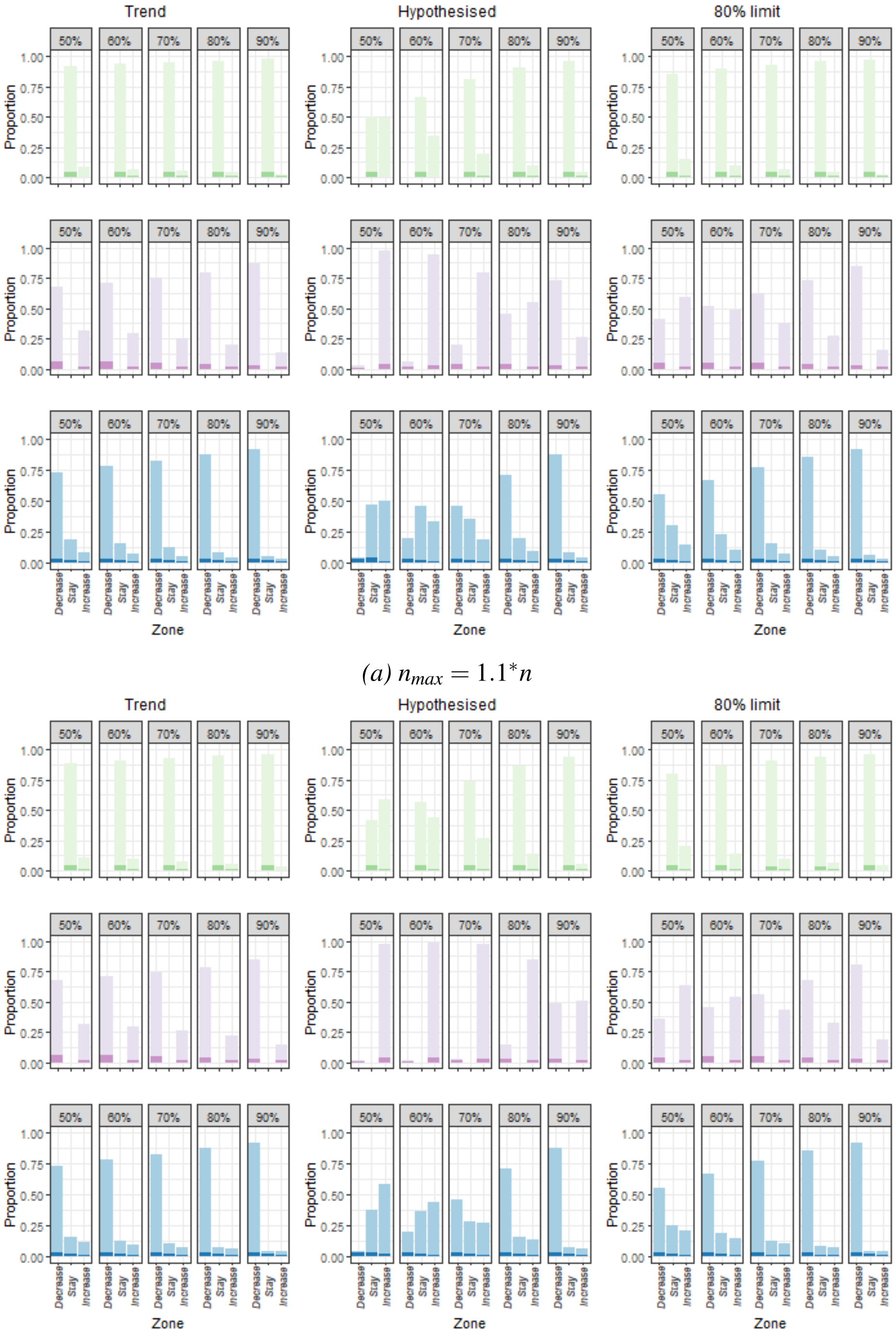
best design to deal with this scenario would be the promising zone deisgn with a futility boundary, due to the lowest values of ASN. The design would also benefit from an earlier interim analysis, as ASN increases with increasing information fraction. Additionally, the combination test with γ_4=0.001 using the current trend assumption could be considered to behave well in the scenario of $\delta = \frac{1}{3}\delta_{plan}$, seeing low values of ASN, and even decreasing in sample size at the 50% information fraction point.

7.4.1.4 True effect = zero

There is always the possibility, even for a superiority trial, that the intervention under investigation may truly be no different to the control group, and a zero effect (or even negative) may be observed. If this is the case, it would be ideal to be able to stop the trial as early as possible, and certainly not to increase in sample size, corresponding with Table 6.1. Additionally, it should be noted that when the true effect is zero, power is to be thought of as Type I error, and should be no more than the pre-determined value (here, 5%).

Figure 7.8 shows the decision of sample size state (decrease, remain or increase) and the significance of trials at the new required n^* value.

For all designs, the total number of significant trials are again much lower than in previous scenarios. Again, the largest numbers stopping for futility (blue), or decreasing in



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure 7.7: Sample size zones from 50000 simulations when $\delta = 0$ and n=264, for two values of n_{max}: comparing three designs and four observed treatment effects

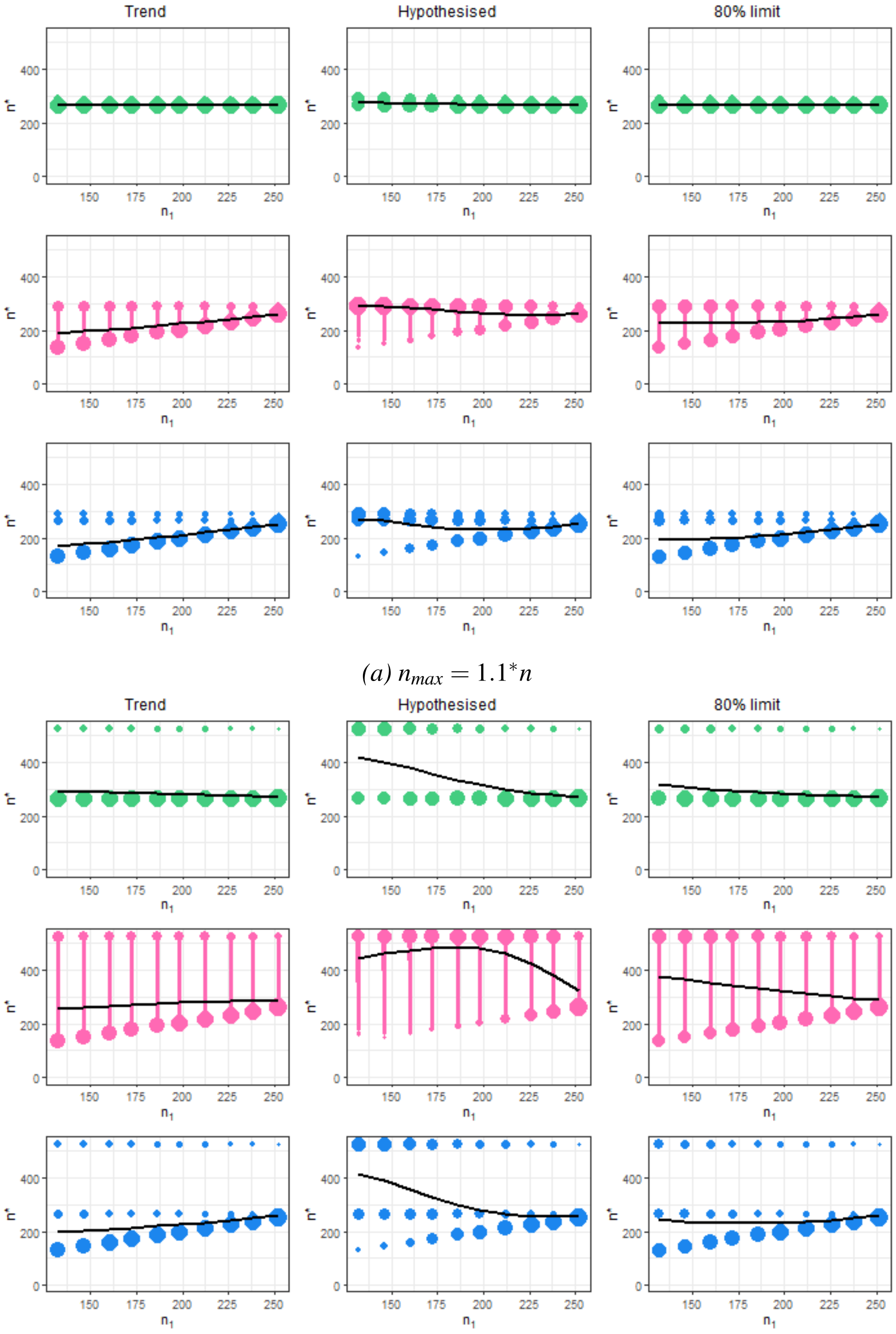
sample size (pink) is observed under the current trend assumption, followed by the 80% limit assumption, which catches up in terms of proportion by 90% information fraction. However, for the hypothesised assumption, a greater proportion of increases can be seen, at the 50 and 60% time points for both values of n_{max}. Again, very few decreases can be seen when using the hypothesised assumption at 50% interim timing.

The promising zone sees equal proportions of remaining and increasing using the hypothesised effect for $n_{max} = 1.1^{*}n$ at 50% interim timing, but more increases can be seen at the same time point for $n_{max} = 2^{*}n$. The design with the most increases is the combination test design, for all time points, and both values of n_{max}. Again, this figure does not show the magnitude of this increase.

Figure 7.8 shows expected sample size for two values of n_{max}, three designs and three assumptions. Again, the promising zone design for $n_{max} = 1.1^{*}n$ remains quite flat, and does not really shift much above the original n patients. A slightly raised line can be seen however for $n_{max} = 2^{*}n$ under the current trend and 80% limit assumptions, but more pronounced in the earlier time points with the 80% limit. A much increased expected sample size line is seen for the hypothesised effect, with a steady decrease up to around 80% through the trial.

Again, ASN is increased for $n_{max} = 2^{*}n$ than $1.1^{*}n$ for the combination test design,but to a much greater extent than in the promising zone design. Whilst few maximum increases are observed using the current trend assumption, the expected sample size line is greater than the n_{rec} line, particularly at earlier time points. The hypothesised effect has greatly increased ASN, which increases up until around the 75% mark, before decreasing to 90% through the trial. Finally, the 80% limit assumption sees a similar pattern to the promising zone design, with a more pronounced slope particularly at 50% data availability.

Promising zone design with an incorporated futility boundary sees the lowest expected sample size lines of the three designs, which is lowest using the current trend assumption. Similarly to the other two designs, the hypothesised assumption sees a large ASN value between 50 and 80% trial duration, but flattens out after this point.



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure 7.8: Sample size from 50000 simulations when $\delta = 0$ *and n=264, for two values of* n_{max}*: comparing three designs and three treatment effect assumptions*

Table 7.9 presents ASN and power values for the four investigated values of γ for short and medium endpoints. When investigated, it was found to be due to small numbers of pipeline patients, often leading to very small sample sizes for the second stage test statistic and therefore a very unstable value. It is for this reason that Type I error is inflated in this table, but comparatively, smaller values will indicate a better design.

δ=0.4		**SHORT**					**MEDIUM**				
δ=0		**Information fraction**					**Information fraction**				
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	273	282	287	291	291	303	315	322	325	306
	Power	7.69	7.38	6.88	5.76	5.43	5.13	5.11	5.05	5.09	5.15
γ_2	ASN	256	267	275	281	285	289	302	312	317	300
	Power	7.91	7.55	6.99	5.82	5.42	5.21	5.14	5.08	5.13	5.14
γ_3	ASN	229	243	255	267	276	265	281	295	304	292
	Power	8.23	7.76	7.16	5.84	5.46	5.26	5.15	5.12	5.10	5.15
γ_4	ASN	199	219	236	253	268	237	259	277	292	284
	Power	8.42	7.95	7.25	5.91	5.54	5.24	5.12	5.10	5.14	5.20
HYPOTHESISED											
γ_1	ASN	465	488	499	478	394	465	488	500	485	403
	Power	5.08	5.06	5.17	5.21	5.35	5.08	5.06	5.18	5.11	5.15
γ_2	ASN	445	475	488	464	380	445	475	490	472	390
	Power	5.05	5.11	5.14	5.26	5.36	5.05	5.13	5.11	5.11	5.16
γ_3	ASN	411	450	464	435	357	411	451	468	448	369
	Power	5.18	5.05	5.15	5.45	5.36	5.18	5.09	5.07	5.18	5.14
γ_4	ASN	377	420	432	400	336	378	423	441	419	348
	Power	5.08	5.11	5.41	5.56	5.40	5.08	5.14	5.18	5.22	5.17
80% Limit											
γ_1	ASN	401	378	353	329	305	418	400	380	359	319
	Power	6.28	6.48	6.49	5.70	5.41	5.04	5.09	5.05	5.12	5.17
γ_2	ASN	376	355	334	314	297	396	380	364	347	312
	Power	6.54	6.70	6.59	5.68	5.41	5.09	5.09	5.05	5.08	5.15
γ_3	ASN	329	315	301	292	285	354	345	336	327	301
	Power	7.11	7.04	6.85	5.70	5.40	5.20	5.08	5.13	5.06	5.13
γ_4	ASN	274	271	269	271	275	305	306	308	309	291
	Power	7.62	7.41	6.98	5.80	5.47	5.23	5.13	5.11	5.11	5.16

*Table 7.9: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when δ=0, n=264, n_{max}=2*n*

Short endpoints with current trend or 80% limits should be avoided, as the inflation in Type I error is considered too high. The hypothesised limit also sees an inflation, but is much lower than the other two assumptions. For medium endpoints, all assumptions see a small inflation in sample size, but are comparatively minimal. The lowest values of Type I error occur for medium endpoints, γ_1=0.0001 using the 80% limit assumption, with the smallest values happening earlier in study duration.

Under the circumstances, it would not be recommended to use the combination test design for a short endpoint as the level of inflation of the Type I error is too high when there are too few patients in stage 2 to be able to split the test statistics according to the pre-specified weights of the combination test design. However, if one were able to incorporate a value of $n_{min} > n_{rec}$, such that the second stage sample size has at least X patients, Type I error may be able to be controlled. This would increase ASN, but without this restriction, the design is not feasible to be practically implemented.

With medium endpoints, the Type I inflation is smaller. This shows the plausibility of the $n_{min} > n_{rec}$ restriction suggestion, as medium endpoints have been designed to have a greater number of pipeline, and therefore n_{rec} patients.

Whilst the smaller values of γ largely correspond to smaller power, they also correspond to greater values of ASN which may be undesirable when $\delta = 0$. Additionally, looking back at Table 7.3, it also corresponds to power values much larger than 90% when $\delta = \delta_{plan}$. This will be discussed further in Section 7.5.

Table 7.10 shows ASN and power for the promising zone design, with and without futility boundaries.

The promising zone design controls Type I error for the current trend assumption at all time points, for hypothesised effect assumption at 90-95%, and for the 80% limit assumption from 55% onwards . Type I error has decreased with the incorporation of a futility boundary and is below the nominal 5% rate at all time points for the current trend and 80% limit assumptions. However, the hypothesised limit has now increased power between 50 and 60% data availability for the hypothesised assumption, before decreasing from this point

δ=0		Information fraction									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	294	292	289	287	285	282	279	276	274	271
	Power	4.96	4.90	4.82	4.83	4.74	4.74	4.72	4.65	4.63	4.62
Hypothesised	ASN	420	403	380	359	335	318	301	288	280	272
	Power	6.06	5.93	5.78	5.66	5.51	5.42	5.24	5.10	4.94	4.82
80% limit	ASN	318	309	302	297	291	286	282	278	275	271
	Power	5.14	4.98	4.89	4.85	4.73	4.72	4.67	4.63	4.59	4.60
Promising zone with Futility											
Trend	ASN	198	202	208	213	220	226	234	242	250	260
	Power	4.60	4.66	4.57	4.62	4.54	4.56	4.57	4.52	4.55	4.57
Hypothesised	ASN	415	391	359	330	299	280	264	257	257	261
	Power	7.40	6.52	5.85	5.61	5.44	5.35	5.18	5.03	4.91	4.80
80% limit	ASN	245	237	232	231	231	233	238	244	251	260
	Power	4.98	4.87	4.74	4.73	4.60	4.59	4.57	4.51	4.52	4.55

Table 7.10: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions and three designs when δ=0, n=264 and and $n_{max} = 2^{*}n$

onwards. However, it still only falls below the 5% rate from 90% onwards.

From power alone, the hypothesised assumption is not appropriate to use as it has inflated Type I error. A modification perhaps to the promising zone region may be able to reduce this level to fall below 5%. However, the design as it currently stands does not work well for the hypothesised assumption.

ASN has been reduced for current trend and 80% limit assumptions by the addition of a futility boundary, as well as decrease power. Looking back to Table 6.1 for evaluating the methodologies, the most fitting design would be promising zone with futility using the current trend assumption, due to the lowest values of ASN, and no inflation of Type I error. However, it has also been shown that 80% limit works well when $\delta = 0$, with no inflation of Type I error (with one exception at 50% information fraction with no futility boundary).

7.5 Discussion

The results of the three designs and three assumptions under four potential trial outcomes ($\delta = \delta_{plan}$, $\frac{2}{3}\delta_{plan}$, $\frac{1}{3}\delta_{plan}$ and 0) have been presented in this chapter, and can be used to answer the specific aims specified at the beginning of this chapter.

The first aim was to appropriately simulate data to be able to suitably generate 50000 trials worth of data. Characteristics from the sampling at every 5% through the trial can be seen in 7.1. For an effect size of 0.4, a mean difference and SD were chosen as 8 and 20 respectively. Therefore, when as planned, the mean $\hat{d}$ should be as close as 8, and $\hat{\sigma}$ as close to 20 as possible. By 100% through the trial, $d - \hat{d}$ is 0.0009, and so can be considered close enough to the true value 8. Similarly, d is multiplied by $\frac{2}{3}$, $\frac{1}{3}$ and 0 respectively. SD is slightly underestimated at all time points, but this does not seem to impact the treatment effect, which is correct to 2 decimal places even at 50% through the trial.

One limitation in the simulations is the lack of bias incorporated to the simulated data, which we have seen in the real-world trial data. Any differences in data re-analysis and simulation results could be down to this. However, trial data was found to be within 1*SE by 57% through the trial, and only interim analyses past 50% are looked at for this reason. On the other hand, the simulations have an advantage over the data re-analysis, in that the true effect is known, not assumed to be the originally observed effect, which could also explain any discrepancies in results.

Conditional power values were calculated using three treatment effect assumptions: trend, hypothesised and 80% limit, seen in Table 7.2. For all true treatment effect scenarios considered, trend sees the lowest mean CP values, hypothesised sees the highest, and the 80% limit is always between these two values. It should be noted that a higher CP are desired when the effect is as planned or high, and lower values when there is little or no treatment effect. When the observed effect is as planned, mean CP>92%, >75% for $\frac{2}{3}\delta_{plan}$, >48% for $\frac{1}{3}\delta_{plan}$, and >17% for zero effect. Even when no effect is seen, the mean CP is above the 10% considered futility boundary for all assumptions.

Four values of γ, the pre-specified parameter required for the combination test design, were considered, and short and medium endpoints were assessed, differing in increasing levels of pipeline patients. Using information from the recruitment table (Table 4.6), pipeline patients were approximately 1-3% for the short endpoint, and 20-22% for the medium outcome. The medium endpoint had higher values of ASN, but did not always correspond to higher values of power.

Additionally, the smaller the value of γ, the higher the ASN and power. This corresponds to the current literature, stating higher values of γ penalise higher sample sizes (Pilz 2019). For direct comparisons between designs, γ_2=0.0002 with a short endpoint was compared for graphical comparisons, chosen prior to seeing the results. However, for different assumptions, alternative values of γ may be of more value to compare. The results have however been presented in tables, and so direct comparisons can still be drawn.

When the observed effect is as planned, both promising zone designs work well in that they maintain power at any interim time point. The trend assumption stops a few more trials early for futility, but results in a lower ASN whilst still maintaining power at the pre-specified limit of 90%. The hypothesised assumption sees the lowest ASN, having seen fewer increases in sample size, but maintains power. The trend assumption with a promising zone design with futility boundary is the best design here in terms of not excessively increasing power, or alternatively the hypothesised assumption with or without futility in terms of lowest ASN. If the value of γ was chosen purely based on power values for designs, $\gamma_3 = 0.0005$ would be most appropriate for the trend and hypothesised assumptions, and γ_4 for the 80% limit assumption. With these values, the combination test design would be able to decrease sample size, whilst maintaining power above the nominal level, without excessively raising power.

Seeing a lower effect than planned but not as high as would have liked ($\delta = \frac{2}{3}\delta_{plan}$), the promising zone designs behave similarly at all time points, but more increases in sample size can be seen. Of the increases, more go on to be significant when $n_{max} = 2^{*}n$ compared to $1.1^{*}n$, which is seen in all assumptions. This indicates such a small allowed sample size increase is too small to fully benefit from the increase. More trials are stopping for futility when allowed, with the most happening at 90% under current trend and 80% limit assumptions. Depending on the value of γ considered for the combination test design, some lower ASN values can be observed, but can also drop in power. As power isn't overly important here, this design may still be of interest looking at this scenario only.

When a much smaller effect is seen ($\delta = \frac{1}{3}\delta_{plan}$), the promising zone design does not work so well. Being unable to decrease in sample size, a large number of studies remain at the original planned n, and few are significant. The addition of a futility boundary however is able to reduce the ASN, whilst showing similar power values to promising zone. It is worthwhile to mention that the 10% boundary used in this thesis can be considered a conservative boundary, and alternative boundaries such as 20% may further reduce the ASN. However, further research would be needed to see the full extent of design operating characteristics and therefore recommend the use, which is one limitation. The combination test using the current trend assumption sees the lowest values of ASN and may also be an appropriate design for consideration looking at this scenario only.

Finally, when the observed effect is zero, the combination test with a short outcome has been shown to inflate Type I error due to the small numbers involved in the calculation of the second stage statistic and therefore high variability. This is an important finding however, as it highlights a flaw in design when sample size is small, or recruitment is particularly slow, resulting in few pipeline patients. With the addition of a minimum pipeline patient number, this may be able to stabilise the second stage test statistic more and therefore not lead to an increase in Type I error. Investigation is beyond the scope of this thesis. However, the minimum number of pipeline principle has inadvertently been shown by the investigation of a medium endpoint, which has been designed to have a higher number of pipeline patients. As Type I error is still above the 5% level, it has much decreased from its short endpoint counterpart, and shows the direction it could go with even more pipeline patients (>25% of the original n patients), or setting a minimum number to still be recruited, if the required level of pipeline patients has not yet been met.

The promising zone design with the current trend, and the 80% limit (but only beyond 55% through) have been shown to not inflate Type I error. The addition of a futility boundary further decreases Type I error, and either assumption can be used at any time point. Furthermore, a futility boundary incorporation reduces the level of ASN, which would be beneficial in the scenario of $\delta = 0$.

Taking all investigated scenarios into account, the simulations suggest that the promising zone with futility boundary, using the current trend assumption is a good design to use. When the observed effect is as planned, it has a relatively low ASN than the corresponding promising zone design, whilst also not excessively increasing power. Additionally, Type

I error is well controlled in the case of $\delta = 0$, and sees the greatest proportion of trials stopping for futility, even at 50% through the trial.

Whilst the hypothesised effect assumption was seen to be better in terms of a lower ASN when $\delta = \delta_{plan}$, or $\delta = \frac{2}{3}\delta_{plan}$, the inflation of Type I error at $\delta = 0$ was too high before 90% through the trial, and would need modification to the design to get this level under control. This again is beyond the scope of the thesis, and could be recommended for further work.

It should be noted that where Type I error is inflated, it is possible to adjust the critical value at the end of the study in order to decrease this value to the nominal rate, if a researcher wished to keep the designs as they are but ensure Type I error is no more than the nominal rate.

7.6 Summary

This chapter has provided the simulation plan and the results from the continuous simulation study of this thesis. Results have been discussed in depth and some recommendations made looking across all four scenarios of investigated δ compared to δ_{plan}. The promising zone with futility has been shown to work well across the scenarios in terms of power and ASN, and can be used at any interim time point, from the 50% investigated from.

An important finding relating to the combination test design has been identified, where the second stage is too small and therefore suffers from huge variability when splitting the stages up. A suggestion of a minimum sample size greater than the recruited number of patients may overcome this problem and has been recommended for further work. As far as I am aware, there is no other research published that indicates that this has been considered previously, and perhaps focus on much larger studies, or longer time points.

The hypothesised effect has also been shown to inflate Type I error and, in my opinion, should not be used in conjunction with the promising zone in its current state. Whilst scenarios may exist where the hypothesised effect may be beneficial, such as confirmatory settings where one may be confident in getting the planned effect, the design choice may be contraindicated in these settings. This is further discussed in Section 9.6. As discussed in the literature review in Chapter 3, there has previously been debate over the choice of future treatment effect, particularly between current trend and the hypothesised effect assumption. This chapter has hopefully provided some evidence to support the argument for the current trend assumption. However, it should be noted that this has only been investigated where no bias has been implemented, and further simulation work should be implemented before making a concrete recommendation.

The next chapter presents a simulation method using binary data rather than continuous, and presents results in terms of ASN and power for the same three designs, three assumptions, and two values of n_{max}.

8 | Simulations for binary outcomes

8.1 Introduction

Chapter 7 presented the simulation plan focussing on continuous outcomes only, and presented results for the continuous outcome simulations. Whilst promising zone with futility was shown to reduce ASN, maintain power and not inflate Type I error rate for the current trend and 80% limit assumptions, this may not be true for binary outcomes. Additionally, the combination test was shown to inflate Type I error when the second stage sample size was very small, and this will again be investigated for a slightly larger sample size. This chapter will present methods for data generating mechanisms for binary outcomes, and provide results for this simulated data. Results will be discussed, and compared to the data re-analysis results in Chapter 6 for binary outcomes.

8.2 Data generating mechanisms

As discussed in Chapter 7, binary data is approximately normal if sample size is large enough, or expected event rates close to 50%. For this reason, event rates have been chosen below this rate, whilst aiming to keep a realistic treatment difference, and still keep the sample size relatively small. Therefore a control group event rate of 20%, and an intervention group of 10% have been chosen, corresponding to a sample size of 556 patients, from an odds ratio of 2.25, two-sided significance level of 5%, and power of 90%.

For binary outcomes, data will be sampled from two uniform distributions. The random number generated between 0 and 1 was then categorised as either a 0 or a 1 depending whether it was less than or equal to the specific event rate (0.1 or 0.2 accordingly), or greater than this value.

Similarly to the simulation plan for continuous outcomes, seed numbers were again chosen as 123 and 1234, and three times the required number of random numbers were generated, corresponding to the maximum allowed sample size investigated. Outcomes from 1 to n_{group} in each sampled distribution correspond to the original planned fixed sample size trial. Repetitions were again set at 50,000, which is sufficient in terms of precision of conditional power values and therefore interim decision making.

All outcomes investigated will be the same as the continuous simulations: focusing on power and ASN for three designs, three assumptions, two endpoints, four values of γ, four values of n_{max} (two in the appendix) and interim time points between 50 and 95% through the original trial duration in terms of data availability.

8.3 Simulation results: Binary outcomes

8.3.1 True odds ratio = 2.25

	Information fraction									
$\hat{\delta}_{obs} = \delta_{plan}7$	**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
$e^{Mean(log(\hat{OR}))}$	2.3036	2.2972	2.2920	2.2882	2.2844	2.2818	2.2804	2.2788	2.2780	2.2776
Mean SE	0.3615	0.3438	0.3283	0.3149	0.3030	0.2923	0.2828	0.2741	0.2661	0.2589
Mean $\hat{\delta}_{obs}$	0.2730	0.2731	0.2732	0.2733	0.2734	0.2735	0.2737	0.2738	0.2740	0.2743
Mean $(log(OR) - log(\hat{OR}))$	0.0236	0.0207	0.0185	0.0168	0.0152	0.0140	0.0134	0.0127	0.0124	0.0122
$\hat{\delta}_{obs} = \frac{2}{3}\delta_{plan}$										
$e^{Mean(log(\hat{OR}))}$	1.6175	1.6172	1.6167	1.6164	1.6192	1.6212	1.6236	1.6274	1.6322	1.6377
Mean SE	0.3332	0.3170	0.3030	0.2907	0.2798	0.2700	0.2612	0.2533	0.246	0.2393
Mean $\hat{\delta}_{obs}$	0.1784	0.1783	0.1783	0.1783	0.1783	0.1782	0.1783	0.1784	0.1786	0.1787
Mean $(log(OR) - log(\hat{OR}))$	-0.3091	-0.3109	-0.3121	-0.313	-0.3141	-0.3148	-0.3151	-0.3155	-0.3154	-0.3155
$\hat{\delta}_{obs} = \frac{1}{3}\delta_{plan}$										
$e^{Mean(log(\hat{OR}))}$	1.2286	1.2332	1.2345	1.2368	1.2389	1.2398	1.2389	1.2398	1.2395	1.2408
Mean SE	0.3151	0.2999	0.2868	0.2752	0.2649	0.2557	0.2474	0.2399	0.2330	0.2267
Mean $\hat{\delta}_{obs}$	0.0856	0.0854	0.0854	0.0854	0.0852	0.0852	0.0852	0.0852	0.0853	0.0855
Mean $(log(OR) - log(\hat{OR}))$	-0.5837	-0.5848	-0.5851	-0.5855	-0.5864	-0.5867	-0.5867	-0.5869	-0.5868	-0.5865
$\hat{\delta}_{obs} = 0$										
$e^{Mean(log(\hat{OR}))}$	0.9268	0.9356	0.9465	0.9554	0.9635	0.9719	0.9790	0.9850	0.9905	0.9952
Mean SE	0.3033	0.2888	0.2761	0.2651	0.2552	0.2464	0.2384	0.2312	0.2245	0.2184
Mean $\hat{\delta}_{obs}$	0.0008	0.0006	0.0006	0.0005	0.0003	0.0003	0.0003	0.0003	0.0004	0.0005
Mean $(log(OR) - log(\hat{OR}))$	-0.8089	-0.8096	-0.8095	-0.8097	-0.8103	-0.8103	-0.8101	-0.8102	-0.8099	-0.8097

Table 8.1: Mean OR, SE, observed treatment effect and difference from the true population value for values between 50% and 95% of the originally planned n=556 when $\delta = \delta_{plan}$

Table 8.1 shows the mean $\hat{\delta}_{obs}$, SE, coefficient transformed onto the exponential scale, and difference between coefficients $log(OR)$ and $log(\hat{OR})$ calculated after every 5% of data available from 50% to 95% through the trial duration (data available from the original n patients). According to the simulation plan, the planned odds ratio is 2.25, with SE=0.25, resulting in a standardised treatment effect δ=0.276.

By 95% through the trial, the odds ratio is 2.28, with SE of 0.259. Repeating the sampling three times with different seed numbers, the mean coefficient (i.e. log(OR)) was the same to one decimal place. Whilst more repetitions may increase this accuracy, due to limitations in access to higher computational power, this will be considered sufficient for the purposes of an illustrative binary outcome example.

The corresponding summary values are also presented when the observed value of δ is $\frac{2}{3}$, $\frac{1}{3}$, and 0 times the planned value of δ in terms of absolute risk difference.

		Information fraction								
Conditional power		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**
δ=δ_{plan}	Trend	0.733	0.752	0.772	0.791	0.810	0.829	0.848	0.866	0.885
	Hypothesised	0.908	0.897	0.889	0.885	0.884	0.885	0.889	0.894	0.901
	80% limit	0.902	0.899	0.898	0.897	0.897	0.899	0.901	0.904	0.908
δ=$\frac{2}{3}\delta_{plan}$	Trend	0.458	0.464	0.473	0.481	0.491	0.501	0.512	0.524	0.537
	Hypothesised	0.768	0.727	0.692	0.662	0.637	0.615	0.599	0.586	0.576
	80% limit	0.710	0.691	0.673	0.657	0.641	0.626	0.612	0.599	0.587
δ=$\frac{1}{3}\delta_{plan}$	Trend	0.215	0.209	0.205	0.200	0.194	0.190	0.186	0.182	0.178
	Hypothesised	0.568	0.495	0.433	0.378	0.331	0.292	0.259	0.230	0.207
	80% limit	0.448	0.411	0.378	0.346	0.315	0.287	0.261	0.236	0.213
δ=0	Trend	0.080	0.072	0.066	0.059	0.053	0.047	0.042	0.037	0.032
	Hypothesised	0.370	0.286	0.219	0.167	0.126	0.096	0.073	0.055	0.041
	80% limit	0.228	0.191	0.160	0.132	0.108	0.088	0.071	0.055	0.043

Table 8.2: Mean conditional power values from 50000 repetitions for three treatment effect assumptions when $\delta = \delta_{plan}$ and n=556

Mean CP values from 50,000 simulations have been calculated using the current trend, hypothesised and 80% optimistic confidence limit from 50% through the trial (i.e. 50% of the originally planned patients with data available), and are presented in Table 8.2.

The hypothesised assumption has the highest observed CP values at the 50% data available time point. However, the 80% limit overtakes the hypothesised assumption in terms of CP; by 55% through when $\delta = \delta_{plan}$, and by 85% through when $\delta = 0$.

The current trend assumption has consistently lower values of CP. In the case of $\delta = 0$, mean CP never goes above 0.08, compared to values of 0.37 and 0.23 for the hypothesised and 80% limit assumptions respectively, which decrease with increasing information fraction. Comparatively, both hypothesised and 80% limit assumptions remain above 0.88 when $\delta = \delta_{plan}$ at all time points. However, CP using the current trend assumption starts at 0.733, and does not reach above 0.88 until 90% trial duration.

8.3.1.1 True effect = planned effect

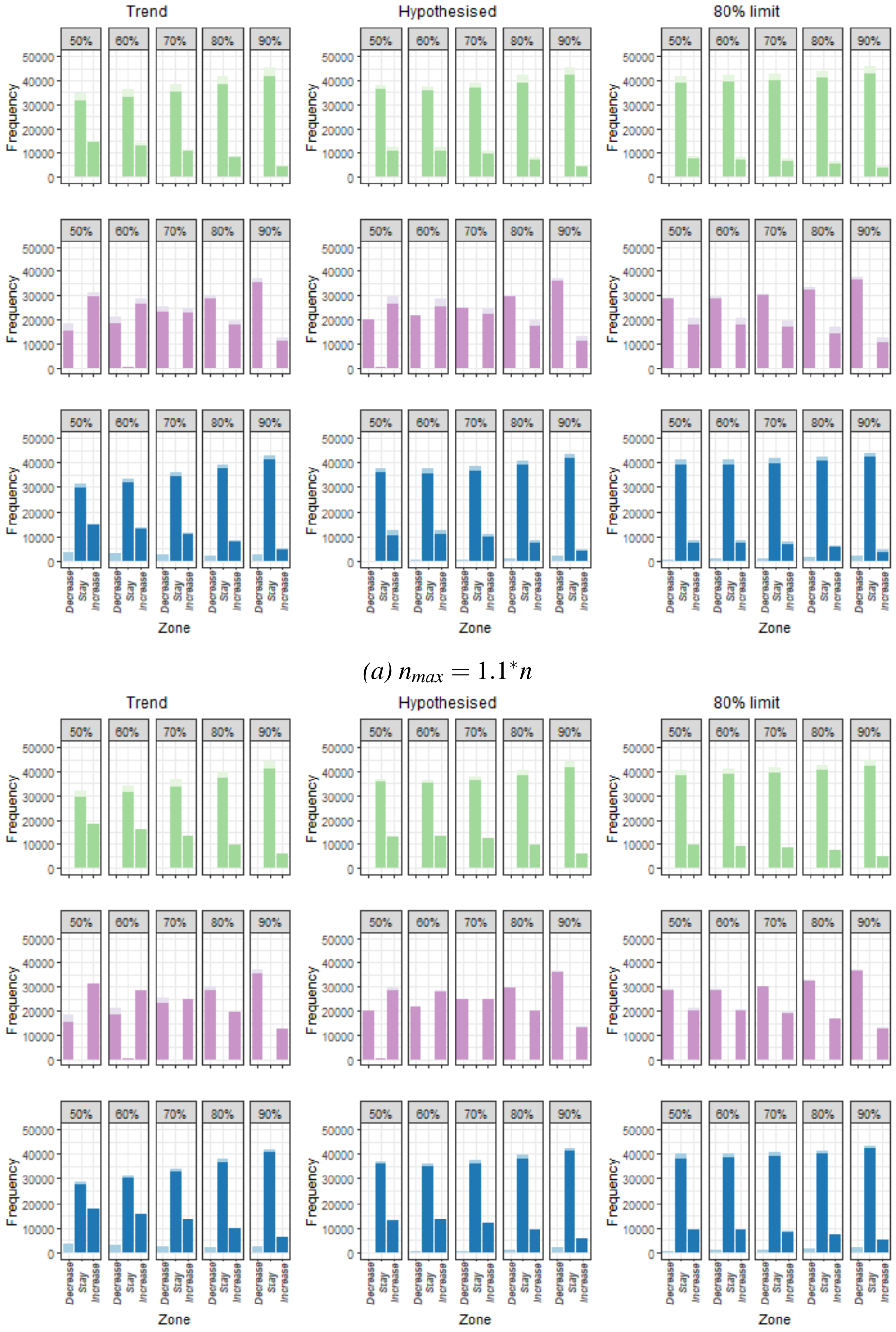
Figure 8.1 shows the sample size states: decrease, remain or increase the sample size in relation to the original n patients required from 50000 repetitions. Additionally, it shows whether or not the trial goes on to be significant, having data available from the new sample size n^* patients, with darker shades representing statistical significance, and lighter shades representing no statistical significance.

The current trend assumption sees the greatest proportion of increases in sample size across all three designs. Additionally, it also sees the greatest proportion of stopping for futility, which decreases with increasing n_1. The proportion of increasing sample size is greatest in the combination test design when γ_2=0.0002. Additionally this design also sees the greatest number of decreases in sample size, which also increases with increasing information fraction. The same pattern for $n_{max} = 1.1^*n$ is observed as for $n_{max} = 2^*n$.

There is a greater difference between 50% and 90% in terms of remaining at the same level of n patients in the current trend assumption for both promising zone designs. This level is much more even between 50% and 90% for the hypothesised and 80% limit assumptions, seeing very small increases, as well as a small increase in the number of trials stopping for futility with increasing values of n_1.

Figure 8.2 shows n^* plotted against n_1 values every 5% of data availability from 50% onwards. Again, colours have been kept consistent: green for promising zone design, pink for the combination test design, and blue for the promising zone design with a futility boundary.

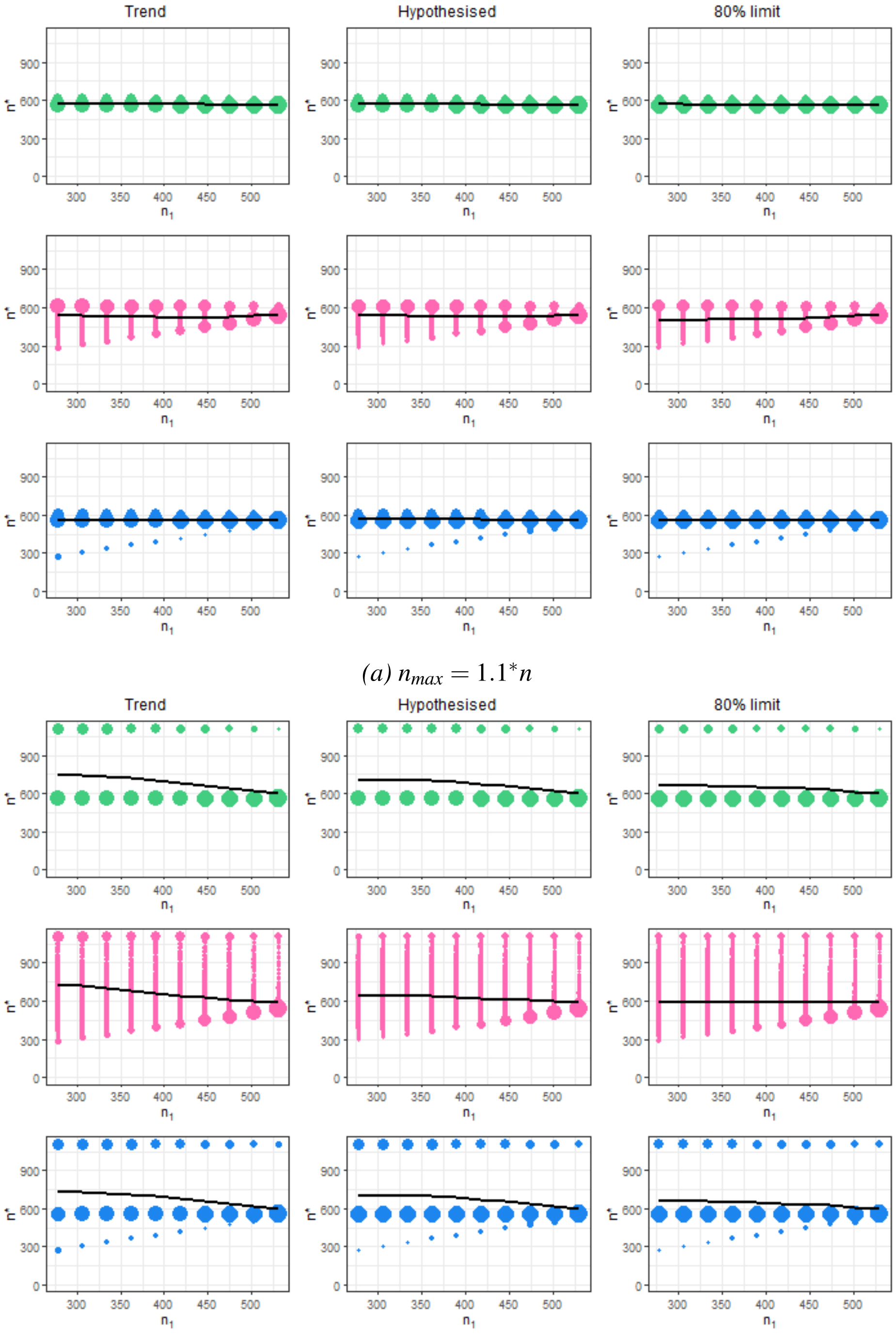
When $n_{max} = 1.1^*n$, there is very little difference in expected sample size for either



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure 8.1: Sample size zones from 50000 simulations when $\delta = \delta_{plan}$ and n=556, for two values of n_{max}: comparing three designs and four observed treatment effects



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure 8.2: Sample size from 50000 simulations when $\delta = \delta_{plan}$ *and n=556, for two values of* n_{max}*: comparing three designs and three treatment effect assumptions*

promising zone design. Very few decreases are observed in the design with futility, and for both designs, increases are so small that the expected sample size line appears virtually flat across values of n_1. Increasing the allowed maximum to twice the original sample size sees an increase in ASN for both designs; more so at the 50% interim timing, before steadily decreasing with increasing n_1 for all three designs.

The expected sample size line for the combination test design for γ_2=0.0002 is slightly below the original n patients when $n_{max} = 1.1^*n$, marginally increasing with increasing n_1. This is not the case however when $n_{max} = 2^*n$ - similarly to the promising zone designs, the expected sample size line is above the original n patients. The current trend sees the highest of these lines, and the 80% limit the lowest ASN for all three designs, and both values of n_{max}.

Table 8.7 provides ASN and power for the four investigated values of γ used in the combination test design. The same values of 0.0001, 0.0002, 0.0005 and 0.001 have been used, as in the simulations for continuous outcomes seen previously in Chapter 7.

Power values are above 90% for all values of γ for the hypothesised and 80% limit assumptions. However, with the current trend assumption, γ_3 sees a drop in power at 50% information fraction, and γ_4 at 50-80% trial duration - observed at both short and medium endpoints. Whilst power has dropped, so has the ASN, seeing as low as 394 patients at the 50% time point with a short endpoint and $\gamma_4 = 0.001$.

Similarly to the continuous simulations, medium endpoints see a larger expected sample size, and almost always larger power (but not always). Additionally, smaller values of γ again increases the ASN and the corresponding power.

The hypothesised assumption sees the largest values of power, despite lower ASN than the current trend assumption comparatively. The maximum observed power is 99.46% with 683 patients, which is much higher than the nominal 90% rate specified. At the largest value of γ investigated, this drops to as low as 93.16%, which is still above the required value of power. The 80% limit assumption sees the closest power to the original 90%, with a value of 90.13% and 90.23% at 50% interim timing and $\gamma_4 = 0.001$ for short and medium endpoints respectively. Additionally, at this value of γ, a sample size saving is observed, with an ASN

		SHORT					MEDIUM				
δ=δ_{plan}		**Information fraction**					**Information fraction**				
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	809	771	722	669	625	813	781	747	704	656
	Power	95.27	96.29	96.98	97.26	97.13	95.37	96.41	97.17	97.37	97.16
γ_2	ASN	730	700	663	627	601	736	715	696	668	635
	Power	93.03	94.66	95.76	96.31	96.41	93.31	94.97	96.14	96.51	96.47
γ_3	ASN	536	537	537	540	553	549	563	583	590	590
	Power	87.21	90.10	92.17	93.73	94.52	88.21	91.21	93.38	94.32	94.77
γ_4	ASN	394	423	452	484	523	417	463	514	543	562
	Power	78.18	83.07	86.64	89.89	91.91	78.18	83.07	86.64	89.89	91.91
HYPOTHESISED											
γ_1	ASN	709	708	683	648	618	709	714	705	682	650
	Power	99.06	99.36	99.46	99.38	98.89	99.06	99.35	99.46	99.38	98.91
γ_2	ASN	638	645	632	613	600	639	654	661	652	633
	Power	98.44	99.00	99.20	99.11	98.56	98.44	99.00	99.22	99.10	98.58
γ_3	ASN	538	554	559	562	572	542	572	601	610	608
	Power	96.54	97.74	98.22	98.17	97.60	96.54	97.76	98.32	98.25	97.64
γ_4	ASN	458	476	492	513	541	467	507	548	568	579
	Power	93.16	94.55	95.31	95.66	95.61	93.23	94.80	95.55	95.79	95.68
80% Limit											
γ_1	ASN	633	634	626	614	604	636	646	657	652	636
	Power	98.24	98.60	98.81	98.84	98.52	98.25	98.60	98.91	98.84	98.54
γ_2	ASN	585	591	590	588	589	591	608	628	630	623
	Power	97.52	98.08	98.29	98.44	98.16	97.57	98.07	98.38	98.45	98.17
γ_3	ASN	510	523	535	548	566	520	550	584	598	602
	Power	95.45	96.23	96.89	97.3	97.18	95.47	96.22	97.04	97.34	97.22
γ_4	ASN	426	452	478	506	540	443	490	538	562	578
	Power	90.13	92.39	93.96	94.86	95.51	90.23	92.48	94.18	94.98	95.51

Table 8.3: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when δ=δ_{plan}, n=556, $n_{max} = 2^{*}n$

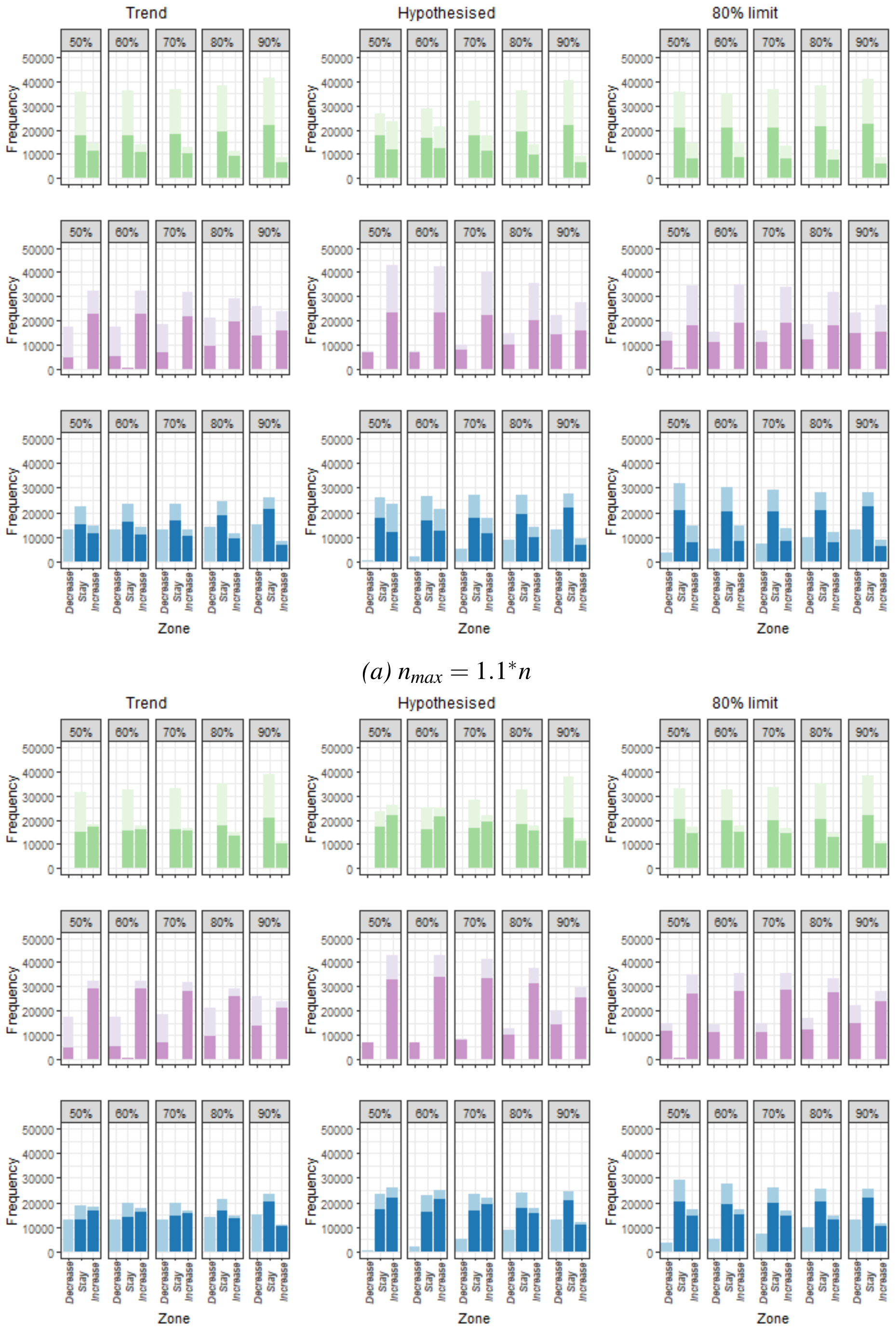
of 426 and 443 patients for short and medium endpoints respectively.

δ=δ_{plan}		**Information fraction**									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	756	746	733	720	705	686	667	646	623	600
	Power	94.09	94.08	94.00	94.07	94.11	94.03	93.93	93.86	93.75	93.44
Hypothesised	ASN	703	709	707	699	691	676	662	643	620	599
	Power	97.59	97.44	97.12	96.64	96.30	95.76	95.47	94.97	94.47	93.85
80% limit	ASN	662	661	659	657	651	645	638	628	614	596
	Power	95.75	95.66	95.63	95.6	95.49	95.24	95.16	94.86	94.42	93.89
Promising zone with Futility											
Trend	ASN	737	731	720	710	697	680	662	642	621	598
	Power	90.71	91.45	91.77	92.29	92.65	92.93	93.17	93.26	93.33	93.19
Hypothesised	ASN	703	709	706	698	690	674	660	641	618	598
	Power	97.59	97.42	97.07	96.57	96.21	95.63	95.31	94.82	94.33	93.69
80% limit	ASN	659	658	656	654	648	642	635	625	612	595
	Power	95.45	95.36	95.32	95.33	95.22	95.00	94.96	94.68	94.28	93.73

Table 8.4: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions and three designs when $\hat{\delta}_{obs} = \delta_{plan}$, n=556 and and $n_{max} = 2^*n$

Table 8.4 similarly shows ASN and power for the promising zone designs: with and without a 10% futility bound. All values of power are above the pre-determined 90% value, ranging from 90.71% (futility design, 50% information fraction and current trend assumption), to 97.59% (both designs, 50% information fraction and the hypothesised assumption). Despite allowing for a futility boundary, ASN always remains above the original planned n patients. This is a desirable characteristic however, as there is no wish to stop early for futility when $\delta = \delta_{plan}$. The later the interim analysis timing, the smaller the value of ASN, which is preferable.

Whilst the current trend assumption, 50% information fraction and futility design has the least excessive increase in power, either of the other designs offer a lower expected sample size, with greater power. In my opinion, the lower ASN criteria would be more favourable than not to excessively increase power, and therefore should be prioritised over the latter criteria where there is a clash in criteria according to Table 6.1. The 80% limit assumption would therefore be most preferable in either design due to the lowest ASN.



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure 8.3: Sample size zones from 50000 simulations when $\delta = \frac{2}{3}\delta_{plan}$ and n=556, for two values of n_{max}: comparing three designs and four observed treatment effects

8.3.1.2 True effect = $\frac{2}{3}$ planned effect

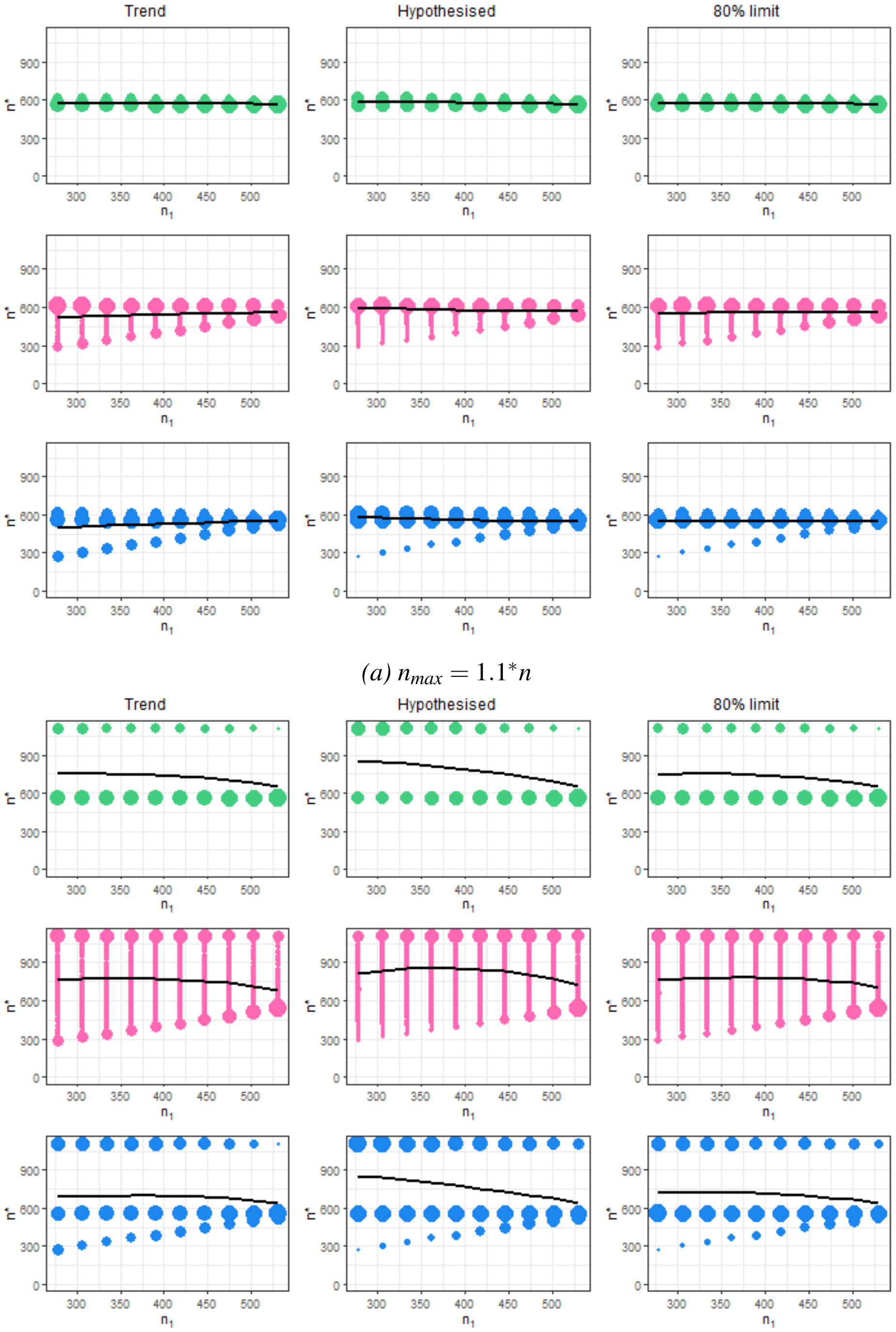
Figure 8.3 shows the proportion of repetitions that decreased, remained at or increased the originally planned sample size n, split by statistical significance at n^* patients with data available when $\delta = \frac{2}{3}\delta_{plan}$. Similarly to the continuous case, for the purposes of this thesis this could be thought of as a smaller than planned effect, but still potentially clinically significant. Therefore, there would be no wish to stop early, and an increase in power would be desirable. Whilst a smaller ASN is always preferable, the increase in power is the criteria to be prioritised in terms of Table 6.1.

The combination test design still has the most number of increases in sample size. However, these do not always result in a significant result, seeing the level of light shading in the bars of any design in the increase sample size stage. Both promising zone designs largely see a significant result for the trial that have increased in sample size. However, large proportions of non-significant trials can be seen when sample size remains using the promising zone design. The addition of a futility boundary is able to move some of these trials to the "decrease sample size" zone, but also moves some of the trials that turn out to be significant without a futility boundary.

The current trend assumption sees the largest proportion of trials with $n^* < n$ for the two designs that allow a decrease in sample size, at all interim timings. Whilst the hypothesised and 80% limit assumptions see nearly the same proportion of trials stopping early by 90% through, at 50% data available this number is much lower than their current trend counterparts.

Whilst there are a similar proportion of increases between the two values of n_{max}, when $n_{max} = 1.1^*n$ many of these increased trials do not become statistically significant at n^* recruited patients. This shows that while an increase in sample size is indicated, such a small sample size increase is insufficient to bring about the benefit from increasing sample size.

Figure 8.4 shows n^* vs n_1 and expected sample size lines for the three designs and three assumptions. Again, the promising zone design sees a flat line when $n_{max} = 1.1^*n$, very



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure 8.4: Sample size from 50000 simulations when $\delta = \frac{2}{3}\delta_{plan}$ *and n=556, for two values of* n_{max}*: comparing three designs and three treatment effect assumptions*

slightly above the original n patients across all three assumptions. However, for $n_{max} = 2^{*}n$, this line is understandably higher, with a number of increases up to the maximum allowed. For all three assumptions the expected sample size line decreases with increasing n_1 patients.

The promising zone design with futility takes a similar shape as the expected sample size line for the regular promising zone design, but decreased values due to the ability to stop early for futility. When $n_{max} = 1.1 * n$, this line even falls below the original n patients.

The combination test design sees a much broader range of sample size between n_{rec} and n_{max}. For the smallest value of n_{max}, the majority of trials have $n^* = n_{max}$, with the next greatest proportion being at $n^* = n_{rec}$, increasing as the number of recruited patients also increases. This pattern is seen across all three assumptions. For the larger value of n_{max} however is able to have a wider range of possible n^* values and therefore see slightly smaller proportions reaching the maximum allowed value, although this is still high.

In all three designs, the hypothesised effect assumption sees the greatest value of expected sample size at all time points. The current trend and 80% limit assumptions have similar values of expected sample sizes, and each take on the lowest value at some time point. This is shown in more detail later in Table 8.6.

Table 8.5 shows ASN and power values for short and medium endpoints for 4 values of γ. Expected sample size values range from 372 (current trend assumption, short endpoint, γ_4=0.001, 50% information fraction) to 917 (hypothesised assumption, medium endpoint, γ_1=0.0001, 70% information fraction). These values of ASN also give the lowest and highest values of power respectively: 40.5% and 84.1%.

Comparing the same γ values across the three assumptions, the hypothesised effect assumption consistently sees the highest values of both ASN and power for both endpoint timings. However, whilst the 80% limit assumption sees the lowest values of ASN, it is the current trend assumption that sees the lowest value of power.

Referring to the criteria for evaluating methodology in Table 6.1, the hypothesised assumption would be the least appropriate assumption to use. However, as power was higher than the planned 90% when $\delta = \frac{2}{3}\delta_{plan}$ even at the largest value of γ, realistically an alter-

δ=0.27		SHORT					MEDIUM				
δ=δ_{plan}		Information fraction					Information fraction				
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	846	853	846	816	762	861	872	868	839	782
	Power	71.14	72.92	74.00	74.01	72.59	71.22	73.12	74.27	74.21	72.67
γ_2	ASN	764	776	773	754	716	784	800	802	782	739
	Power	66.56	68.79	70.03	70.64	69.70	66.79	69.26	70.58	71.05	69.86
γ_3	ASN	526	555	578	595	606	555	590	621	634	633
	Power	52.99	56.03	58.97	61.54	63.03	53.95	57.49	60.72	62.72	63.53
γ_4	ASN	372	415	458	497	536	412	465	517	547	569
	Power	40.53	44.73	48.99	53.11	56.45	40.53	44.73	48.99	53.11	56.45
HYPOTHESISED											
γ_1	ASN	884	914	911	878	809	884	915	917	889	826
	Power	81.53	83.39	84.05	83.62	80.82	81.53	83.38	84.07	83.62	81.05
γ_2	ASN	812	852	856	831	769	812	854	866	846	789
	Power	78.65	81.5	82.64	82.37	78.83	78.65	81.5	82.77	82.37	79.12
γ_3	ASN	691	735	747	731	697	692	743	769	755	721
	Power	72.25	76.47	78.39	77.74	74.60	72.26	76.65	78.83	77.83	74.87
γ_4	ASN	555	574	584	590	594	565	599	619	627	623
	Power	62.16	65.13	66.70	67.16	65.94	62.54	65.99	66.97	67.39	66.14
80% Limit											
γ_1	ASN	823	833	831	817	772	825	840	850	833	790
	Power	79.64	80.53	81.02	81.03	78.57	79.63	80.54	81.57	81.01	78.70
γ_2	ASN	765	778	780	773	738	770	787	798	793	758
	Power	77.01	78.27	79.10	79.19	76.90	77.03	78.26	79.15	79.19	77.05
γ_3	ASN	642	666	679	686	676	658	682	708	715	700
	Power	69.89	72.22	73.45	74.13	72.84	70.40	72.19	73.63	74.13	72.95
γ_4	ASN	474	508	542	568	590	493	539	583	607	618
	Power	56.66	59.81	62.48	64.77	65.45	56.70	60.01	62.62	64.96	65.41

*Table 8.5: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when $\delta=\frac{2}{3}\delta_{plan}$, n=556, n_{max}=2*n*

native value would have been sought, which would likely have reduced these values.

In terms of power from Table 8.7 above, γ_2 or γ_3 may have been appropriate choices when using the current trend, whereas γ_4 may have been more appropriate for the 80% limit assumption. Whilst this means a more comparable value of ASN and power between the assumptions compared to direct values of γ, the 80% limit assumption still sees the lowest ASN, as well as having similar power, and therefore is the most preferable assumption when $\delta = \frac{2}{3}\delta_{plan}$.

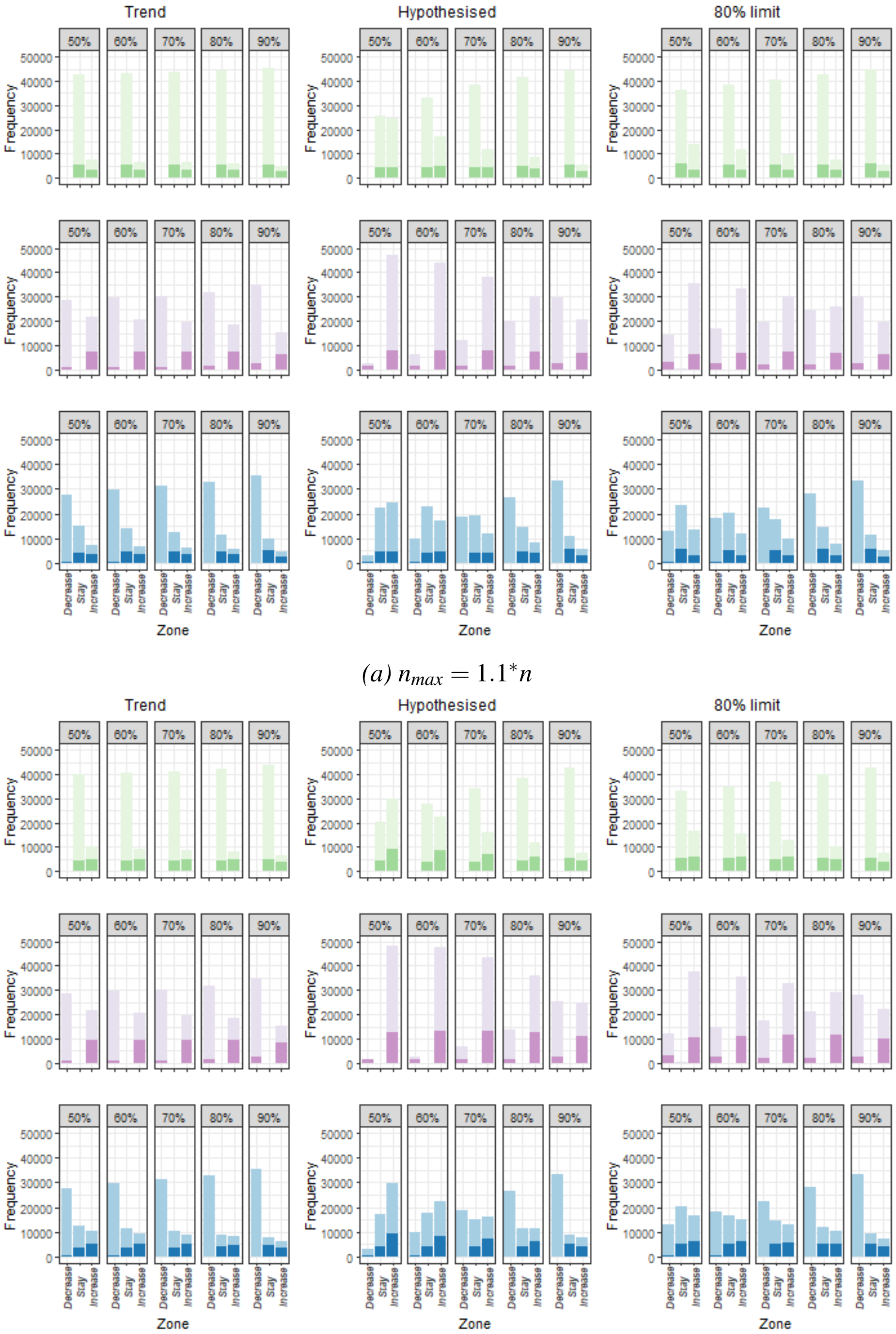
δ=$\frac{2}{3}\delta_{plan}$		**Information fraction**									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	761	757	750	747	743	733	720	705	679	650
	Power	64.03	63.76	63.34	63.1	62.99	62.73	62.42	62.04	61.58	60.60
Hypothesised	ASN	849	849	834	815	797	771	751	723	692	651
	Power	77.55	76.76	74.89	72.79	71.29	69.21	67.6	65.66	64.09	61.63
80% limit	ASN	747	753	750	749	741	732	719	702	683	648
	Power	69.57	69.68	69.45	69.17	68.67	67.89	66.84	65.61	64.19	61.9
Promising zone with Futility											
Trend	ASN	688	693	694	697	699	696	690	681	663	641
	Power	59.31	59.86	59.94	60.31	60.5	60.74	60.82	60.78	60.75	60.15
Hypothesised	ASN	846	843	825	802	781	753	731	705	677	643
	Power	77.57	76.74	74.83	72.7	71.13	69.01	67.29	65.33	63.79	61.36
80% limit	ASN	726	731	727	724	717	709	698	684	669	639
	Power	69.15	69.25	68.99	68.61	68.21	67.47	66.39	65.20	63.89	61.63

Table 8.6: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions and three designs when $\hat{\delta}_{obs} = \frac{2}{3}\delta_{plan}$, n=556 and $n_{max} = 2^{*}n$

Table 8.6 shows ASN and power for the promising zone designs. The addition of a futility boundary has again decreased ASN. This effect is greatest under the current trend, seeing a decrease of 73 patients at the 50% interim time point, but is very small under the hypothesised effect, ranging between 3 and 8 patients between 50 and 90% data availability.

By 95% through the trial, both designs and all three assumptions reach power levels of around 60-62%. The hypothesised assumption sees the highest power, almost reaching 78% at the 50% information fraction. However, the increase in power compared to the other two assumptions also has an associated higher ASN of 846 or 849 patients with and without a futility boundary respectively. The 80% limit assumption appears to be a good middle ground - not overly increasing ASN, but manages to increase power. The earlier the interim timing, the greater the power, but also the greater the ASN and so a trial may wish to weigh up these two aspects on a case-by-case basis.

8.3.1.3 True effect = $\frac{1}{3}$ planned effect



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure 8.5: Sample size zones from 50000 simulations when $\delta = \frac{1}{3}\delta_{plan}$ and n=556, for two values of n_{max}: comparing three designs and four observed treatment effects

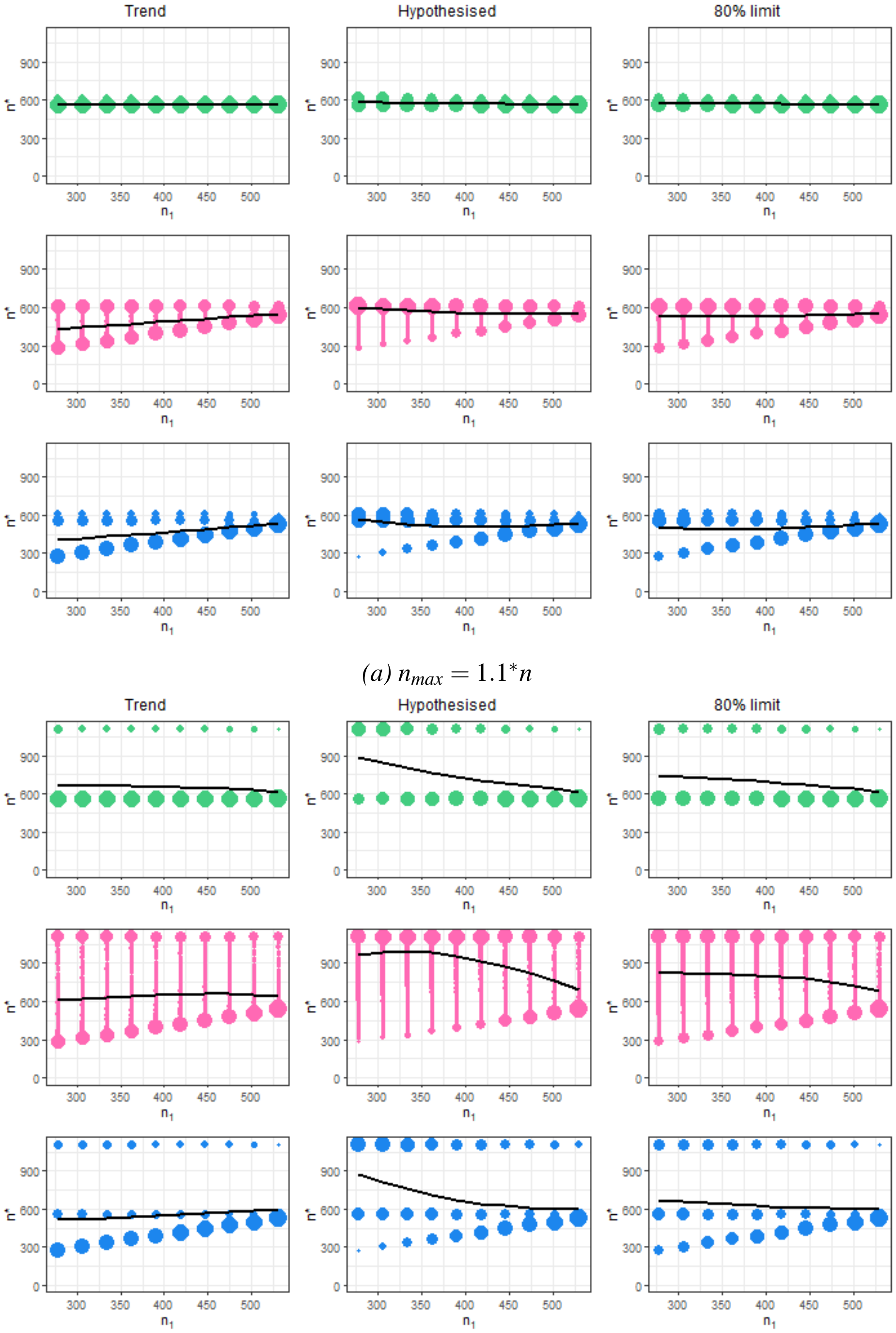
Figure 8.5 shows the sample size states (decrease/remain/increase) when $\delta = \frac{1}{3}\delta_{plan}$.

This scenario is considered both smaller than planned, and too small to have any potential to be clinically relevant. Therefore, a low ASN would be most desirable, whether below the original sample size (combination test or promising zone with futility) or as minimal as allowed (promising zone). Additionally, a comparatively high power is no longer of high importance when evaluating the methodologies, as the effect size is too small to be considered clinically meaningful. Therefore the prioritised measure for evaluation is the expected sample size.

Comparatively to when $\delta = \delta_{plan}$ or $\delta = \frac{2}{3}\delta_{plan}$, all graphs are now predominantly lightly shaded, indicating far more trials go on to be non-statistically significant, which is as expected. The current trend assumption sees a large number of decreases in sample size at all time points for the combination test and promising zone with futility designs, for both values of n_{max}. The hypothesised assumption actually sees very few decreases in sample size earlier in the trial, but does greatly increase by 90% through the trial, particularly around 70/80% data availability. The 80% limit assumption also sees a steep increase in sample size decrease, but takes on a higher proportion than the hypothesised effect even at 50% information fraction.

For the promising zone design, the current trend sees a huge proportion of trials remaining at $n^* = n$ patients, and very few increases at any time point investigated. On the other hand, the hypothesised assumption sees almost equal proportions of sample size remaining versus increasing when $n_{max} = 1.1^* n$ at the 50% time point. When $n_{max} = 2^* n$, the proportion of increases is even higher. Also for the hypothesised effect, the combination test sees almost all trials increase in sample size at 50/60% trial duration, but does decrease with increasing values of n_1.

Figure 8.6 shows n^* with progressing n_1, and plots the ASN line from 50000 repetitions. For all three designs, the current trend assumption sees the lowest ASN line; although all three assumptions have very similar values of ASN by 95% data availability. The 80% limit assumption sees relatively flat ASN lines when $n_{max} = 1.1^* n$ for all three designs but sees a slight downwards slope for the promising zone designs when $n_{max} = 2^* n$. A downwards



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure 8.6: Sample size from 50000 simulations when $\delta = \frac{1}{3}\delta_{plan}$ and n=556, for two values of n_{max}: comparing three designs and three treatment effect assumptions

slope is also seen in this case for the combination test design, but is a little steeper.

The hypothesised assumption consistently sees the highest ASN: marginal in the case of $n_{max} = 1.1^{*}n$, but much more so when $n_{max} = 2^{*}n$. Whilst still seeing some trials decrease in sample size, the majority of repetitions are seeing the maximum allowed sample size which, in the case of $n_{max} = 2^{*}n$, drastically brings up the expected sample size line. Between 65% and 95% data availability, this line starts decreasing as more values of $n^{*} = n_{rec}$ are seen.

The promising zone designs see very few trials with $n^{*} = n_{max}$ for the current trend and 80% limit assumptions. The hypothesised assumption however sees much larger proportions reaching this value between 50-60% trial duration.

Table 8.7 investigates the value of γ for the combination test design for the different assumptions when $n_{max} = 2^{*}n$. Sample size decreases can be seen for γ_4 for the 80% limit assumption, and γ_3 and γ_4 for the current trend assumption. No decreases in sample size are observed for any value of γ under the hypothesised effect assumption.

The greatest magnitude of expected sample size increase is seen for γ_1 under the hypothesised assumption, seeing as many as 1041 patients; an increase of 87% compared to the original n value. At this value of γ, almost 30% power is observed, which is similar to the power values using the 80% limit assumptions but with a lower ASN (by around 100 patients).

The lowest ASN is 326 patients (just 59% of the original planned sample size), observed at γ_4 using the current trend assumption, with an associated power drop to 11.71%. This drop in power is not a concern as $\delta = \frac{1}{3}\delta_{plan}$ and there is no real wish to find a significant result as this is not a clinically relevant difference. Increasing the sample size is not a desirable characteristic for a design at this level of observed treatment effect.

Table 8.8 similarly gives ASN and power values for the promising zone designs from 50% data availability onwards. Only the current trend assumption in the design allowing for a trial to stop for futility at the 10% CP cut-off point sees a decrease in sample size, which is lowest at the 50% interim time point (516 patients, 93% of the originally planned n patients.

		SHORT					**MEDIUM**				
$\delta=\frac{1}{3}\delta_{plan}$		**Information fraction**					**Information fraction**				
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	689	706	714	712	692	728	755	768	758	723
	Power	22.60	23.08	23.29	23.11	22.38	22.59	23.11	23.40	23.19	22.43
γ_2	ASN	615	634	652	661	655	661	689	713	712	687
	Power	20.68	20.96	21.39	21.76	21.16	20.74	21.09	21.69	21.95	21.29
γ_3	ASN	426	466	506	542	573	482	532	579	600	608
	Power	14.97	16.06	16.80	17.94	18.38	15.41	16.68	17.65	18.57	18.66
γ_4	ASN	326	376	428	477	525	391	454	512	543	564
	Power	11.71	12.97	14.12	15.39	16.79	11.71	12.97	14.12	15.39	16.79
HYPOTHESISED											
γ_1	ASN	1015	1039	1011	928	802	1015	1041	1024	945	826
	Power	28.45	29.60	29.85	29.61	27.45	28.44	29.49	29.84	29.54	27.53
γ_2	ASN	963	993	956	878	760	963	998	977	907	787
	Power	27.50	29.03	29.40	28.98	26.57	27.39	28.96	29.41	29.04	26.65
γ_3	ASN	843	862	821	758	683	847	882	868	793	716
	Power	25.29	26.99	27.54	26.90	24.60	25.17	27.02	27.82	26.83	24.72
γ_4	ASN	612	595	580	574	578	641	648	639	628	614
	Power	20.80	21.96	22.03	21.62	20.82	20.92	22.31	22.18	21.73	20.89
80% Limit											
γ_1	ASN	895	876	856	832	760	906	901	896	862	786
	Power	28.18	28.45	28.29	28.09	26.51	28.11	28.45	28.46	28.08	26.52
γ_2	ASN	834	819	803	779	723	849	843	838	818	751
	Power	27.25	27.35	27.40	27.15	25.68	27.18	27.28	27.34	27.22	25.69
γ_3	ASN	663	688	689	681	662	714	723	736	729	694
	Power	23.99	24.83	25.10	24.87	23.94	24.45	24.75	24.99	24.89	23.97
γ_4	ASN	454	482	517	547	574	493	535	580	601	609
	Power	18.42	19.14	19.99	20.49	20.62	18.34	19.10	19.90	20.43	20.61

*Table 8.7: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when $\delta=\frac{1}{3}\delta_{plan}$, n=556, n_{max}=2*n*

δ=$\frac{1}{3}\delta_{plan}$		Information fraction									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	669	664	661	658	653	650	646	639	628	612
	Power	19.09	18.97	18.86	18.62	18.57	18.35	18.10	17.85	17.76	17.40
Hypothesised	ASN	887	850	805	769	735	708	685	663	641	616
	Power	26.92	26.00	24.84	23.64	22.42	21.47	20.68	19.53	18.81	17.93
80% limit	ASN	744	738	727	716	701	684	670	656	638	615
	Power	22.91	22.76	22.57	22.23	21.73	21.09	20.65	19.81	19.07	18.21
Promising zone with Futility											
Trend	ASN	516	520	530	541	550	562	573	583	589	593
	Power	17.25	17.31	17.37	17.30	17.38	17.31	17.25	17.15	17.27	17.10
Hypothesised	ASN	870	817	762	715	673	644	626	614	605	597
	Power	27.13	26.14	24.96	23.73	22.41	21.39	20.53	19.35	18.57	17.73
80% limit	ASN	671	663	647	636	627	613	608	605	602	596
	Power	22.87	22.73	22.47	22.05	21.55	20.91	20.43	19.60	18.84	18.01

Table 8.8: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions and three designs when $\hat{\delta}_{obs} = \frac{1}{3}\delta_{plan}$, n=556 and $n_{max} = 2^{*}n$

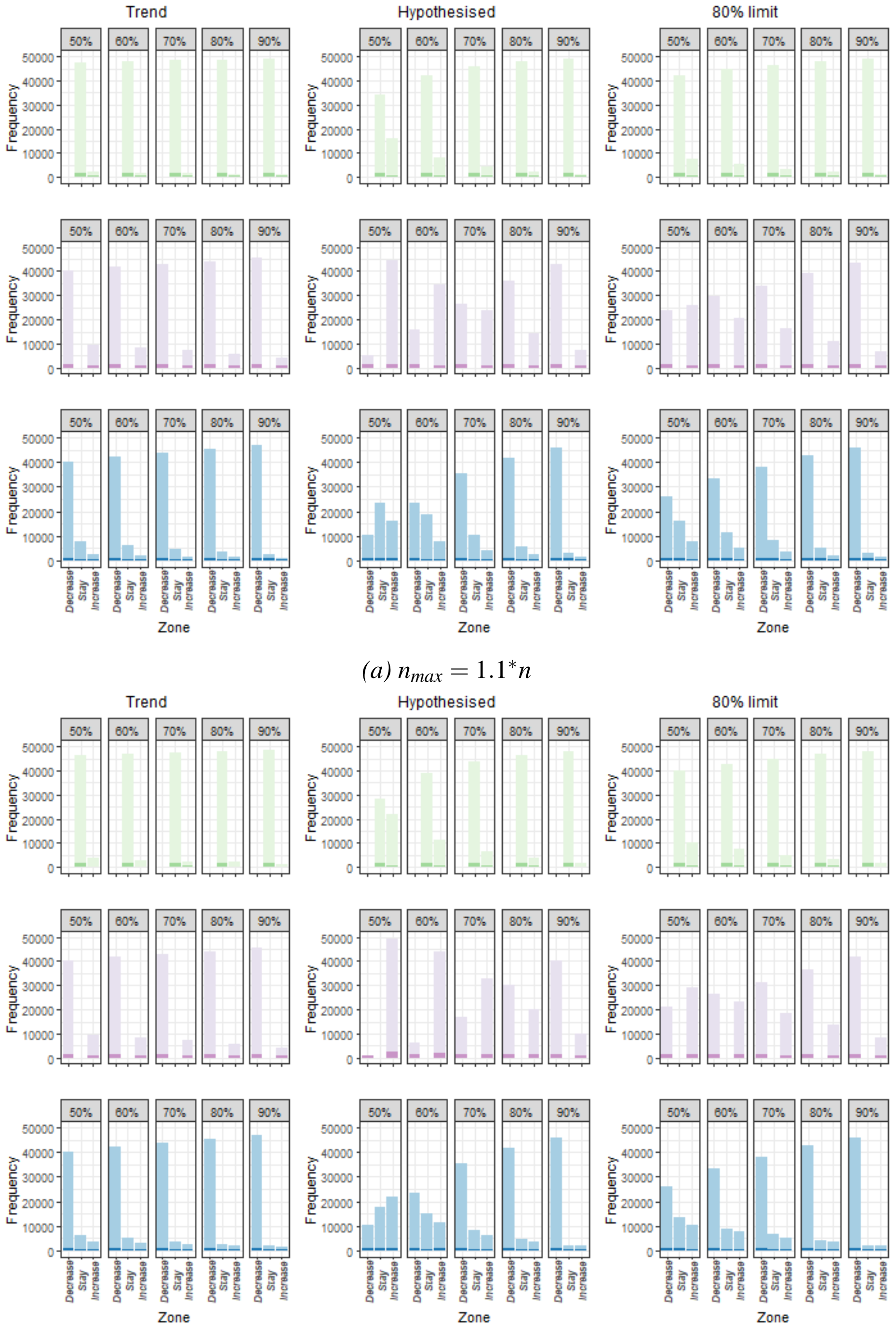
With this design and assumption, power does not see any values above 17.4% power. With no futility boundary however, power ranges between 17.4 (95% through) to 19.1% (50% through), under the current trend assumption.

Both ASN and power are greater in the 80% limit and hypothesised assumptions. The hypothesised assumption sees the greatest values of ASN up until 85% through the trial, where the 80% limit assumption sees greater power with a smaller ASN.

As seen in the previous examples, the addition of a futility boundary has decreased both ASN and power. The 80% limit assumption sees only the smallest of power decreases when incorporating a futility boundary, but sees a much lower ASN. For example, at the 60% information fraction, the 80% limit assumption sees a 16% inflation of sample size compared to 37% for the hypothesised assumption, and a 5% deflation using the current trend assumption.

Due to the decrease in sample size, the promising zone design with a futility boundary using the current trend could be considered as a preferable design when $\hat{\delta}_{obs} = \frac{1}{3}\delta_{plan}$. Additionally, the combination test design with $\gamma_4 = 0.001$ could also be considered, using either the current trend or 80% limit assumptions.

8.3.1.4 True effect = zero



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure 8.7: Sample size zones from 50000 simulations when $\delta = 0$ and n=556, for two values of n_{max}: comparing three designs and four observed treatment effects

Figure 8.7 shows the sample size state (increase/decrease/remain) when $\delta = 0$ and additionally shows statistical significance of trials at n^{*} patients with data available. As expected now that $\delta = 0$, almost all trials are not statistically significant regardless of sample size state.

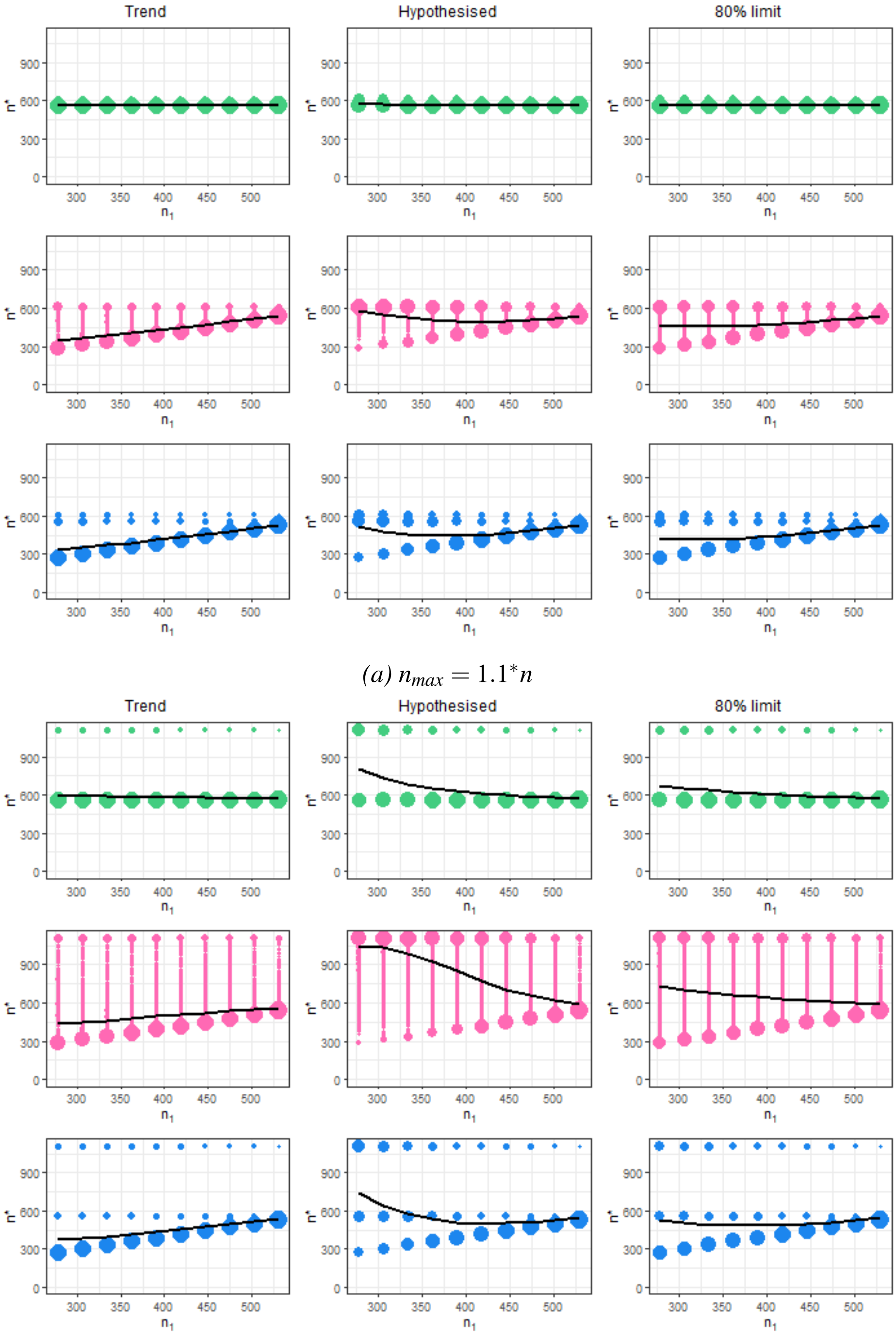
For the promising zone design under the current trend, there are almost no increases in sample size whatsoever. The hypothesised assumption has more increases particularly at the 50% interim time point. By 90% data available, the proportion of increases for the hypothesised has dropped, and almost no increases are seen at this time point. The 80% limit sees some increases early in the trial duration, lying somewhere between the other two assumptions.

The same pattern of increases are observed in the promising zone with futility as the design without, across all three assumptions. However, instead of the remaining trials falling in the remain zone, the trial is able to decrease in sample size if CP falls below the 10% mark. The current trend sees the largest proportion stopping early for futility, even at 50% trial duration. The hypothesised assumption on the other hand, sees the fewest trials stopping early for futility, which increases with increasing n_1. As for the proportion of increases seen, the proportion stopping for futility under the 80% limit assumption is somewhere between the other two assumptions, increasing with increasing n_1 values.

The combination test design sees the greatest proportion of increases in sample size, which does decrease with increasing n_1. The current trend assumption sees the fewest increases, with the majority of trials actually decreasing in sample size even at 50% available data. The hypothesised effect assumption sees very few decreases in sample size at 50% trial duration, which gradually increases, overtaking the proportion of increases by 80% through. Again the 80% limit assumption sees somewhere in between the states observed under the current trend and hypothesised effect assumptions, but overtakes the proportion of increases by 60% duration.

Figure 8.8 shows the magnitude of the new required total sample size n^* and the mean sample size from 50000 simulations. There is very little change in sample size when $n_{max} = 1.1^*n$ for the promising zone design under all any of the three assumptions.

The shapes of the expected sample size lines are very similar between the combination test and promising zone design with futility for this same value of n_{max}. The current trend



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure 8.8: Sample size from 50000 simulations when $\delta = 0$ and n=556, for two values of n_{max}: comparing three designs and three treatment effect assumptions

sees mainly $n^* = n_{rec}$ patients and very few increases, which means the expected line is steadily increasing as more data is available in the first stage of the design. The hypothesised assumption starts off seeing large proportions of sample size increase, and far fewer values of $n^* = n_{rec}$, which reverses with increasing n_1 values. This leads to a curve starting at its highest point, dropping somewhat in ASN until around 75% through, before a gradual increase to the largest investigated n_1. Finally, the 80% limit assumption sees an almost flat line somewhere between n_{rec} and n patients until around 70% through trial duration, before gradually increasing for the remaining values of n_1, up until around n patients.

For $n_{max} = 2^*n$, the current trend assumption sees the same expected sample size lines for the three designs as when $n_{max} = 1.1^*$, but shifted marginally higher due to the larger increases allowed in this design. The 80% limit assumption sees a small downwards slope for the promising zone design, starting just greater than n patients. A more pronounced slope is seen for the combination test design due to its slightly larger proportion of increases. The promising zone design with futility sees almost equal expected sample size at 50% and 90% data availability, but dips slightly in between these two time points. The hypothesised effect sees similar curves for the two promising zone designs; more pronounced in the futility design. The combination test however starts almost at $n_{max} = 2^*n$ at 50-55% information fraction, before a steep downwards slope to n patients at 95% through the trial.

Table 8.9 show ASN and power for four values of γ used in the combination test design. In the case of $\delta = 0$, power is the Type I error rate and should not be inflated past the nominal rate of 5%. A smaller ASN is desirable, but must not coincide with an increase in power.

The current trend assumption sees no increases in Type I error rate at any value of γ. Additionally, it sees the lowest value of ASN, with decreases in sample size seen for γ_2-γ_4, and between 50-80% trial duration for γ_1 for a short endpoint. The smallest ASN is 297 patients, just 53% of the original sample size.

Comparatively, Type I error is slightly inflated at early time points when using the 80% limit assumption. For γ_4 however, this is only an inflation of 0.07 and 0.09% for short and medium endpoints respectively. Later through the trial, Type I error is reduced, and even

		SHORT					**MEDIUM**				
δ=0		**Information fraction**					**Information fraction**				
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	490	506	526	547	566	554	584	612	618	609
	Power	4.12	4.15	4.15	4.32	4.29	4.17	4.25	4.15	4.32	4.38
γ_2	ASN	437	462	495	524	552	505	546	586	597	596
	Power	4.01	4.01	4.13	4.28	4.28	4.08	4.13	4.16	4.30	4.40
γ_3	ASN	339	385	433	479	526	415	474	530	556	571
	Power	4.01	3.99	4.15	4.34	4.37	4.17	4.23	4.29	4.44	4.50
γ_4	ASN	297	351	407	459	513	378	446	508	538	558
	Power	4.09	4.20	4.25	4.44	4.54	4.09	4.20	4.25	4.44	4.54
HYPOTHESISED											
γ_1	ASN	1077	1046	923	759	643	1078	1058	956	803	680
	Power	5.52	6.51	5.47	5.08	4.93	5.43	6.18	5.44	5.05	5.00
γ_2	ASN	1044	990	855	705	619	1047	1013	898	773	659
	Power	6.34	6.25	5.41	5.14	4.97	5.84	6.27	5.4	5.15	5.05
γ_3	ASN	919	829	702	618	578	937	874	789	685	621
	Power	6.93	5.75	5.28	5.17	4.97	6.57	5.9	5.32	5.14	5.05
γ_4	ASN	577	518	493	505	533	639	606	587	578	577
	Power	5.55	5.26	5.15	5.02	4.84	5.71	5.40	5.12	4.98	4.92
80% Limit											
γ_1	ASN	790	729	689	658	617	821	796	756	721	657
	Power	5.58	5.14	5.04	5.01	4.91	5.61	5.26	5.02	5.00	4.99
γ_2	ASN	729	679	647	618	598	766	733	714	687	639
	Power	5.50	5.20	5.01	5.04	4.94	5.53	5.28	4.98	5.03	5.02
γ_3	ASN	553	569	563	559	567	642	634	642	631	610
	Power	5.31	5.18	5.03	5.11	4.91	5.45	5.26	4.96	5.10	4.98
γ_4	ASN	386	413	452	490	531	448	493	542	564	575
	Power	5.07	5.02	4.91	4.92	4.85	5.09	5.07	4.84	4.88	4.91

Table 8.9: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when δ=0, n=556, $n_{max} = 2^{*}n$

sees a decrease in sample size for short endpoints. This reduction is not as great as under the current trend, seeing a minimum sample size of 386 patients, 69% of the original n patients. Fewer decreases are seen for medium endpoints however.

The hypothesised effect assumption sees the greatest inflation of Type I error rate and the highest ASN values. Whilst Type I error rate decreases with increasing information fraction, it still is largely greater than 5% until 90% trial duration. For medium endpoints, this inflation varies between smaller and greater than Type I error rates seen in the short endpoint case with the same values of γ.

δ=0		**Information fraction**									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	596	591	589	586	583	581	579	575	572	568
	Power	4.35	4.33	4.31	4.26	4.23	4.20	4.18	4.17	4.15	4.22
Hypothesised	ASN	800	732	683	652	627	610	597	586	577	570
	Power	5.14	4.88	4.68	4.60	4.51	4.40	4.38	4.30	4.26	4.29
80% limit	ASN	670	656	640	625	613	603	593	585	577	570
	Power	4.79	4.76	4.62	4.62	4.58	4.54	4.49	4.39	4.36	4.33
Promising zone with Futility											
Trend	ASN	373	385	402	419	438	458	479	499	521	543
	Power	3.89	3.90	3.92	3.95	3.93	4.07	3.98	3.99	3.97	4.18
Hypothesised	ASN	743	643	578	536	510	502	505	513	528	545
	Power	5.34	4.83	4.60	4.54	4.4	4.42	4.31	4.20	4.14	4.28
80% limit	ASN	524	508	491	485	487	491	500	512	528	545
	Power	4.67	4.62	4.49	4.52	4.44	4.54	4.41	4.29	4.24	4.32

Table 8.10: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions and three designs when $\hat{\delta}_{obs} = 0$, n=556 and $n_{max} = 2^{*}n$

Table 8.10 similarly shows ASN and power for the promising zone designs with and without a futility boundary. Only two instances of an inflation of Type I error are observed across both designs: both of which happen using the hypothesised effect assumption at the 50% time point. Whilst previously, the addition of a futility boundary has decreased power, in these two cases where an inflation occurs, the opposite pattern is seen. From 55% onwards however, all power is below the nominal 5% rate, and is greatly reduced under the current trend assumption with a futility boundary.

ASN again decreases with the incorporation of a futility boundary. Under the current trend assumption at 50%, 373 patients are required, which is 67% of the originally planned sample size. ASN is below the original n patients for the futility boundary design at all time points for the current trend and 80% limit assumptions, and from 65% onwards for the hypothesised effect assumption. The greatest increase in sample size is observed using the hypothesised treatment effect, with a ASN of 800 patients; an increase of 44%.

It may appear that earlier interim analyses may be acceptable as the Type I error, especially for the current trend assumption is much lower than the nominal level of 5%. However, interim analyses were chosen only from the 50% time point earlier as it was shown that the treatment effect was not very stable before this time point when real data was re-analysed. Researchers wishing to implement an earlier interim timing should proceed with caution, and consider adding some level of noise within simulation work.

From the simulations in the case where $\delta = 0$, the current trend and 80% limit assumptions are fine to use at any time point from 50% onwards. Additionally, the hypothesised time point may also be used from 55% onwards due to no further inflation of Type I error from this point onwards. In terms of ASN design, the current trend assumption sees the lowest values and is therefore the most preferable assumption. Also, the addition of a futility boundary sees the lowest values of ASN and would therefore be preferential compared to no boundary. If one wishes to use the 80% limit or hypothesised assumption with the futility design, it is recommended to conduct an interim analysis as 65 and 75% data availability for the two assumptions respectively. This is because no Type I error inflation, and the lowest values of ASN are observed.

8.4 Discussion

The results from 50000 simulations of a binary outcome with event rates planned at 10% and 20% have been reported in this chapter, also considering different values of the true effect δ compared to that planned δ_{plan}.

The first aim of this chapter was to appropriately simulate binary outcome data with two event rates: 0.1 and 0.2 when δ=δ_{plan} resulting in a planned odds ratio of 2.25, SE=0.25 and δ=0.276. Table 8.1 summarises characteristics of the simulated data from 50000 repetitions.

Whilst there are slight variations to that planned in the second decimal place, increasing the number of simulations was not possible due to the lack of high computational power. As simulations were repeated with three seed numbers and were still within a decimal place of each other, it was considered that this was sufficient.

The same four values of γ were investigated for the combination test design as seen in the continuous simulations. Results were broadly similar in that higher values of ASN and power were seen with smaller values of γ. Additional larger values of γ may have been appropriate here for the hypothesised assumption as power was higher than the 90% pre-specified limit when $\delta = \delta_{plan}$, and also slightly inflated when $\delta = 0$.

In Chapter 7 a highly inflated Type I error rate for the short term endpoint when n=264 was observed, and therefore seeing between 1-3% pipeline patients as minimum restriction on sample size lead to a very unstable second stage test statistic. However, Type I error rate for short term endpoints has been decreased in Table 8.9, even for a short-term endpoint. It is likely to be due to the increased sample size (n=556) and therefore more pipeline patients for the minimum restricted sample size. On the other hand, another contributing factor may be the difference in behaviour between binary and continuous data, particularly as event rates as deliberately been chosen to be far from 0.5. More work is required to investigate both binary and continuous outcomes and the effect of a minimum sample size restriction on Type I error rate.

When $\delta = \delta_{plan}$, the promising zone design with futility with the current trend assumption at a 50% information fraction saw the closest power to 90%. However, the 80% limit assumption saw a lower ASN with the same design, with higher power, and is therefore the most favourable design to use in this instance. Additionally, the combination test design at 60% information fraction using the current trend and γ_3, or at 50% information fraction using the 80% limit assumption and γ_4 have power close to 90%, whilst actually decreasing ASN compared to the original n patients.

When $\delta = \frac{2}{3}\delta_{plan}$, all designs see the hypothesised effect assumption with the greatest power, but this does come with an associated higher ASN. This problem does not have a single solution and each trial should decide the price of power increase versus expected sample size in order to determine which assumption to use in their scenario. However this investigation has provided the evidence for researchers to make a more informed choice around the design and assumption that is most appropriate.

When $\delta = \frac{1}{3}\delta_{plan}$, the promising zone design with futility design using the current trend assumption sees the lowest power and lowest ASN, which are good design characteristics when the treatment effect observed is too small to be of any value. Additionally, the combination test using γ_4 behaves well here for the current trend and 80% limit assumptions. However it should be noted that when γ_4 is used when planned effect = observed with the current trend, that power falls below the 90% level.

When $\delta = 0$, the promising zone design with futility boundary is again to be recommended with either the current trend at 50% information fraction, or 80% limit assumption at 65%, due to where the lowest ASN is observed. The combination test design with the current trend does not inflate Type I error rate at any time, or additionally with the 80% limit assumption using γ_4 from 70% onwards.

Collating the results of the binary simulations at each scenario, I would recommend the use of a promising zone design with futility boundary using the current trend assumption, or using the 80% limit assumption at around 65% through the trial duration (or as close as possible to this point), under the constraints imposed by this research (10% futility boundary, equal randomisation etc). Whilst power is increased beyond 90% when the planned effect is observed with the 80% limit, it has the lowest ASN here. Additionally, when $\delta = 0$, Type I error is not inflated and whilst the current trend assumption sees the lowest ASN here, it is still low, and behaves as a good middle-ground between the trend and hypothesised assumptions.

Additionally, a combination test design using the current trend assumption could be recommended. The specific value of γ should be based on a case-by-case basis, and I would recommend simulations as a way to investigate the magnitude required to have reasonable power, not inflate Type I error, and give a reasonable ASN.

8.5 Continuous vs Binary simulations

There are some differences in the results between the continuous and binary simulation work. However, it should be noted that this could be due to the difference in sample size considered (n=264 for continuous, and n=556 for binary).

When the result is as planned, trials with binary outcomes see a higher proportion of sample size increases and therefore higher relative ASN. However, power is also much higher in all designs except promising zone with futility using the current trend assumption, and some values of γ for the combination test design using the current trend assumption. This increased power is most pronounced under the hypothesised assumption at a 50% interim time point (93.32% vs 97.59%).

The same pattern is observed for $\delta = \frac{2}{3}\delta$, mainly under the hypothesised assumption again. Whilst higher power is observed here for all designs, a greater ASN is also observed. As power when $\delta = \delta_{plan}$ is much higher than the nominal level of 90% power, the increased ASN may not be considered worthwhile and an alternative assumption may be more favourable.

Simulations of continuous endpoints do not recommend a combination test design when the sample size is small due to the observed increase in Type I error when $\delta = 0$. This level of inflation is not observed for the combination test design using simulations of binary endpoints. Whilst some small inflation is observed typically with an earlier interim timing, some values of γ are able to control Type I error level from a 70% interim time point for the 80% limit assumption, 80% interim timing for the hypothesised assumption, and any time point under the current trend.

Both chapters of simulation results are able to recommend the use of a promising zone design with a 10% futility boundary using either the current trend or 80% limit assumption. However, a later interim timing is also recommended when using the 80% confidence limit assumption (from 65% interim timing onwards). For binary trials, the combination test design can also be recommended, but further research is needed for continuous trials with small sample sizes.

8.6 Summary

This chapter has expanded on the work of Chapter 7 to extend to simulations of binary outcomes where event rates are not close to 0.5. Data generating mechanisms have been presented, and a summary of the characteristics such as the odds ratio and standard error observed in 50000 simulations have been reported.

Investigating different event rates in the intervention group to that hypothesised has enabled the comparison at various possible outcomes of a trial, and therefore better evaluate the methodology. Whilst some designs work well when the effect observed is equal to that planned, and other designs may perform better for a zero effect, it is of interest to find a well-rounded design. A researcher may have the aim of showing an effect, it may be the case that there is no effect, or even the opposite, and a good design should be able to cope with observing something other than that planned.

The combination test with either the current trend or 80% limit assumptions have been seen to perform well in the four scenarios, but the timing of the interim analysis and the choice of γ parameter will impact ASN. If using this design, I would recommend a later interim timing of around 70-80% where possible, and suggest simulations as a good way to choose the γ parameter in order to ensure that power and ASN are appropriate.

The promising zone design with a futility boundary using either the current trend or the 80% limit assumptions has also been shown to not reduce power below the pre-specified rate, not inflate Type I error, and have lower ASN than the corresponding regular promising zone designs. The recommended interim analysis timings are 50 and 65% patients with data available for the current trend and 80% limit assumptions respectively.

The next chapter provides a case study of designing a prospective clinical trial with a binary outcome.

9 | Prospective case study: The RIPOSTE trial

9.1 Introduction

This thesis has so far reviewed the current literature and background to SSR designs in great detail, provided results from 21 retrospective case studies implementing uSSR designs, and presented results from 50000 simulations for both continuous and binary outcomes. Logistical features such as the timing of the interim analysis and the maximum value allowed have been explored, as well as delving into the CP calculation and investigating alternative future treatment effects. The 80% limit assumption has been shown to work well in both the retrospective case studies and simulations, and provides a good middle-ground between the current trend and hypothesised assumption when the effect is smaller than planned. The addition of a futility boundary gives the additional advantage of being able to decrease the sample size, as well as the combination test design provided a sufficient second stage sample size is recruited.

Within the clinical trials research unit at the University of Sheffield, a grant application was being developed which allowed for the possible application of the methods described in this thesis to a prospectively planned research study. This chapter provides details of the RIPOSTE study, specific simulation results, and comments from experienced trialists and clinicians on the logistical features of an uSSR design.

9.2 Aims

This chapter aims to show the planning stage of a prospective study if implementing a uSSR. Whilst statistical features of the designs have been explored in depth, alternative opinions of study team members would help to point out any logistical issues that may prevent these designs being used in practice.

9.3 Trial background

Emergency laparotomy procedures are associated with 90 day mortality rates of around 14%. It is thought that by using remote ischaemic preconditioning during the emergency laparotomy could protect vital organs after surgery. If used during the emergency laparotomy, it is thought that 90 day mortality could be as low as 9.1%.

The RIPOSTE study aims to test this hypothesis, and will be putting in a grant application to the NIHR in the near future. The study team were interested in an AD, and whilst also planning a fixed sample size design, agreed to consider a uSSR design. The details of the fixed sample size calculation are below:

> Event rates are assumed to be 14% and 9.1% for the control and intervention groups respectively, resulting in an absolute risk reduction of 4.9%, or a relative risk reduction of 35%. With planned power set to 90%, a 2-sided significance level of 5% and 1:1 randomisation ratio, 1786 patients with data available will be required in total (893 per group)

The trial will also allow for a 5% loss to follow up rate, resulting in an actual total planned sample size of 1880. As the simulated data will be based on data available only, the comparable fixed sample size of 1786 patients will be used. When implementing the design in practice, the sample size for recruitment targets will need to be inflated to allow for the 5% loss to follow up.

An initial discussion regarding the use of an AD was carried out, where the original fixed sample size calculation, target differences and recruitment details were shared. The meeting began with an overview of the promising zone design and conditional power, and a brief summary of the systematic review carried out in Chapter 3. A list of questions were prepared in advance, which provided the structure of the discussion following the introduction. Specific questions included:

1. What are your target recruitment rates for sites and patients?
2. What event rates would still be considered clinically meaningful?
3. Would an interim analysis timing of 60/70% data availability be an appropriate time point?
4. What would be the maximum constraint considered for a sample size increase?
5. Would you consider a futility boundary included in the design?
6. Would you still put in an application for less established methodology to what is currently used (e.g. an 80% optimistic confidence limit in the CP calculation)?
7. What would convince you to put in an application with a SSR?

9.4 Meeting with the trial team

9.4.1 Study design

Prior to the initial discussion, a clinician from the study team commented on the prospect of using an AD, saying:

> "I would like to keep the design as simple as possible. Particularly as NHS costings [...] are difficult to sort for even a simple trial. My peers would probably be less interested in delivering a complex design as unfamiliarity and complexity breeds contempt amongst surgeons"

It was decided to concentrate on getting the principles of uSSR across and therefore a method that was easy to communicate was chosen in the first instance. Due to the complexity of the combination test design and the decision surrounding the parameter of γ, this design was excluded from initial discussion, but could be revisited in later discussions if relevant. Therefore, only the promising zone designs with/without a futility boundary were taken forward for discussion with the clinicians.

9.4.2 Recruitment and pipeline patients

It is hoped that recruitment will take place across 25-30 sites, with approximately 2 patients per site per month. Assuming target recruitment is met for the maximum of 30 sites, it is estimated that there will be approximately 180 pipeline patients (recruited but not yet reached the 90 day outcome time point). This is roughly 10% of the original n patients, which falls somewhere in between the "short" and "medium" endpoints investigated in previous chapters.

9.4.3 Clinically important differences

Whilst the target difference is set at 4.9%, following expected event rates of 14% and 9.1% in the control and intervention arms respectively, values of clinical relevance and the MCID were also discussed. A clinician stated

> "We'd want to be hitting at least 10% in the intervention arm. [...] This is not a big study compared to say drug trials, but for a surgical study, this is quite a big trial, and 2% [absolute reduction] just wouldn't cut it".

Therefore a simulation will be run at the initial estimate of 9.1%, the MCID of 10%, a non-clinically relevant difference of 12% and no difference of 14% will be investigated for the intervention arm, whilst keeping the control arm consistently at 14% event rate. Restricting the control event rate is a limitation of this investigation, as there is also uncertainty surrounding this control group event rate. For initial investigations however, this was thought to be sufficient.

9.4.4 Timing of the interim analysis

Following the results of the simulation work presented in Chapter 8, it was recommended that an interim analysis take place between 60-70% through the trial. This suggestion was met very positively with the study team.

> "Certainly for a study this size, that feels absolutely reasonable. If you see futility with around 1200 people, you can stop [the trial] and no-one is going to

> argue with you. I can't see a surgeon arguing over a negative result when it's based on over 1000 patients"

They went on to comment on a recently published study that stopped early:

> "I've recently seen a study that stopped for efficacy after around 140 patients, which dealt with the primary but which wasn't necessarily the most interesting of the outcomes. [...] With 1000 patients, we're going to get some value from the secondary outcomes which we're also interested in. I know it's not what we're powering for, but if you're putting this much money into something and calling it early, actually 1000 patients will still give you that detail that you'd want to see."

The general consensus from everyone was that another small equivocal study is of no use to anyone as it would be too small to offer any evidence on decisions for policy making. On the other hand, if no effect is seen with a sufficient sample size, it is likely to be accepted that there is no true effect. Because of this positive feedback, an interim analysis between 60-70% through the trial will be considered. This would result in between 1072 to 1251 patients with data available, and between 1252 to 1431 patients recruited at the point of the interim analysis, assuming the target recruitment rates are met.

9.4.5 Maximum constraint on sample size

Following from the simulation work that had a particular emphasis on $n_{max} = 1.1^{*}n$ and $n_{max} = 2^{*}n$, it was recommended that a sample size increase of twice the original sample size would be appropriate from a statistical point of view. This recommendation however was met with some concern.

> "I'm thinking about the pragmatic things [...], which is keeping my colleagues on side for a bit longer. We're starting off with 1800 patients, and to say to them that we need to double the sample size because we think we've seen something, but need to double to get there, I would really struggle to carry people with me. I do not think it would be about funding."

> “If we were talking about a study starting with 600/700 patients and I turn around and said we need to double it, then that’s a different conversation. I think if you’re looking at big numbers such as over 1000, then a 50% increase is probably the limit you can get people to. Under 1000, I think you could make a case for doubling it, as a gut feeling.”

The clinician then went on to describe two current studies and reactions to a doubling of sample size. The first aims to recruit 5000 patients, and is already 2 years over their recruitment target time frame, having only recruited 3000. The other is a very large 12000 patient study.

> “The reason I say that is there are far bigger studies running that have really struggled [...]. It’s not that the condition is uncommon - it’s the same we’re looking at here. If we went from 5000 to 10000, I think we’d be laughed out of the building”
> “Then again, there’s another study recruiting around 12000 patients. I think with this study, people may sigh but would probably accept it. So much has been invested in it and it’s an important current problem and so they would probably have the justification if they said that was what was needed.”

In light of this, only a 50% increase in sample size will be considered for the simulations for the RIPOSTE study. It was mentioned that when only a 10% increase was seen, the study dropped in power. Whilst an increase was indicated, not as many of the studies went on to see a significant result as when $n_{max} = 2^{*}n$ as the increase was too low to see any benefit. A maximum increase of 50% has been considered, but provided in the appendix.

9.4.6 Funding considerations

The future treatment effect assumption was raised, and that the current version of promising zone design used the observed current trend so far assumption. The hypothesised assumption was briefly mentioned, but due to the simulation work of Chapter 8 and data re-analysis results from Chapter 6 was not suggested for implementation. Finally it was highlighted that the 80% optimistic confidence limit was not a currently used assumption for this design. Whilst it exists in the current literature, to my knowledge it has not been implemented in conjunction with a SSR outside of this thesis. When asked if this would be a deterrent from applying for funding using this assumption, the team replied

> "From the perspective of a researcher: As you know, doing things that do not fit the established pattern can be risky. So if I were putting in [a funding application], I would want to know that it was going to a panel that was open-minded enough to accept something different to what they normally do"

> "On the other side: If I were reviewing a grant that had this in, I'd actually look quite favourably on it. One of the questions is about value for money for the funder, and I feel there's a very strong argument for this when you put an emphasis on early termination. If you put the emphasis on "I may need to ask you for a lot more money", that then changes the tone of your contract negotiation."

An opposing view commented

> "The concern is that with an emphasis on futility, it can be seen as a sign of your lack of confidence in the proposed study"

> "[From a clinical perspective] I do like the idea. I can see where you're coming from. I can see the value for money. There are just a few little bits of information that we'd need to be bold enough to put in a grant application"

A question was raised as to whether a reviewer either from a panel member or academic journal side had been asked for opinions.

> "I can imagine that if we did a sample size re-estimation partway through, no matter how well justified, you might get a bit of a hard time, and I just wonder if it's worth understanding some of the challenges that could be a barrier to this"

The team requested that this design be raised with a funding body panel member before final consideration of the study design. It was agreed that this would be a really useful next step to understand the inside perspective of the funding body, their understanding, and what they may value in an application.

9.4.7 Additional comments

When asked "What would convince you to put in an application with SSR?", the response from a clinician was

> "It would be really interesting to see about half a dozen or so surgical trial specific case studies, apply the methods, and see the outcome. That would give me a bit of security, particularly as [surgical trials] often look at things that are this lower end of event rates, and we don't want to be in this position where every surgical study needs its sample size inflated partway through, or every surgical trial being stopped partway through."

9.5 Simulation results

Following the meeting with the trial team, simulations were run to ensure that no Type I error, or reduction in power is observed. The results are based on 25,000 simulations using the promising zone with or without a futility boundary, with $n_{max} = 1.5^{*}n$, and investigates three interim timings between 60 and 70% data availability. The corresponding CP_{min} values are 0.399,0.394 and 0.389 for 60,65 and 70% interim timing respectively. The upper boundary of the promising zone is kept consistent at $1-\beta$ (90%). Event rates in the intervention group (π_B) are chosen as the original estimate (9.1%), the MCID (10%), a non-clinically relevant value (12%) and no effect (14%), whilst keeping the event rate in the control group (π_A) constant at 14%. The current trend assumption with the futility boundary drops power below the pre-planned 90% value at the 60% data available interim timing. Additionally, the largest expected sample size is largest when the treatment effect is exactly as planned (14% vs 9.1%), rather than increasing the sample size when smaller than planned but still promis-

		π_B=9.1%			π_B=10%			π_B=12%			π_B=14%		
		60%	65%	70%	60%	65%	70%	60%	65%	70%	60%	65%	70%
No Futility													
Current trend	ASN	2058	2037	2016	2093	2078	2061	1970	1962	1955	1829	1825	1823
	Power	91.99	92.05	92.06	77.28	77.26	77.26	24.94	24.84	24.71	4.70	4.66	4.63
80% limit	ASN	1962	1952	1947	2043	2035	2029	2057	2041	2023	1898	1879	1862
	Power	93.70	93.51	93.36	80.63	80.32	80.26	27.74	27.45	27.05	4.99	4.96	4.89
Futility													
Current trend	ASN	2012	1999	1985	1981	1982	1978	1601	1630	1665	1223	1285	1353
	Power	89.42	90.02	90.47	73.94	74.40	74.74	22.93	23.08	23.28	4.26	4.25	4.25
80% limit	ASN	1949	1938	1934	2004	1995	1989	1851	1833	1823	1424	1428	1449
	Power	93.20	93.20	93.02	80.19	79.81	79.78	27.62	27.32	26.89	4.81	4.76	4.70

Table 9.1: Simulation results in terms of ASN and power for the RIPOSTE study using the promising zone design with and without a futility boundary, with $n_{max} = 1.5^{*}n$, π_A=14%, and π_B=9.1, 10, 12 and 14% respectively.

ing. Furthermore, the 80% limit assumption sees smaller ASN compared to the current trend assumption when observed effect is exactly as planned, and has higher power at all three investigated interim timings. For these reasons, the 80% limit assumption is recommended over the current trend assumption when a futility boundary is incorporated.

From the conversation with the study design, it seemed that the futility boundary and therefore allowing a decrease in patients should the data suggest this at the interim stage, seemed to be the real attraction to the design. Seeing a CP value of <10% at the interim analysis and yet continuing to the pre-planned sample size does not sound reasonable, nor provide the value for money that was discussed. Whilst clinicians would be blinded to the exact CP calculation (particularly the chief investigator), and would only know "remain as planned" or "increase to n^* patients", knowing that you may be continuing a trial that is able to show futility does not seem logical.

Figure 9.1 shows the results from 25,000 simulations at four scenarios: as planned (9.1%), the MCID (10%), too small an effect to be clinically relevant (12%) and no effect (14%). Similarly to the simulation results, the darker blue shades in the zone plots represent a significant result with n^* patients with data available.

Whilst the expected sample size lines are greater than the the originally planned n patients for π_B = 9.1 and 10%, so is the power. Even at 10% (the MCID, an increase in sample size is largely found to be a significant result at n^* patients, despite being smaller than that

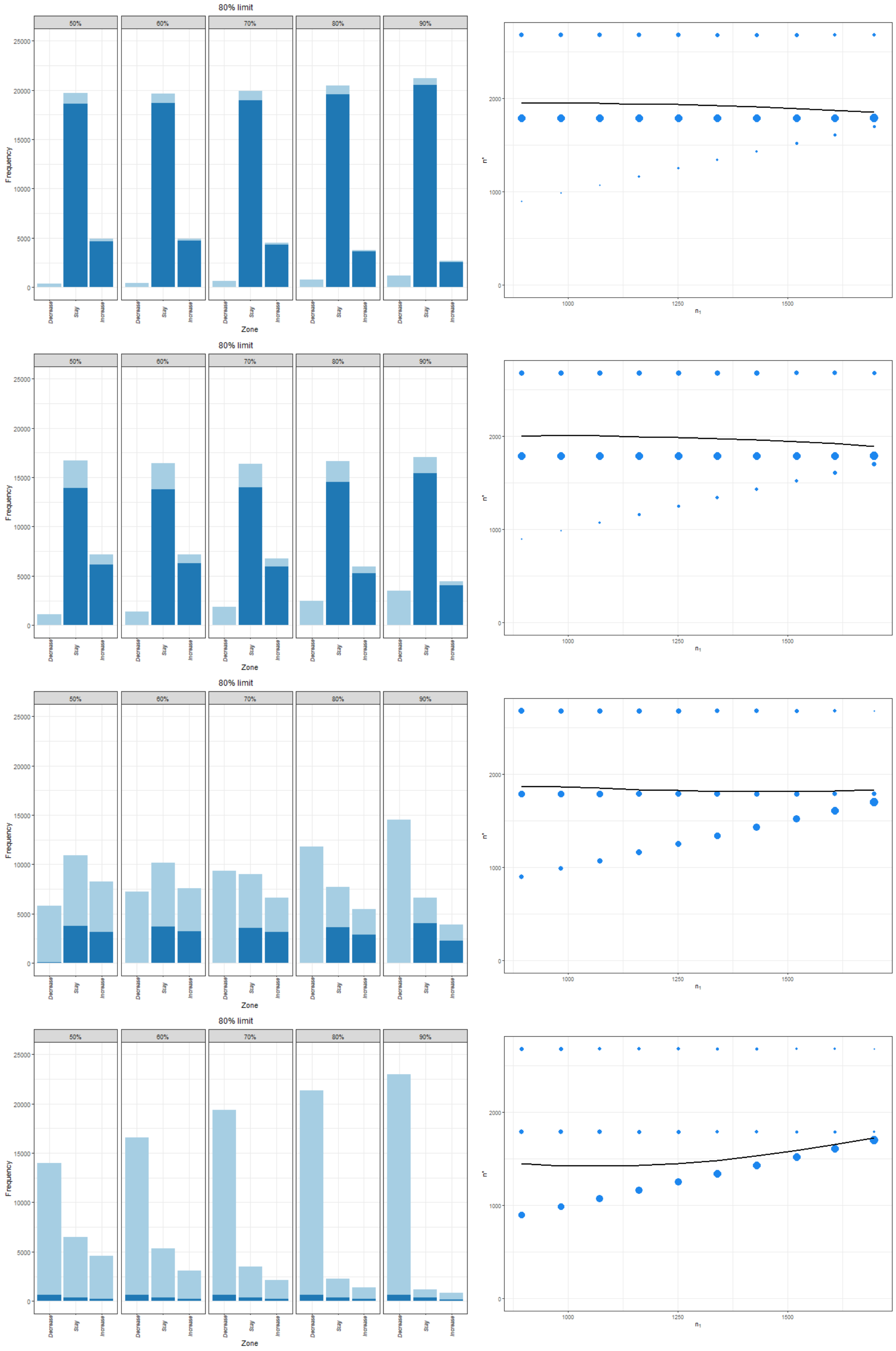


Figure 9.1: Sample size zones and expected sample size when the intervention group is (Row 1) 9.1%, (Row 2) 10%, (Row 3) 12% and (Row 4) 14%

planned. At 12%, the expected sample size line is still slightly above the original n patients, but a large number are stopping for futility at this point. By 14% (i.e. no difference between the control and intervention arms), there are very few increases, and the trial largely stops for futility. This brings the expected sample size line down below the original n patients.

The RIPOSTE study aims to detect an absolute risk reduction of up to 4.9% (14% vs 9.1%) with 90% power and a 5% two-sided significance level requires 1786 patients. Using a promising zone design with a futility boundary, with a 70% interim time point (data available), $n_{max} = 1.5^{*}n$ is expected to have between 1449 and 1989 patients . The maximum possible sample size is 2679 patients, and a minimum of 1430 patients. This design has been shown through simulation to reduce Type I error rate (to around 4.7%), and has power of 93% for detecting the anticipated 4.9% difference, and 80% power for detecting a 4% MCID.

9.6 Meeting with a HTA panel member

At the request of the trial team in the initial discussion, a meeting was set up with a HTA panel member, one of the NIHR funding streams in the UK. Again, a set of questions were prepared prior to the meeting, including:

1. Have you seen any study proposals implementing a uSSR?
2. What would be your opinion of a study including a futility boundary within a uSSR design?
3. Would you want to see the costings set out for possible sample sizes (e.g. a minimum/maximum value)?
4. Would you want to see simulation results in the grant application?
5. Would you currently consider funding a study using uSSR

A sample size re-estimation was not a new concept to the panel member, who had seen such designs previously. He went on to explain that the HTA generally require an internal pilot stage, and at this point (usually about 1/3 through the trial), that any uncertainty in initial estimates used in the sample size calculation would be addressed by the DMSC.

> "If there is uncertainty about baseline event rates, the HTA may check they are in accordance with the initial estimates at [the end of the internal pilot]. There would be a SSR carried out by the DMEC. [...] The HTA are particularly worried about needing a much larger sample size than stated at this point. They would have already sunk a lot of the costs into the trial, and so it puts the HTA in a difficult position."

The discussion continued, and when asked when a panel may consider a SSR more or less favourably than a fixed sample size design, the response was

> "If it's a well established research area, with a good registry process, you should know the baseline event rates and so we would certainly question why you don't have a fixed sample size and therefore cost."

> "The HTA are particularly interested in new research areas, and so if this is the case and initial estimates are uncertain, we would certainly welcome such a design. It shows you are aware of this and have come up with a way to mitigate the risk."

The two opposing views on how a futility boundary would look to a panel; value for money, or showing a lack of confidence in the study were discussed.

> "I wouldn't be worried about it showing a lack of confidence in the study. HTA are generally favourable and I certainly encourage the use of a futility analysis."

> "The main point however, is who does the futility analysis? The trial team are likely to be aware that a futility analysis is taking place and would know it hasn't met the criteria if they do not hear anything. We're likely to put this into the DMECs hands, who mostly want overwhelming evidence to stop. Whereas, the HTA would generally be fine with less strict criteria."

The discussion moved on to costings, and whether a range of costs would be looked on more or less favourably by a panel, than say one fixed “expected” cost.

> “We would want to see a properly thought out trial in terms of costings. We would want to see the worst case scenario. You do need clarity, otherwise it undermines your application. A lot of it is about having confidence in your design. We need to believe that the team know what they’re doing.”

When asked about providing simulation results in the application itself, the panel member responded

> “There is a limit on the number of pages, so being succinct would be a strength. [The simulation work] shouldn’t take space away from important clinical information.”

> “Applications would undergo strong statistical scrutiny. There are lots of statisticians who sit on the panel, who are very keen to replicate [the statistical aspects of the design]. You need to be transparent, and if it has been carefully thought through, [the design] would be well received.”

9.7 Final comments

The comments from the HTA panel member were taken back to the study team. After much discussion, the trial team decided to change the primary outcome, with existing literature being used to inform the baseline event rate in the sample size calculation. As it was felt that at least the minimum clinically important difference could be met, the study would not be using a SSR approach in the trial design. However, they quite liked the idea of a SSR approach, and some final comments from the trial team included:

> [Clinician] “I would probably consider it [for future trials], particularly in those situations where baseline rates might not be quite tacked down, which can happen in emergency surgery where the cohort literature is poor”

> [Statistician] “If we had continued with the initial choice of primary we could well have used the uSSR approach developed. I will consider this approach in similar situations in future.”

Overall, the discussions were hugely positive towards SSR designs, particularly when incorporating a futility boundary. Whilst already having been shown to be statistically sound in previous chapters, this chapter has highlighted that the logistical hurdles associated with this design are no huge barrier to the implementation of a trial, or indeed potential acquisition of funding. It was a good exercise to work with an ongoing study team, to communicate the concepts and get insights from different perspectives.

The suggested interim analysis timing of 60-70% informed by previous work matched the clinical opinion. However, initial opinions did not match for the maximum restriction of sample size. From previous chapters, a 100% increase was recommended, but when put in logistical terms, it was decided a 50% increase would be better. Looking at the results, there is no statistical issue from using a 50% restriction. Whilst power is slightly reduced, it is still above 90% when $\delta_{obs} = \delta_{plan}$. Therefore the discussion has been really useful for making future recommendations, discussed in Chapter 10, making allowances for the logistical implications too.

Whilst not all trials may be suitable for a uSSR design, this chapter has certainly highlighted that this design could offer a great advantage in new and emerging research areas. It is hoped that the work in this chapter will help researchers wishing to apply a uSSR design in the future.

10 | Discussion

10.1 Introduction

Chapters 5, 6 have presented results from data re-analysis using real clinical trial data, and Chapters 7 and 8 from the simulation studies for continuous and binary outcomes. This chapter brings together the results and uses them to make recommendations for trialists looking to implement an uSSR in a future study. The work will be discussed in relation to current literature, particularly that published since the start of the thesis. Any limitations will also be considered, before recommending areas for future research.

10.2 Summary of results

Chapter 5 reported the data re-analysis for 21 outcomes from 14 real-world clinical trial datasets across both publicly funded and industry settings. The observed effect at the original *n* patients was taken as the assumed true value of the population, as this is otherwise unknown. The treatment effect was calculated after every 10 patients, observed in the same order as patients entered the trial, and the reverse order, to see estimate stability of the observed treatment effect. Additionally, the treatment effect from 1000 random orders of patients were calculated, and the median plotted and compared to the original order. From an average of 57% of the original sample size with data available, the observed treatment effect remained within ±1*SE from the original observed effect. For this reason, only interim analyses after 50% were considered for the simulation study, as this did not add in any bias in the data generating mechanisms. Beyond this point, it is assumed the effect is sufficiently stable, but it should be noted that this improves even further with more data.

Chapter 5 saw little reason to believe that a futility boundary used in conjunction with the current trend or hypothesised effects would have much value, if any. The trend assumption appeared noisy, even around halfway through the trial, and a boundary would have incorrectly stopped a significant trial where the assumed true treatment effect was non-zero. On the other hand, the hypothesised limit was slow to decrease CP values, and a very late interim analysis would be needed to be of any benefit. However, this would decrease the potential saving of sample size, as more patients will have been recruited by this point. Finally, the optimistic limits were shown to get to one or zero faster and could offer a good middle ground between trend and hypothesised assumptions.

A limitation of Chapter 5 was the limited trial data, with only 6 non-statistically significant trials. In order to assess all trials at a user specified level, Chapter 6 transformed the data to see how CP differed, and therefore the impact on SSR decisions. Trials with binary outcomes were found to be very noisy in terms of CP values, and therefore continuous and binary outcomes were split up. Trials with continuous outcomes were quick to reach one when observed and planned effects were equal, and to reach zero when no effect or a zero effect was observed. Trials with binary outcomes were noisy for observed equalling planned and half planned, but behaved similarly to the trials with continuous outcomes for zero and negative effects. Chapter 6 confirmed the results of Chapter 5 in terms of the addition of a futility bound for trend and hypothesised, and that 80/90% limits could be a good compromise between the two. Again, the trend assumption behaved well when a very small effect was observed, and hypothesised when as planned or smaller than planned but not zero or negative. However, 80/90% limits were able to quickly reach the final value of CP earlier in all four investigated scenarios.

In light of the data re-analysis chapters, the assumption that the future treatment effect follows the observed current trend is not a good one to use in the CP calculation. It could result in unnecessarily high sample size increase when the treatment effect is close to that planned, but not when a very small effect is observed. As it is the current recommended assumption to use, it has been left in for the simulation investigation. The hypothesised effect could be used, but would result in high expected sample size, as a futility analysis is unlikely to benefit the design. Finally, 80 and 90% limits behave similarly, with 90% limit behaving more closely to the hypothesised assumption. Therefore the 80% limit was chosen to take forward in the investigation, and would likely benefit from a futility boundary.

The simulation study reported in Chapters 7 and 8 concluded that a promising zone design with a futility boundary is the most recommended design in terms of ASN and power across the four investigated scenarios. It also concluded however, that the current trend assumption would also be fine to use, and in many cases the best in terms of ASN and power. It is likely that this is due to the lack of bias introduced into the simulated data. The simulations are not affected by the noisy real-world data and therefore a discrepancy in results has come up between the retrospective data analysis and simulated data. In both however, the 80% limit assumption was seen to be a good "middle-ground" across the four values of observed effect compared to that planned, and appear to be a more realistic assumption to use regardless of observed bias in real-world trial data. Simulations incorporating high levels of noise could be investigated for further work to confirm this finding but is considered beyond the scope of this thesis.

An issue was found for continuous outcomes, with the hypothesised effect assumption inflating Type I error rate for either promising zone design, which was not seen in the binary outcomes beyond the 55% interim timing. For 264 patients and a continuous outcome, Type I error was inflated for all assumptions using the combination test design. This inflation was smaller for a medium endpoint, where more pipeline patients were observed. When investigated, this was due to the unstable second stage test statistic when very few pipeline patients were observed. This has not been reported in the literature as far as I can find, with much larger numbers of pipeline patients having been explored (i.e. with longer outcomes), and is an important finding of this thesis. A minimum restriction of the second stage sample size would likely resolve this problem, and further work into this is encouraged.

A thesis published earlier this year that also compared combination test and promising zone in the case of an AdGSD (i.e. ≥ 2 interim analyses) reported the combination test design being most efficient in terms of ASN, which is also what is seen in the research of this thesis (Jimenez 2020). However, the choice of γ and the impact on power and Type I error when $\delta = \delta_{plan}$ and $\delta = 0$ needs to be carefully considered when designing this kind of trial.

In light of the findings from the previously discussed data re-analysis and simulation work, the next section presents recommendations in terms of the initially set out aims and objectives of the thesis, with the reasoning of each recommendation briefly described.

10.3 Recommendations

It should be noted that the recommendations outlined below follow on from the work of this thesis, and apply to the setting outlined previously only; for continuous or binary endpoints with 90% nominal power, two groups and a 1:1 randomisation allocation, considering a 10% futility boundary. Researchers wishing to implement uSSR designs outside of this scenario should take into consideration the differences in settings before using any of the following recommendations, or adapt the following as appropriate.

Recommendation 1: When should a uSSR design be considered

- Limited prior information: When there is a rich source of data in which to estimate a treatment effect for a sample size calculation, the designs considered here may not be appropriate to use. It is recommended that researchers assess the uncertainty surrounding each of the design parameters, as this will be required in order to justify design choice. If there is only uncertainty around a nuisance parameter, blinded methods would be recommended over unblinded due to the increased risk of bias being introduced.

Recommendation 2: Timing of interim analysis

- No interim analysis before 50% through: Before halfway through the trial, the estimate has been seen to be biased, and may not be stable enough for appropriate interim decision making. Therefore, post 50% is highly recommended.

- The later the interim analysis the better: When the observed effect is close to planned, a later interim analysis is best in terms of a low ASN; however the opposite is seen when a very small effect is observed. A recommended timing therefore would be approximately 60-70% where possible, based from the simulation work.

Recommendation 3: Maximum allowed sample size

- A 10% maximum increase is too small: When only a small increase of 10% is allowed, the number of trials that go on to be significant with n^* patients is small. Despite indicating an increase is necessary, the allowed sample size increase is insufficient to maintain the power. If only a small increase is able to be considered, these designs would not be appropriate to use and an alternative should be sought.

- Larger values of n_{max}: The larger the allowed increase in n_{max}, the larger the ASN. A value of twice the original n is seen to be sufficient, and most trials that do increase in sample size, do go on to be significant after n^* recruited patients. A moderate increase of 50% is the smallest increase that should realistically be considered given the results of this thesis. Other values have not been investigated and further research is recommended.

Recommendation 4: CP future treatment effect assumption

- Trend assumption should be used with caution: This assumption saw conflicting results between the data re-analysis and simulation work. The data re-analysis suggested the trend assumption was too noisy, even around the 50% mark, and was slow to increase to CP$\approx$1 even when the observed effect was exactly as planned. However, the simulations did not account for bias, and showed this would be a good assumption to use in a design. Therefore, it should only be used with extreme caution, and as late an interim time point as possible.

- Hypothesised assumption can be used in a trial with a binary endpoint: Type I error rate is inflated in the continuous case, but is controlled for binary outcomes past 55% through the trial. This assumption consistently sees the highest ASN however, and rarely stops early when the observed effect is small. This assumption should only be considered if there is no real wish to stop a trial early, and the later the interim analysis the better.

- The 80% limit poses a good compromise between trend and hypothesised assumptions: Whilst the current trend assumption is best to use when for a zero or small effect, and the hypothesised assumption for a close to planned effect, the 80 and 90% limits have been seen to offer a good compromise between the two, whether the observed effect is close to that planned, a small effect, or a zero effect. This result was confirmed in the simulation work for the 80% limit, and in light of this would be the most recommended assumption of the three.

- Get input from DMSC/TSC: Researchers could pre-assess their level of uncertainty at the planning stage to aid the choice of assumption to use and get input from the DMSC/TSC for the ultimate decision. With very limited certainty surrounding δ, the decision may lean in favour of the current trend or 80% confidence limit; however with slightly less uncertainty, it could be thought that the planned effect from the protocol would be most favourable to the committees.

Recommendation 5: SSR design

- The promising zone design with futility boundary is the most practical design: Whilst the promising zone design is also fine to use, the allowance for early stopping reduces ASN without any big compromise in power. It is easy to implement for researchers, in terms of design and analysis, and offers a “safety net” for when the trial is nowhere near what is hoped.

- Combination test design can reduce expected sample size: The combination test has also been seen to be advantageous, with later interim timing of around 70-80% and either trend or 80% limit. Again, this was discovered in the situation where no bias was considered, and the trend assumption should be used with caution. If low numbers of pipeline patients are expected, a minimum second stage sample size should be implemented in order to preserve Type I error.

Recommendation 6: Values of γ

- Larger γ values penalize large sample size increases: In line with the current literature, larger values of γ correspond to smaller ASN values and also power. If a value of a smaller than hoped but still clinically meaningful effect is known, the design can be optimised in terms of γ for ASN and power. However, one limitation of this thesis is that these values are unknown. A number of values have been investigated by simulation, and this method would be recommended when designing this kind of trial. The combination test design is more complicated, but not unmanageable, to implement this design practically.

Recommendation 7: Pipeline patients

- Shorter outcomes have lower ASN: Shorter outcomes have less pipeline patients and therefore lower minimum sample size values for the combination test design. When the recommendation is to stop recruitment immediately, the minimum sample size of n_{rec} patients is chosen, and therefore compared to a medium outcome with more pipeline patients, has a lower ASN.

- Small trials and the combination test design: Small trials that are unlikely to see many pipeline patients, or trials with slow recruitment should consider a minimum sample size to use instead of the n_{rec} design. A small second stage sample size may lead to an unstable test statistic and therefore has been shown to increase Type I error in the continuous case.

- Longer outcomes are less likely to have a sample size saving: Unless a surrogate endpoint can be used for interim decision purposes, a long time to primary outcome data would usually result in too many patients in the pipeline, and therefore no sample size saving for the case of a futility boundary, or the combination test design. However, if recruitment is extremely slow, this may not be such an issue, but a sample size increase could heavily prolong the recruitment time for this scenario.

10.4 Suggestions for future work

Whilst the thesis has made clear recommendations in one setting (e.g. 10% futility boundary, continuous/binary endpoints, 80% nominal power etc.), there are other scenarios where this thesis is not able to necessarily recommend specific trial details, and more work is needed. Specific suggestions for future work include:

- Stopping boundaries: The use of additional levels of futility boundary beyond only the 10% boundary considered here. The addition of an efficacy boundary should also be considered, such as a 99% CP boundary for overwhelming evidence.
- Only 90% nominal power has been considered here and so a lower level of 80% should be investigated. Additionally, a trial with 90% nominal power but using an 80% acceptable level for no further sample size increase (i.e. a lower favourable zone boundary in the promising zone design) could be investigated
- Allocation ratios that differ from 1:1 equal randomisation could be considered, as well as adding in additional treatment arms (i.e. more than 2 groups)
- Whilst survival endpoints have been used in the literature, it would be useful to have some clear recommendations such as those given in this thesis specifically for survival endpoints
- More guidance would be welcomed around pipeline patients, such as an acceptable ratio of accrual time compared to time to primary outcome data collection
- It would be useful to research further the estimation of the treatment effect and corresponding confidence intervals in the final analysis, regardless of the design used and corresponding design features

10.5 Summary

This chapter has summarised the results from the three main studies in this thesis. Combining the findings from data re-analysis and simulation work, recommendations have been made in an easy to follow way. The aims of the thesis have been used as recommendation headings, and therefore it is clear to see that the thesis objectives have been achieved. The thesis provides illustrative examples of several SSR methods using real-world data, and should hopefully provide thought-provoking considerations for trialists looking to implement a uSSR in their studies going forward. Future work has been suggested in Section 10.4 and it is hoped that the thesis sparks additional research in this field. The next, and final, chapter summarises the aims, methods, and results of the work achieved in this thesis.

11 | Conclusion

This thesis has compared a number of factors in relation to uSSR design implementation, and considerations for the practical application of these designs. The promising zone design was first introduced in 2011, and has been utilised by at least 29 trials, as found in the trial systematic review from 2018, and an updated search more recently (Chapter 3). The current literature around promising zone design has been comprehensively reviewed, and strengths and limitations discussed in detail. Having seen alternative designs to the promising zone for uSSR during the literature review in Chapter 3, the combination test was chosen as a comparative design. Following the literature review, specific thesis aims were to:

- Compare existing methodologies for uSSR using CP calculations, with a focus on promising zone and combination test designs
- Incorporate stopping boundaries in each methodological framework and compare interim decision making
- Investigate the future treatment effect assumption used in the CP equation
- Explore CP values when observed effect sizes are equal to, or different by some amount to the target effect size
- Make recommendations for the planning of a future trial using SSR including operational considerations such as when an interim analysis should be carried out, and the maximum sample size increase to consider

Chapter 4 describes methods for obtaining the real-world trial data from publicly funded and industry settings, as well as a detailed analysis plan for the data re-analysis. Finally, it summarised trial characteristics, such as recruitment rates and observed vs planned effect sizes. The original data re-analysis results are seen in Chapter 5. CP values are observed graphically, split between a significant or non-significant final result from the original analysis of n patients. Finally stability of the estimate was investigated, and led to only investigating interim analyses of post 50% through the trial. An additional framework (the stepwise methodology) was used as an illustrative example in Chapter 5. As this design is less powerful, and used predominantly when there is an issue with the potential for the back calculation of the treatment effect following knowledge of the interim decision of n^*, it was not taken forward in Chapters 6 and 7.

Due to the varying observed vs planned effect sizes seen in Chapter 5, original trial data were transformed a number of times, to specify an end result of that planned, half that planned, zero effect, and a negative effect (Chapter 6). Therefore, CP curves could be further investigated, knowing the decision at the end of the trial. Together with Chapters 4 and 5, this chapter informed the simulation plan seen in Chapter 7.

It was thought from the data re-analysis results, that an optimistic 80 or 90% confidence limit would provide a good compromise between trend and hypothesised assumptions, and able to work well whether the observed effect was close to that planned, or much smaller. Additionally, this assumption would likely benefit from a futility boundary, able to correctly stop when the observed effect was low.

This was confirmed in the simulation work, with results from continuous and binary outcomes reported in Chapter 7 and Chapter 8 respectively. Recommendations made in Chapter 10 are summarised, combining results from Chapters 4 - 8 to provide insight to trialists looking to implement an uSSR in future trials. An interim timing of around 60-70% is recommended where logistically possible. A maximum increase of around 1.5 or 2 times the original sample size is recommended, and a maximum increase of 10% has been found to be too low to have a benefit.

The trend assumption should be used with caution, and has been seen to be influenced heavily on the estimate stability, which could impact interim decisions. Additionally, the hypothesised effect could be used in the CP calculation for binary trials, but only if there is no wish to stop early. The promising zone design with a futility boundary is recommended, and suitable regardless of the observed effect size (between as planned and zero). The combination test design can also be considered when designing the trial, which requires a pre-specified constant of γ. If a value of the MCID is known, the design can be optimised, according to the current literature (Jennison 2015; Pilz 2019) for power. This has not been investigated here, which is one limitation of the work of the thesis. However, this could be recommended for future research.

A long time to primary outcome is unlikely to be beneficial in terms of sample size saving, as most patients will likely have been recruited. However, if a surrogate endpoint could be used instead for interim decision making, this could be implemented. Shorter outcomes also result in a lower ASN value, due to the lower number of pipeline patients at the interim time point, and therefore the lower the minimum allowed sample size, which this design allows. As discussed, pipeline patients can impact Type I error for small studies, and a restriction imposed on this minimum value is required.

Finally, the comments from clinicians and other study team members were very positive about the statistical aspects of the design, and the prospect of implementing such a design in practice.; the main drawback in publicly funded trials is the funding. With limited experience of the promising zone design in the UK, and not being the "normal" application they may receive, it is unknown how a panel might react to seeing a design where the final sample size (and therefore cost) is not known at the start of the trial. However, the incorporation of a futility boundary may offer the value for money design that may sway a panel in favour of the SSR design.

Overall, the research in this thesis has achieved what was set out in the original aims of the thesis, and provides practical recommendations to be considered when planning a future trial with uSSR. The promising zone design with a futility analysis, using a CP calculation with an 80% optimistic limit is suitable, even when the observed effect is much smaller than planned. Therefore this design is strongly recommended, with an interim analysis of approximately 60% through the trial duration.

References

WHO (2018). *Health Topics: Clinical Trials*. URL: http://www.who.int/topics/clinical%7B%5C_%7Dtrials/en/ (visited on 06/08/2018).

Sackett, D., W. Rosenberg, J. Gray, R. Haynes, and W. Richardson (1996). "Evidence based medicine: What it is and what it isn't - It's about integrating individual clinical expertise and the best external evidence". In: *British medical Journal* 312.1, pp. 71–72. URL: http://discovery.ucl.ac.uk/1315238/.

Roberts, C. and D. J. Torgerson (1999). "Baseline imbalance in randomised controlled trials". In: *Bmj* 319.7203, p. 185. URL: http://www.bmj.com/content/319/7203/185.1.abstract.

Shore, B. J., A. Y. Nasreddine, and M. S. Kocher (2012). "Overcoming the funding challenge: The cost of randomized controlled trials in the next decade". In: *Journal of Bone and Joint Surgery - Series A* 94.SUPPL. 1, pp. 101–106.

Collier, R. (2009). "Rapidly rising clinical trial costs worry researchers." In: *CMAJ : Canadian Medical Association Journal* 180.3, pp. 277–278.

Lindsey, H. (2009). "NerveCenter: Events, people, and issues in academic neurology". In: *Annals of Neurology* 65.3, pp. 9–11.

Noordzij, M., F. W. Dekker, C. Zoccali, and K. J. Jager (2011). "Kidney Disease and Population Health: Sample size calculations". In: *Nephron - Clinical Practice* 118.4, pp. 319–323.

Chen, J., D. L. DeMets, and G. Lan (2004). *Increasing the sample size when the unblinded interim result is promising*.

Herbert, R. D. (2000). "How to estimate treatment effects from reports of clinical trials. II: Dichotomous outcomes". In: *Australian Journal of Physiotherapy* 46.4, pp. 309–313. URL: http://dx.doi.org/10.1016/S0004-9514(14)60334-2.

Chow, S. C. and R. Corey (2011). “Benefits, challenges and obstacles of adaptive clinical trial designs”. In: *Orphanet Journal of Rare Diseases* 6.1, p. 79. URL: `http://www.ojrd.com/content/6/1/79`.

Bhatt, D. L. and C. Mehta (2016). “Adaptive Designs for Clinical Trials”. In: *New England Journal of Medicine* 375.1. Ed. by J. M. Drazen, D. P. Harrington, J. J. McMurray, J. H. Ware, and J. Woodcock, pp. 65–74. URL: `http://www.nejm.org/doi/10.1056/NEJMra1510061`.

Gallo, P., C. Chuang-Stein, V. Dragalin, B. Gaydos, M. Krams, and J. Pinheiro (2006). “Adaptive designs in clinical drug development - An Executive Summary of the PhRMA Working Group”. In: *Journal of Biopharmaceutical Statistics* 16.3, pp. 275–283.

Mehta, C. and S. Pocock (Dec. 2011). “Adaptive increase in sample size when interim results are promising: A practical guide with examples”. In: *Statistics in Medicine* 30.28, pp. 3267–3284. URL: `http://doi.wiley.com/10.1002/sim.4102`.

Lachin, J. M. (2005). “A review of methods for futility stopping based on conditional power”. In: *Statistics in Medicine* 24.18, pp. 2747–2764.

Glimm, E. (2012). “Comments on Adaptive increase in sample size when interim results are promising”. In: *Statistics in Medicine* 31, pp. 98–99.

Jennison, C. and B. W. Turnbull (2015). “Adaptive sample size modification in clinical trials: Start small then ask for more?” In: *Statistics in Medicine*.

Pilz, M., K. Kunzmann, C. Herrmann, G. Rauch, and M. Kieser (2019). “A variational approach to optimal two-stage designs”. In: *Statistics in Medicine* 38.21, pp. 4159–4171.

Laterre, P. F. and B. François (2015). “Strengths and limitations of industry vs. academic randomized controlled trials”. In: *Clinical Microbiology and Infection* 21.10, pp. 906–909. URL: `http://dx.doi.org/10.1016/j.cmi.2015.07.004`.

Mossialos, E., M. Mrazek, and T. Walley, eds. (2005). *Regulating Pharmaceuticals in Europe: Striving for Efficiency, Equity and Quality*. Vol. 14. 3. Open University Press, pp. 227–228.

Altman, D. (1980). “Statistics and ethics in medical research: III How large a sample?” In: *Bmj* 281.6251, pp. 1336–1338. URL: http://www.bmj.com/cgi/doi/10.1136/bmj.281.6251.1336.

Streiner, D. L. (2007). “Alternatives to placebo-controlled trials.” In: *The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques* 34 Suppl 1, S37–41. URL: http://www.ncbi.nlm.nih.gov/pubmed/17469680.

Togo (2016). “Is the Use of Placebos Ethically Justified and are there Any Alternatives that can be of Equal Benefit in Advancing Current Medical Knowledge?” In: *Journal of Clinical Research & Bioethics* 07.02, pp. 7–9. URL: https://www.omicsonline.org/open-access/is-the-use-of-placebos-ethically-justified-and-are-there-any-alternatives-that-can-be-of-equal-benefit-in-advancing-current-medica-2155-9627-1000267.php?aid=70915.

Chen, Z., Y. Zhao, Y. Cui, and J. Kowalski (2012). “Methodology and application of adaptive and sequential approaches in contemporary clinical trials”. In: *Journal of Probability and Statistics* 2012.

Schulz and Grimes (2005). “Epidemiology 1: Sample size calculations in randomised trials: mandatory and mystical.” In: *Lancet (London, England)* 365.9467, pp. 1348–53. URL: http://www.ncbi.nlm.nih.gov/pubmed/15823387.

Teare, M., M. Dimairo, N. Shephard, A. Hayman, A. Whitehead, and S. J. Walters (2014). “Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study”. In: *Trials* 15.1, p. 264. URL: http://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-15-264.

Viele, K., S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, S. Micallef, S. Roychoudhury, and L. Thompson (2013). “Use of historical control data for assessing treatment effects in clinical trials”. In: *Pharmaceutical Statistics* 13.1, pp. 41–54. URL: http://doi.wiley.com/10.1002/pst.1589%7B%5C%%7D5Cnpapers3://publication/doi/10.1002/pst.1589.

Shih and Long (1998). “Blinded sample size re-estimation with unequal variances and center effects in clinical trials”. In: *Communications in Statistics - Theory and Methods* 27.2, pp. 395–408.

Biau, D. J., B. M. Jolles, and R. Porcher (2010). “P value and the theory of hypothesis testing: An explanation for new researchers”. In: *Clinical Orthopaedics and Related Research* 468.3, pp. 885–892.

Akobeng, A. (2016). “Understanding type I and type II errors, statistical power and sample size”. In: *Acta Paediatrica* 105.6, pp. 605–609. URL: `http://doi.wiley.com/10.1111/apa.13384`.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. 4th. Edinburgh Oliver & Boyd.

Sterne, J. A., G. D. Smith, and D. R. Cox (2001). “Sifting the evidence—what’s wrong with significance tests?” In: *Bmj* 322.7280, p. 226.

Julious, S. A. (2004). “Tutorial in biostatistics: Sample sizes for clinical trials with Normal data”. In: *Statistics in Medicine* 23.12, pp. 1921–1986.

Sedgwick, P. (2014). “Pitfalls of statistical hypothesis testing: Type I and type II errors”. In: *BMJ (Online)* 349.July, pp. 2–3.

Cook, J., J. Hislop, T. Adewuyi, K. Harrild, D. Altman, C. Ramsay, C. Fraser, B. Buckley, P. Fayers, I. Harvey, A. Briggs, J. Norrie, D. Fergusson, I. Ford, and L. Vale (Jan. 2014). “Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review”. In: *Health Technology Assessment* 18.28. URL: `https://www.journalslibrary.nihr.ac.uk/hta/hta18280/`.

Cook, J., J. Hislop, D. Altman, P. Fayers, A. Briggs, C. Ramsay, J. Norrie, I. Harvey, B. Buckley, D. Fergusson, I. Ford, and L. Vale (2015). “Specifying the target difference in the primary outcome for a randomised controlled trial: Guidance for researchers”. In: *Trials* 16.1, pp. 1–7.

Biau, D. J., S. Kernéis, and R. Porcher (2008). “Statistics in brief: The importance of sample size in the planning and interpretation of medical research”. In: *Clinical Orthopaedics and Related Research* 466.9, pp. 2282–2288.

Schulz, K., D. Altman, and D. Moher (Dec. 2010). “CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials”. In: *BMC Medicine* 8.1, p. 18. URL: http://bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-8-18.

Yusuf, S., R. Collins, and R. Peto (Oct. 1984). “Why do we need some large, simple randomized trials?” In: *Statistics in Medicine* 3.4, pp. 409–420. URL: http://doi.wiley.com/10.1002/sim.4780030421.

Fukunaga, S., M. Kusama, and S. Ono (2014). “The effect size, study design, and development experience in commercially sponsored studies for new drug applications in approved drugs”. In: *SpringerPlus* 3.1, pp. 1–9.

Dantzig, G. B. (1940). “On the Non-Existence of Tests of "Student’s" Hypothesis Having Power Functions Independent of σ”. In: *The Annals of Mathematical Statistics* 11.2, pp. 186–192. URL: http://www.jstor.org/stable/2235875.

Stein, C. (1945). “A Two-Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance”. In: *The Annals of Mathematica* 16.3, pp. 243–258.

Mills, E., A. W. Chan, P. Wu, A. Vail, G. Guyatt, and D. Altman (2009). “Design, analysis, and presentation of crossover trials”. In: *Trials* 10, pp. 1–6.

Li, T., T. Yu, B. S. Hawkins, and K. Dickersin (2015). “Design, analysis, and reporting of crossover trials for inclusion in a meta-analysis”. In: *PLoS ONE* 10.8, pp. 1–12.

Moser, B. K. and S. Halabi (2015). “Sample size requirements and study duration for testing main effects and interactions in completely randomized factorial designs when time to event is the outcome”. In: *Communications in Statistics - Theory and Methods* 44.2, pp. 275–285.

Lesaffre, E. (2008). “Superiority, equivalence, and non-inferiority trials”. In: *Bulletin of the NYU Hospital for Joint Diseases* 66.2, pp. 150–154.

Hahn, S. (2012). “Understanding noninferiority trials”. In: *Korean Journal of Pediatrics* 55.11, p. 403. URL: https://synapse.koreamed.org/DOIx.php?id=10.3345/kjp.2012.55.11.403.

Greene, C. and L. Morland (2008). “Noninferiority and equivalence designs: Issues and implications for mental health research”. In: *J Trauma Stress* 21.5, pp. 433–439. URL: http://onlinelibrary.wiley.com/doi/10.1002/jts.20367/abstract.

Flight, L. and S. A. Julious (2016a). “Practical guide to sample size calculations: Superiority trials”. In: *Pharmaceutical Statistics* 15.1, pp. 75–79.

— (Jan. 2016b). “Practical guide to sample size calculations: non-inferiority and equivalence trials”. In: *Pharmaceutical Statistics* 15.1, pp. 68–74. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84954379510%7B%5C&%7DpartnerID=40%7B%5C&%7Dmd5=e3879cb601d09dfd5fafb2dbd72bc41c%20http://doi.wiley.com/10.1002/pst.1709.

Julious, S. A. (2009). *Sample Sizes for Clinical Trials*. 1st ed. Chapman and Hall CRC.

Halpern, S. D. (2005). “Adding nails to the coffin of underpowered trials”. In: *Journal of Rheumatology* 32.11, pp. 2065–2066.

Edwards, S. J., R. J. Lilford, D. Braunholtz, and J. Jackson (1997). “Why ’underpowered’ trials are not necessarily unethical”. In: *Lancet* 350.9080, pp. 804–807.

Halpern, S. D., J. H. T. Karlawish, and J. a. Berlin (2002). “The Continuing Unethical Conduct of Underpowered Clinical Trials”. In: *JAMA* 288.3.

Turner, R. M., S. M. Bird, and J. P. T. Higgins (2013). “The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews”. In: *PLoS ONE* 8.3, pp. 1–8.

Griffiths, M. (Nov. 1997). ““Underpowered” trials”. In: *The Lancet* 350.9088, p. 1406. URL: http://linkinghub.elsevier.com/retrieve/pii/S0140673605651896.

Christy, C.-S., A. Keaven, G. Paul, and C. Sylva (2006). “Sample Size Reestimation : A Review and Recommendations”. In: *Drug Information Journal* 40, pp. 475–484.

Bauer, P., F. Bretz, V. Dragalin, F. König, and G. Wassmer (2016a). “Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls”. In: *Statistics in Medicine* 35.3, pp. 325–347.

Pallmann, P., A. W. Bedding, B. Choodari-Oskooei, M. Dimairo, L. Flight, L. V. Hampson, J. Holmes, A. P. Mander, L. Odondi, M. R. Sydes, S. S. Villar, J. M. Wason, C. J. Weir,

G. M. Wheeler, C. Yap, and T. Jaki (2018). “Adaptive designs in clinical trials: Why use them, and how to run and report them”. In: *BMC Medicine* 16.1, pp. 1–15.

FDA (2010). *Draft Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*. Tech. rep. URL: `https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf`.

— (2016). *Guidance for Industry: Adaptive Designs for Medical Device Clinical Studies*. Tech. rep. URL: `https://www.fda.gov/downloads/medicaldevices/deviceregulationandguida guidancedocuments/ucm446729.pdf`.

Mahajan, R. and K. Gupta (2010). “Adaptive design clinical trials: Methodology, challenges and prospect”. In: *Indian Journal of Pharmacology* 42.4, p. 201. URL: `http://www.ijp-online.com/text.asp?2010/42/4/201/68417`.

Lin, M., S. Lee, B. Zhen, J. Scott, A. Horne, G. Solomon, and E. Russek-Cohen (2016). “CBER’s Experience With Adaptive Design Clinical Trials”. In: *Therapeutic Innovation & Regulatory Science* 50.2, pp. 195–203. URL: `http://journals.sagepub.com/doi/10.1177/2168479015604181`.

Hatfield, I., A. Allison, L. Flight, S. A. Julious, and M. Dimairo (Dec. 2016). “Adaptive designs undertaken in clinical research: a review of registered clinical trials”. In: *Trials* 17.1, p. 150. URL: `http://dx.doi.org/10.1186/s13063-016-1273-9%20http://www.trialsjournal.com/content/17/1/150`.

Coffey, C. S., B. Levin, C. Clark, C. Timmerman, J. Wittes, P. Gilbert, and S. Harris (2012). “Overview, hurdles, and future work in adaptive designs: Perspectives from a National Institutes of Health-funded workshop”. In: *Clinical Trials* 9.6, pp. 671–680.

Kairalla, J., C. S. Coffey, M. Thomann, and K. E. Muller (2012). “Adaptive trial designs: a review of barriers and opportunities.” In: *Trials* 13.1, p. 145. URL: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3519822%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Dtool=pmcentrez%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Drendertype=abstract`.

Vandemeulebroecke, M. (2008). “Group sequential and adaptive designs - A review of basic concepts and points of discussion”. In: *Biometrical Journal* 50.4, pp. 541–557.

Chang, M. and J. Balser (2016). “Adaptive Design -Recent Advancement in Clinical Trials Adaptive Design Methods in Clinical Trials”. In: *J Bioanal Biostat* 1.11, p. 14. URL: `http://www.avensonline.org/wp-content/uploads/JBABS-01-0003.pdf`.

Maca, J., V. Dragalin, and P. Gallo (2014). “Adaptive Clinical Trials: Overview of Phase III Designs and Challenges”. In: *Therapeutic Innovation and Regulatory Science* 48.1, pp. 31–40.

Wassmer, G. and W. Brannath (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer.

Mauer, M., L. Collette, and J. Bogaerts (June 2012). “Adaptive designs at European Organisation for Research and Treatment of Cancer (EORTC) with a focus on adaptive sample size re-estimation based on interim-effect size”. In: *European Journal of Cancer* 48.9, pp. 1386–1391. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0959804911010732`.

Levin, G. P., S. C. Emerson, and S. S. Emerson (2014). “An evaluation of inferential procedures for adaptive clinical trial designs with pre-specified rules for modifying the sample size”. In: *Biometrics* 70.3, pp. 556–567.

Koh, W. J. H. (2017). “Adaptive designs in the time to event setting: The potential for benefit and risk.” In: URL: `http://0-search.ebscohost.com.liucat.lib.liu.edu/login.aspx?direct=true%7B%5C&%7Ddb=psyh%7B%5C&%7DAN=2016-58395-008%7B%5C&%7Dsite=ehost-live%7B%5C&%7Dscope=site`.

Bowden, J. and A. Mander (2014). “A review and re-interpretation of a group-sequential approach to sample size re-estimation in two-stage trials”. In: *Pharmaceutical Statistics* 13.3, pp. 163–172.

Wang, S. J., H. M. James Hung, and R. O’Neill (2012). “Paradigms for adaptive statistical information designs: Practical experiences and strategies”. In: *Statistics in Medicine* 31.25, pp. 3011–3023.

Huskins, W. C., V. G. Fowler, and S. Evans (2018). “Adaptive Designs for Clinical Trials: Application to Healthcare Epidemiology Research”. In: *Clinical Infectious Diseases*.

Proschan, M. (2005). “Two-stage sample size re-estimation based on a nuisance parameter: A review”. In: *Journal of Biopharmaceutical Statistics* 15.4, pp. 559–574.

Wittes, J. and E. Brittain (1990). “The role of internal pilot studies in increasing the efficiency of clinical trials”. In: *Statistics in Medicine* 9.1-2, pp. 65–72.

Mütze, T., H. Schmidli, and T. Friede (2018). “Sample size re-estimation incorporating prior information on a nuisance parameter”. In: *Pharmaceutical Statistics* 17.2, pp. 126–143. arXiv: arXiv:1703.06957v1.

Pritchett, Y. L., S. Menon, O. Marchenko, Z. Antonijevic, E. Miller, M. Sanchez-Kam, C. C. Morgan-Bouniol, H. Nguyen, and W. R. Prucka (2015). “Sample Size Re-estimation Designs In Confirmatory Clinical Trials—Current State, Statistical Considerations, and Practical Guidance”. In: *Statistics in Biopharmaceutical Research* 7.4, pp. 309–321.

Gould, L. (2001). “Sample size re-estimation: recent developments and practical considerations”. In: *Statistics in Medicine* 20.17-18, pp. 2625–2643.

Stein, C. (1950). “Unbiased Estimates with Minimum Variance”. In: *Annals of Mathematical Statistics* 21.3, pp. 406–415.

Birkett, M. A. and S. J. Day (1994). “Internal pilot studies for estimating sample size”. In: *Statistics in Medicine* 13.23-24, pp. 2455–2463.

Gould, L. and W. J. Shih (Jan. 1992). “Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance”. In: *Communications in Statistics - Theory and Methods* 21.10, pp. 2833–2853. URL: http://www.tandfonline.com/doi/abs/10.1080/03610929208830947.

Proschan, M. (2006). “Adaptive Sample Size Methods”. In: *Statistical Monitoring of Clinical Trials*. New York, NY: Springer New York, pp. 185–211. URL: http://link.springer.com/10.1007/978-0-387-44970-8%7B%5C_%7D11.

Friede, T. and M. Kieser (2002). “On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation”. In: 176.March 2001, pp. 165–176.

Gould, L. and W. J. Shih (2005). “Letter to the Editor: On the inappropriateness of an EM algorithm based procedure for blinded”. In: *Statistics in Medicine*, pp. 147–154.

Waksman, J. A. (2007). “Assessment of the Gould – Shih procedure for sample size re-estimation”. In: *Pharmaceutical Statistics* November 2006, pp. 53–65.

Kieser, M. and T. Friede (Dec. 2003). “Simple procedures for blinded sample size adjustment that do not affect the type I error rate”. In: *Statistics in Medicine* 22.23, pp. 3571–3581. URL: `http://doi.wiley.com/10.1002/sim.1585`.

Friede, T. and M. Kieser (2004). “Sample size recalculation for binary data in internal pilot study designs”. In: *Pharmaceutical Statistics* 3.4, pp. 269–279.

Wachtlin, D. and M. Kieser (2013). “Blinded Sample Size Recalculation in Longitudinal Clinical Trials Using Generalized Estimating Equations”. In: *Therapeutic Innovation & Regulatory Science* 47.4, pp. 460–467. URL: `http://journals.sagepub.com/doi/10.1177/2168479013486658`.

Friede, T. and H. Schmidli (2010). “Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis”. In: *Statistics in Medicine* 29.10, pp. 1145–1156.

Lewis, J. A. (Aug. 1999). “Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline”. In: *Statistics in Medicine* 18.15, pp. 1903–1942. URL: `http://doi.wiley.com/10.1002/%7B%5C%%7D28SICI%7B%5C%%7D291097-0258%7B%5C%%7D2819990815%7B%5C%%7D2918%7B%5C%%7D3A15%7B%5C%%7D3C1903%7B%5C%%7D3A%7B%5C%%7D3AAID-SIM188%7B%5C%%7D3E3.0.CO%7B%5C%%7D3B2-F`.

Bauer, P. and K. Kohne (1994). “Evaluation of Experiments with Adaptive Interim Analyses”. In: *Biometrics* 50.4, pp. 1029–1041.

Lehmacher, W. and G. Wassmer (1999). “Adaptive Sample Size Calculations in Group Sequential Trials”. In: 55.4, pp. 1286–1290.

European Medicines Agency (2007). *Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned With an Adaptive Design*. Tech. rep.

Gaffney, M. and J. H. Ware (Sept. 2017). “An evaluation of increasing sample size based on conditional power”. In: *Journal of Biopharmaceutical Statistics* 27.5, pp. 797–808. URL: `http://www.tandfonline.com/action/journalInformation?journalCode=`

lbps20%20https://www.tandfonline.com/doi/full/10.1080/10543406.2017.1289943.

US Food and Drug Administration (FDA) (2019). “Adaptive Designs for Clinical Trials of Drugs and Biologics FDA Guidance for Industry”. In: *Biostatistics* November, pp. 1–33. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.

Wan, H., S. Ellenberg, and K. Anderson (2015). “Stepwise two-stage sample size adaptation”. In: *Statistics in Medicine* 34.1, pp. 27–38.

Denne, J. S. (2001). “Sample size recalculation using conditional power”. In: *Statistics in Medicine* 20.17-18, pp. 2645–2660.

Sully, B. G. O., S. A. Julious, and J. Nicholl (2014). “An investigation of the impact of futility analysis in publicly funded trials”. In: *Trials* 15.1, pp. 1–9. URL: Trials.

Herson, J., M. Buyse, and J. T. Wittes (2012). *Designs for Clinical Trials*. URL: http://link.springer.com/10.1007/978-1-4614-0140-7.

Mehta, C. R. and A. S. Tsiatis (2001). “Flexible sample size considerations using information based interim monitoring”. In: *Drug Information Journal* 35, pp. 1095–1112.

Liu, Y. and M. Hu (2016). “Testing multiple primary endpoints in clinical trials with sample size adaptation”. In: *Pharmaceutical Statistics* 15.1, pp. 37–45.

Wald, A. (1947). *Sequential Analysis*. 1st. John Wiley and Sons.

Anderson, T. W. (1960). “A Modification of the Sequential Probability Ratio Test to Reduce the Sample Size”. In: *Annals of Mathematical Statistics* 31.1, pp. 165–197.

Whitehead, J. and D. Jones (1979). “The Analysis of Sequential Clinical Trials”. In: *Biometrika* 66.3, pp. 443–452.

Bassler, D., V. M. Montori, M. Briel, P. Glasziou, and G. Guyatt (2008). “Early stopping of randomized clinical trials for overt efficacy is problematic”. In: *Journal of Clinical Epidemiology* 61.3, pp. 241–246.

Hampson, L. V. and C. Jennison (2012). “Group sequential tests for delayed responses (with discussion)”. In: *J. R. Statist. Soc. B* 75.1, pp. 3–54.

Pocock, S. (1977). “Group Sequential Methods in the Design and Analysis of Clinical”. In: *Biometrika* 64.2, pp. 191–199.

Proschan, M. (1999). “Biometrika Trust Properties of Spending Function Boundaries”. In: *Biometrika* 86.2, pp. 466–473.

Chow, S. C., H. Wang, and J. Shao (2008). *Sample size calculations in clinical research*. Second Edi. Chapman & Hall/CRC.

Pampallona, S. and A. A. Tsiatis (1994). “Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis”. In: *Journal of Statistical Planning and Inference* 42.1-2, pp. 19–35.

O’Brien, P. and T. Fleming (1979). “A Multiple Testing Procedure for Clinical Trials”. In: *Biometrics* 35.3, pp. 549–556.

Proschan, M. and S. Hunsberger (Dec. 1995). “Designed Extension of Studies Based on Conditional Power”. In: *Biometrics* 51.4, p. 1315. URL: `https://www.jstor.org/stable/2533262?origin=crossref`.

Posch, M. and P. Bauer (1999). “Adaptive two stage designs and the conditional error function”. In: *Biometrical Journal* 41.6, pp. 689–696.

Li, G., W. J. Shih, T. Xie, and J. Lu (2002). “A sample size adjustment procedure for clinical trials based on conditional power.” In: *Biostatistics* 3.2, pp. 277–287. URL: `http://www.ncbi.nlm.nih.gov/pubmed/12933618`.

Westberg, M. (1985). “Combining Independent Statistical Tests”. In: *Journal of the Royal Statistical Society* 34.3, p. 287.

Wassmer, G. (1998). “A Comparison of Two Methods for Adaptive Interim Analyses in Clinical Trials”. In: *Biometrics* 54.2, pp. 696–705.

Bauer, P. and J. Rohmel (1995). “An adaptive method for establishing a dose-response relationship”. In: *Statistics in Medicine* 14.14, pp. 1595–1607.

Fisher, L. D. (1998). “SELF-DESIGNING CLINICAL TRIALS”. In: 1562.October 1997, pp. 1551–1562.

Jennison, C. and B. W. Turnbull (2003). “Mid-course sample size modification in clinical trials based on the observed treatment effect”. In: *Statistics in Medicine* 22.6, pp. 971–993.

Hedges, L. and I. Olkin (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Cui, L., H. M. Hung, and S. J. Wang (1999). “Modification of sample size in group sequential clinical trials”. In: *Biometrics* 55.3, pp. 853–857.

Burman, C. F. and C. Sonesson (2006). “Are flexible designs sound?” In: *Biometrics* 62.3, pp. 664–669.

Basu, D. (1969). “Role of the Sufficiency and Likelihood Principles in Sample Survey Theory Stable URL : https”. In: *The Indian Journal of Statistics* 31.4, pp. 441–454.

Elsäßer, A., J. Regnstrom, T. Vetter, F. Koenig, R. J. Hemmings, M. Greco, M. Papaluca-amati, and M. Posch (2014). “Adaptive clinical trial designs for European marketing authorization : a survey of scientific advice letters from the European Medicines Agency”. In: pp. 1–10.

Shih, W. J., G. Li, and Y. Wang (2016). “Methods for flexible sample-size design in clinical trials: Likelihood, weighted, dual test, and promising zone approaches”. In: *Contemporary Clinical Trials*.

Gao, P., J. H. Ware, and C. Mehta (2008). “Sample size re-estimation for adaptive sequential design in clinical trials”. In: *Journal of Biopharmaceutical Statistics* 18.6, pp. 1184–1196.

Hung, H. M. J., S.-j. Wang, P. Yang, K. Jin, and J. Lawrence (2016a). “Statistical challenges in a regulatory review of cardiovascular and CNS clinical trials”. In: 26.1, pp. 37–43.

Mehta, C. and L. Liu (Feb. 2016a). “An objective re-evaluation of adaptive sample size re-estimation: commentary on ‘Twenty-five years of confirmatory adaptive designs’”. In: *Statistics in Medicine* 35.3, pp. 350–358. URL: `http://doi.wiley.com/10.1002/sim.6614`.

Brannath, W. and P. Bauer (2004). “Optimal conditional error functions for the control of conditional power”. In: *Biometrics* 60.3, pp. 715–723.

Hawkins, D. and R. Wagers (1982). “Online Bibliographic Search Strategy Development”. In: *Online*, pp. 12–19.

Schlosser, R. W., O. Wendt, K. L. Angermeier, and M. Shetty (2005). “Searching for evidence in augmentative and alternative communication: Navigating a scattered literature”. In: *AAC: Augmentative and Alternative Communication* 21.4, pp. 233–255.

Bauer, P. and F. Koenig (2006). “The reassessment of trial perspectives from interim data - A critical view”. In: *Statistics in Medicine* 25.1, pp. 23–36.

Pepe, M. S. and G. L. Anderson (1992). “Two-stage Experimental Designs: Early Stopping with a Negative Result”. In: *Applied Statistics* 41.1, pp. 181–190.

Lan, K. K. G. and D. C. Trost (1997). “Estimation of Parameters and Sample Size Re-Estimation”. In: *Annual meeting, American Statistical Association: Biopharmaceutical Section;* Anaheim; CA, pp. 48–51.

Liu, L., S. Hsiao, and C. R. Mehta (2017a). “Efficiency Considerations for Group Sequential Designs with Adaptive Unblinded Sample Size Re-assessment”. In: *Statistics in Biosciences*, pp. 1–15.

Committee on National Statistics (2010). “Appendix A, Clinical Trials: Overview and Terminology”. In: *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academy of Sciences. Chap. Appendix A. URL: https://www.ncbi.nlm.nih.gov/books/NBK209903/.

Deley, M.-C. L., K. V. Ballman, J. Marandet, and D. Sargent (June 2012). “Taking the long view: how to design a series of Phase III trials to maximize cumulative therapeutic benefit”. In: *Clinical Trials: Journal of the Society for Clinical Trials* 9.3, pp. 283–292. arXiv: NIHMS150003. URL: http://journals.sagepub.com/doi/10.1177/1740774512443430.

Togo, K. and M. Iwasaki (2013). “Optimal timing for interim analyses in clinical trials”. In: *Journal of Biopharmaceutical Statistics* 23.5, pp. 1067–1080.

Liu, Y. and P. Lim (July 2017b). “Sample size increase during a survival trial when interim results are promising”. In: *Communications in Statistics - Theory and Methods* 46.14, pp. 6846–6863. URL: https://www.tandfonline.com/doi/full/10.1080/03610926.2015.1137596.

Levin, G. P., S. C. Emerson, and S. S. Emerson (2013). “Adaptive clinical trial designs with pre-specified rules for modifying the sample size: Understanding efficient types of adaptation”. In: *Statistics in Medicine* 32.8, pp. 1259–1275.

Chen, J., S. Yuan, and X. Li (2018a). “Statistical inference following sample size adjustment based on the 50%-conditional-power principle”. In: *Journal of Biopharmaceutical Statistics* 28.3, pp. 575–587. URL: https://doi.org/10.1080/10543406.2017.1372766.

Antonijevic, Z. (2016). “The Impact of Adaptive Design on Portfolio Optimization”. In: *Therapeutic Innovation and Regulatory Science*.

Wang, S.-J., W. Brannath, M. Brückner, H. M. James Hung, and A. Koch (2013). “Unblinded Adaptive Statistical Information Design Based on Clinical Endpoint or Biomarker”. In: *Statistics in Biopharmaceutical Research* 5.4, pp. 293–310. URL: http://www.tandfonline.com/doi/abs/10.1080/19466315.2013.791639.

Mehta, C. R. (2012). *Designs for Clinical Trials*. URL: http://link.springer.com/10.1007/978-1-4614-0140-7.

Brannath, W., G. Gutjahr, and P. Bauer (2012). “Probabilistic foundation of confirmatory adaptive designs”. In: *Journal of the American Statistical Association* 107.498, pp. 824–832.

Broberg, P. (2013). “Sample size re-assessment leading to a raised sample size does not inflate type i error rate under mild conditions”. In: *BMC Medical Research Methodology* 13.94.

Freidlin, B. and E. L. Korn (2017). “Sample size adjustment designs with time-to-event outcomes: A caution”. In: *Clinical Trials*.

Mehta, C. R. (2013). “Adaptive clinical trial designs with pre-specified rules for modifying the sample size: A different perspective”. In: *Statistics in Medicine* 32.8, pp. 1276–1279.

Posch, M. and P. Bauer (2013). “Adaptive Budgets in Clinical Trials”. In: *Statistics in Biopharmaceutical Research* 5.4, pp. 282–292.

Uozumi, R. and C. Hamada (2017). “Adaptive Seamless Design for Establishing Pharmacokinetic and Efficacy Equivalence in Developing Biosimilars”. In: *Therapeutic Innovation and Regulatory Science*.

Bayar, M. A., G. Le Teuff, S. Michiels, D. J. Sargent, and M. C. Le Deley (2016). “New insights into the evaluation of randomized controlled trials for rare diseases over a long-term research horizon: a simulation study”. In: *Statistics in Medicine* 35.19, pp. 3245–3258.

Hung, H. M. J., S. J. Wang, and P. Yang (2014). “Some challenges with statistical inference in adaptive designs”. In: *Journal of Biopharmaceutical Statistics* 24.5, pp. 1059–1072. URL: `http://dx.doi.org/10.1080/10543406.2014.925911%20https://doi.org/10.1080/10543406.2014.925911`.

Hung, H. M. J. (2016b). “Rejoinder to Dr . Cyrus R . Mehta”. In: *Journal of Biopharmaceutical Statistics* 26.2, p. 405.

Turnbull, B. W. (2017). “Adaptive designs from a Data Safety Monitoring Board perspective: Some controversies and some case studies”. In: *Clinical Trials* 14.5, pp. 462–469.

Müller, H. and H. Schäfer (2001). “Adaptive group sequential design for clinical trials: Combining the advantages of adaptive and of classic group sequential approaches.” In: *Biometrics* 57.1, pp. 886–891.

Chen, C., K. Anderson, D. V. Mehrotra, E. H. Rubin, and A. Tse (2018b). “A 2-in-1 adaptive phase 2/3 design for expedited oncology drug development”. In: *Contemporary Clinical Trials* 64, pp. 238–242. URL: `https://doi.org/10.1016/j.cct.2017.09.006`.

Tamhane, A. C., Y. Wu, and C. R. Mehta (2012). “Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (II): Sample size re-estimation”. In: *Statistics in Medicine* 31.19, pp. 2041–2054.

Bauer, P., F. Bretz, V. Dragalin, F. König, and G. Wassmer (2016b). “Authors’ response to comments”. In: *Statistics in Medicine* 35.3, pp. 364–367. URL: `http://doi.wiley.com/10.1002/sim.6823`.

Mehta, C. R. (2016b). “Comments on "Some Challenges with Statistical Inference in Adaptive Designs" by Hung, Wang, and Yang”. In: *Journal of Biopharmaceutical Statistics*

26.2, pp. 402–404. URL: http://dx.doi.org/10.1080/10543406.2015.1099541%20https://doi.org/10.1080/10543406.2015.1099541.

Zhang, L., L. Cui, and B. Yang (Aug. 2016). “Optimal flexible sample size design with robust power”. In: *Statistics in Medicine* 35.19, pp. 3385–3396. URL: http://doi.wiley.com/10.1002/sim.6931.

Hsiao, S. T., L. Liu, and C. R. Mehta (2018). “Optimal promising zone designs”. In: *Biometrical Journal* September, pp. 1–12.

Gao, P., L. Liu, and C. Mehta (2013a). “Adaptive designs for noninferiority trials”. In: 55, pp. 310–321.

Senchaudhuri, E. (2015). *Why you should not power for superiority upfront - Promising Zone trials with adaptive switch.* (Visited on 09/23/2018).

Joshua Chen, Y. H., C. Li, and K. K. Gordon Lan (2015). “Sample size adjustment based on promising interim results and its application in confirmatory clinical trials”. In: *Clinical Trials.*

Menon, S., J. Massaro, M. J. Pencina, J. Lewis, and Y. C. Wang (2013). “Comparison of operating characteristics of commonly used sample size re-estimation procedures in a two-stage design”. In: *Communications in Statistics: Simulation and Computation.*

Cui, L., L. Zhang, and B. Yang (2017). “Optimal adaptive group sequential design with flexible timing of sample size determination”. In: *Contemporary Clinical Trials* 63.April, pp. 8–12. URL: http://dx.doi.org/10.1016/j.cct.2017.04.005.

Pocock, S. J., T. C. Clayton, and G. W. Stone (2015). “Challenging Issues in Clinical Trial Design Part 4 of a 4-Part Series on Statistics for Clinical Trials”. In: *Journal of the American College of Cardiology* 66.25, pp. 2886–2898.

Magirr, D., T. Jaki, F. Koenig, and M. Posch (2016). “Sample size reassessment and hypothesis testing in adaptive survival trials”. In: *PLoS ONE* 11.2, pp. 1–14.

Dimairo, M., P. Pallmann, J. Wason, S. Todd, T. Jaki, S. Julious, A. Mander, C. Weir, F. Koenig, M. Walton, J. Nicholl, E. Coates, K. Biggs, T. Hamasaki, M. Proschan, J. Scott, Y. Ando, D. Hind, and D. Altman (2020). “The Adaptive designs CONSORT Extension

(ACE) statement: A checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design". In: *The BMJ* 369.

Wang, Y., G. Li, and W. J. Shih (2010). "Estimation and Confidence Intervals for Two-Stage Sample-Size-Flexible Design with LSW Likelihood Approach". In: *Statistics in Biosciences* 2.2, pp. 180–190.

Jennison, C. and B. W. Turnbull (1984). "Repeated confidence intervals for group sequential clinical trials". In: *Controlled Clinical Trials* 5.1, pp. 33–45.

Mehta, C., P. Bauer, M. Posch, and W. Brannath (2007). "Repeated confidence intervals for adaptive group sequential trials". In: *Statistics in Biosciences* 26, pp. 5422–5433.

Jennison, C. and B. W. Turnbull (1989). "Interim Analyses: The repeated confidence interval approach". In: *Journal of the Royal Statistical Society* 51.3, pp. 305–361.

Gao, P., L. Liu, and C. Mehta (2013b). "Exact inference for adaptive group sequential designs". In: *Statistics in Medicine* 32.23, pp. 3991–4005.

Deal, J. A., A. M. Goman, M. S. Albert, M. L. Arnold, S. Burgard, T. Chisolm, D. Couper, N. W. Glynn, T. Gmelin, K. M. Hayden, T. Mosley, J. S. Pankow, N. Reed, V. A. Sanchez, A. Richey Sharrett, S. D. Thomas, J. Coresh, and F. R. Lin (2018). "Hearing treatment for reducing cognitive decline: Design and methods of the Aging and Cognitive Health Evaluation in Elders randomized controlled trial". In: *Alzheimer's and Dementia: Translational Research and Clinical Interventions* 4, pp. 499–507. URL: https://doi.org/10.1016/j.trci.2018.08.007.

Colwell, J. C., J. Pittman, R. Raizman, and G. Salvadalena (2018). "A Randomized Controlled Trial Determining Variances in Ostomy Skin Conditions and the Economic Impact (ADVOCATE Trial)". In: *Journal of Wound, Ostomy and Continence Nursing* 45.1, pp. 37–42. URL: http://insights.ovid.com/crossref?an=00152192-201801000-00007.

Tripathy, D., S. Tolaney, A. Seidman, C. Anders, N. Ibrahim, H. Rugo, C. Twelves, V. Dieras, V. Muller, A. Hannah, M. Tagliaferri, and J. Cortes Castan (2018). "Annals of Oncology". In: *ATTAIN: Phase III study of etirinotecan pegol (EP) vs treatment of* *physicians choice (TCP) in patients with metastatic breast cancer*. Vol. 29. October, p. 118.

Miller, E., P. Gallo, W. He, L. A. Kammerman, K. Koury, J. Maca, Q. Jiang, M. K. Walton, C. Wang, K. Woo, C. Fuller, and Y. Jemiai (2017). “DIA’s Adaptive Design Scientific Working Group (ADSWG): Best Practices Case Studies for “Less Well-understood” Adaptive Designs”. In: *Therapeutic Innovation and Regulatory Science*.

Schindler, C., A. L. Birkenfeld, M. Hanefeld, U. Schatz, C. Köhler, M. Grüneberg, D. Tschöpe, M. Blüher, C. Hasslacher, and S. R. Bornstein (Feb. 2018). “Intravenous Ferric Carboxymaltose in Patients with Type 2 Diabetes Mellitus and Iron Deficiency: CLEVER Trial Study Design and Protocol”. In: *Diabetes Therapy* 9.1, pp. 37–47. URL: `http://link.springer.com/10.1007/s13300-017-0330-z`.

Hirakawa, A., T. Hatakeyama, D. Kobayashi, C. Nishiyama, A. Kada, T. Kiguchi, T. Kawamura, and T. Iwami (2018). “Real-time feedback, debriefing, and retraining system of cardiopulmonary resuscitation for out-of-hospital cardiac arrests: A study protocol for a cluster parallel-group randomized controlled trial”. In: *Trials* 19.1, pp. 1–9.

White, W. B., C. P. Cannon, S. R. Heller, S. E. Nissen, R. M. Bergenstal, G. L. Bakris, A. T. Perez, P. R. Fleck, C. R. Mehta, S. Kupfer, C. Wilson, W. C. Cushman, and F. Zannad (Oct. 2013). “Alogliptin after Acute Coronary Syndrome in Patients with Type 2 Diabetes”. In: *New England Journal of Medicine* 369.14, pp. 1327–1335. URL: `http://www.nejm.org/doi/10.1056/NEJMoa1305889`.

Churilov, L., H. Ma, B. C. Campbell, S. M. Davis, and G. A. Donnan (2018). “Statistical Analysis Plan for EXtending the time for Thrombolysis in Emergency Neurological Deficits (EXTEND) trial”. In: *International Journal of Stroke* 0.0, pp. 1–8.

Campbell, B., P. J. Mitchell, B. Yan, M. W. Parsons, S. Christensen, L. Churilov, R. J. Dowling, H. Dewey, M. Brooks, F. Miteff, C. Levi, M. Krause, T. J. Harrington, K. C. Faulder, B. S. Steinfort, T. Kleinig, R. Scroop, S. Chryssidis, A. Barber, A. Hope, M. Moriarty, B. Mcguinness, A. A. Wong, A. Coulthard, T. Wijeratne, A. Lee, J. Jannes, J. Leyden, T. G. Phan, W. Chong, M. E. Holt, R. V. Chandra, C. F. Bladin, M. Badve, H. Rice, L. D. Villiers, H. Ma, P. M. Desmond, G. A. Donnan, and S. M. Davis (2014).

"Protocols A multicenter , randomized , controlled study to investigate EXtending the time for Thrombolysis in Emergency Neurological Deficits with Intra-Arterial therapy (EXTEND-IA)". In: 9.January, pp. 126–132.

Campbell, B., P. Mitchell, L. Churilov, N. Yassi, T. Kleinig, B. Yan, R. J. Dowling, S. Bush, H. Dewey, V. Thijs, M. Simpson, M. Brooks, H. Asadi, T. Wu, D. Shah, T. Wijeratne, T. Ang, F. Miteff, C. Levi, M. Krause, T. Harrington, K. Faulder, B. Steinfort, P. Bailey, H. Rice, L. de Villiers, R. Scroop, W. Collecutt, A. Wong, A. Coulthard, P. A. Barber, B. McGuinness, D. Field, H. Ma, W. Chong, R. V. Chandra, C. F. Bladin, H. Brown, K. Redmond, D. Leggett, G. Cloud, A. Madan, N. Mahant, B. O'Brien, J. Worthington, G. Parker, P. M. Desmond, M. W. Parsons, G. A. Donnan, and S. M. Davis (2018). "Tenecteplase versus alteplase before endovascular thrombectomy (EXTEND-IA TNK): A multicenter, randomized, controlled study". In: *International Journal of Stroke* 13.3, pp. 328–334.

Boyle, Y., D. Fernando, H. Kurz, S. R. Miller, M. Zucchetto, and J. Storey (Dec. 2014). "The effect of a combination of gabapentin and donepezil in an experimental pain model in healthy volunteers: Results of a randomized controlled trial". In: *Pain* 155.12, pp. 2510–2516. URL: `http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage%7B%5C&%7Dan=00006396-201412000-00012%20http://dx.doi.org/10.1016/j.pain.2014.09.003`.

O'Connor, O. A., M. Özcan, E. D. Jacobsen, J. M. R. Vidal, J. Trotman, J. Demeter, T. Masszi, J. Pereira, R. Ramchandren, F. A. D'Amore, F. Foss, W.-S. Kim, J. P. Leonard, C. S. Chiatton, C. D. Ullmann, and A. R. Shustov (2015). "First Multicenter, Randomized Phase 3 Study in Patients (Pts) with Relapsed/Refractory Peripheral T-Cell Lymphoma: Alisertib Vs Investigator's Choice". In: *Blood* 126.23, p. 341. URL: `http://www.bloodjournal.org/content/126/23/341?sso-checked=true`.

Niikura, N., Y. Ota, N. Hayashi, M. Naito, K. Kashiwabara, K.-i. I. Watanabe, T. Yamashita, H. Mukai, and M. Umeda (Sept. 2016). "Evaluation of oral care to prevent oral mucositis in estrogen receptor-positive metastatic breast cancer patients treated with everolimus (Oral Care-BC): randomized controlled phase III trial". In: *Japanese Journal of Clinical*

Oncology 46.April 2018, pp. 879–882. URL: https://academic.oup.com/jjco/article-lookup/doi/10.1093/jjco/hyw077.

Costanzo, M. R., R. Augostini, L. R. Goldberg, P. Ponikowski, C. Stellbrink, and S. Javaheri (2015). "Design of the remede System Pivotal Trial: A Prospective, Randomized Study in the Use of Respiratory Rhythm Management to Treat Central Sleep Apnea". In: *Journal of Cardiac Failure* 21.11, pp. 892–902. URL: http://dx.doi.org/10.1016/j.cardfail.2015.08.344.

De Valk, H. W., S. Lablanche, E. Bosi, P. Choudhary, J. D. Silva, J. Castaneda, L. Vorrink, S. De Portu, and O. Cohen (2018). "Study of MiniMed 640G Insulin Pump with SmartGuard in Prevention of Low Glucose Events in Adults with Type 1 Diabetes (SMILE): Design of a Hypoglycemia Prevention Trial with Continuous Glucose Monitoring Data as Outcomes". In: *Diabetes Technology and Therapeutics* 20.11, pp. 758–766.

Chauhan, N., J. Bukovcan, E. Boucher, D. Cosgrove, J. Edeline, B. Hamilton, L. Kulik, F. Master, R. Salem, H. Kevin Kim, B. El-Rayes, J. Schlaak, S. Cohen, J. L. Raoul, J. Strosberg, B. Arslan, R. Shridhar, M. Johnson, M. Maluccio, D. Brown, J. Geschwind, D. Cosgrove, M. Choti, A. Zhu, L. Roberts, B. Rilling, R. Murthy, A. Kaseb, V. Mazzaferro, L. Kulik, M. Mulcahy, R. Salem, B. Lewandowski, A. Benson, M. Bloomston, T. Saab, H. Khabiri, D. Sze, W. Messersmith, J. Durham, N. Fidelman, and H. C. Zhao (2018). "Intra-arterial TheraSphere yttrium-90 glass microspheres in the treatment of patients with unresectable hepatocellular carcinoma: Protocol for the STOP-HCC phase 3 randomized controlled trial". In: *Journal of Medical Internet Research* 20.8.

Meretoja, A., L. Churilov, B. C. V. Campbell, R. I. Aviv, N. Yassi, C. Barras, P. Mitchell, B. Yan, H. Nandurkar, C. Bladin, T. Wijeratne, N. J. Spratt, J. Jannes, J. Sturm, J. Rupasinghe, J. Zavala, A. Lee, T. Kleinig, R. Markus, C. Delcourt, N. Mahant, M. W. Parsons, C. Levi, C. S. Anderson, G. A. Donnan, and S. M. Davis (June 2014). "The Spot Sign and Tranexamic Acid on Preventing ICH Growth – AUStralasia Trial (STOP-AUST): Protocol of a Phase II Randomized, Placebo-Controlled, Double-Blind, Multicenter Trial". In: *International Journal of Stroke* 9.4, pp. 519–524. URL: http://journals.sagepub.com/doi/10.1111/ijs.12132.

Mehta, C. R., L. Liu, and C. Theuer (2019). "An adaptive population enrichment phase III trial of TRC105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma (TAPPAS trial)". In: *Annals of Oncology* 30.1, pp. 103–108.

Muller, C., N. W. Cheung, H. Dewey, L. Churilov, S. Middleton, V. Thijs, E. I. Ekinci, C. Levi, R. Lindley, G. Donnan, M. Parsons, and C. Bladin (2018). "Treatment with exenatide in acute ischemic stroke trial protocol: A prospective, randomized, open label, blinded end-point study of exenatide vs. standard care in post stroke hyperglycemia". In: *International Journal of Stroke* 13.8, pp. 857–862.

Forni, C., F. D'Alessandro, P. Gallerani, R. Genco, A. Bolzon, C. Bombino, S. Mini, L. Rocchegiani, T. Notarnicola, A. Vitulli, A. Amodeo, G. Celli, and P. Taddia (Jan. 2018). "Effectiveness of using a new polyurethane foam multi-layer dressing in the sacral area to prevent the onset of pressure ulcer in the elderly with hip fractures: A pragmatic randomised controlled trial". In: *International Wound Journal* October 2017, pp. 1–8. URL: `http://doi.wiley.com/10.1111/iwj.12875`.

Ravandi, F., E. K. Ritchie, H. Sayar, J. E. Lancet, M. D. Craig, N. Vey, S. A. Strickland, G. J. Schiller, E. Jabbour, H. P. Erba, A. Pigneux, H.-A. Horst, C. Recher, V. M. Klimek, J. Cortes, G. J. Roboz, O. Odenike, X. Thomas, V. Havelange, J. Maertens, H.-G. Derigs, M. Heuser, L. Damon, B. L. Powell, G. Gaidano, A.-M. Carella, A. Wei, D. Hogge, A. R. Craig, J. A. Fox, R. Ward, J. A. Smith, G. Acton, C. Mehta, R. K. Stuart, and H. M. Kantarjian (Sept. 2015). "Vosaroxin plus cytarabine versus placebo plus cytarabine in patients with first relapsed or refractory acute myeloid leukaemia (VALOR): a randomised, controlled, double-blind, multinational, phase 3 study". In: *The Lancet Oncology* 16.9, pp. 1025–1036. URL: `http://linkinghub.elsevier.com/retrieve/pii/S1470204515002016`.

Taylor, D., A. Grobler, and S. Abdool Karim (2012). "An adaptive design to bridge the gap between Phase 2b/3 microbicide effectiveness trials and evidence required for licensure". In: *Clinical Trials*.

Wan, H. (2013). "Issues In Group Sequential / adaptive Designs Issues In Group Sequential / adaptive Designs". In:

Edwards, J. M., S. J. Walters, C. Kunz, and S. A. Julious (2020). “A systematic review of the “promising zone” design”. In: *Trials* 21.1, pp. 1–10.

Pilz, M., M. Kieser, and K. Kunzmann (2020a). “Comments on “Adaptive sample size modification in clinical trials: Start small then ask for more?”” In: *Statistics in Medicine* 39.1, pp. 97–98.

Pilz, M., S. Kilian, and M. Kieser (2020b). “A note on the shape of sample size functions of optimal adaptive two-stage designs”. In: *Communications in Statistics - Theory and Methods* 0.0, pp. 1–8. URL: `https://doi.org/10.1080/03610926.2020.1776875`.

Huang, Z., F. Samuelson, L. Tcheuko, and W. Chen (2020). “Adaptive designs in multi-reader multi-case clinical trials of imaging devices”. In: *Statistical Methods in Medical Research* 29.6, pp. 1592–1611.

Wang, X., T. Xu, S. Zhong, Y. Zhou, and L. Cui (2019). “An efficient sample size adaptation strategy with adjustment of randomization ratio”. In: *Biometrical Journal* 61.3, pp. 769–778.

Gao, L., H. Zhu, and L. Zhang (2020). “Sequential monitoring of response-adaptive randomized clinical trials with sample size re-estimation”. In: *Journal of Statistical Planning and Inference* 205, pp. 129–137. URL: `https://doi.org/10.1016/j.jspi.2019.06.007`.

Wang, J., J. Li, Y. Shu, and X. Su (2020). “A Practical Perspective: Application of the Generalized Approach for Adaptive Design”. In: *Therapeutic Innovation and Regulatory Science* 54.1, pp. 167–170. URL: `https://doi.org/10.1007/s43441-019-00041-1`.

Niewczas, J., C. U. Kunz, and F. König (2019). “Interim analysis incorporating short- and long-term binary endpoints”. In: *Biometrical Journal* 61.3, pp. 665–687.

Casula, D., A. Callegaro, P. Nakanwagi, V. Weynants, and A. K. Arora (2020). “Evaluation of an Adaptive Seamless Design for a Phase II/III Clinical Trial in Recurrent Events Data to Demonstrate Reduction in Number of Acute Exacerbations in Patients With Chronic Obstructive Pulmonary Disease (COPD)”. In: *Statistics in Biopharmaceutical Research* 12.3, pp. 273–278. URL: `https://doi.org/10.1080/19466315.2020.1764382`.

Jimenez, J. L. (2020). “Innovative adaptive designs in oncology clinical trials with drug combinations”. PhD thesis.

Asakura, K., S. R. Evans, and T. Hamasaki (2020). “Interim Monitoring for Futility in Clinical Trials With Two Co-Primary Endpoints Using Prediction”. In: *Statistics in Biopharmaceutical Research* 12.2, pp. 164–175. URL: https://doi.org/10.1080/19466315.2019.1677494.

Herrmann, C., M. Pilz, M. Kieser, and G. Rauch (2020). “A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation”. In: *Statistics in Medicine* 39.15, pp. 2067–2100.

National Institure for Health Research (2019). *NIHR position on the sharing of research data*. URL: https://www.nihr.ac.uk/documents/nihr-position-on-the-sharing-of-research-data/12253.

Ulug, P., R. J. Hinchliffe, M. J. Sweeting, M. Gomes, M. T. Thompson, S. G. Thompson, R. J. Grieve, R. Ashleigh, R. M. Greenhalgh, and J. T. Powell (2018). “Strategy of endovascular versus open repair for patients with clinical diagnosis of ruptured abdominal aortic aneurysm: The IMPROVE RCT”. In: *Health Technology Assessment* 22.31, pp. 1–122.

Rothwell, J. C., S. A. Julious, and C. L. Cooper (2018). “A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal Suzie Cro”. In: *Trials* 19.1, pp. 1–13.

Bosanquet, K., J. Adamson, K. Atherton, D. Bailey, C. Baxter, J. Beresford-Dent, J. Birtwistle, C. Chew-Graham, E. Clare, J. Delgadillo, D. Ekers, D. Foster, R. Gabe, S. Gascoyne, L. Haley, J. Hamilton, R. Hargate, C. Hewitt, J. Holmes, A. Keding, H. Lewis, D. McMillan, S. Meer, N. Mitchell, S. Nutbrown, K. Overend, S. Parrott, J. Pervin, D. A. Richards, K. Spilsbury, D. Torgerson, G. Traviss-Turner, D. Trépel, R. Woodhouse, and S. Gilbody (2017). “Collaborative care for screen-positive elders with major depression (CASPER plus): A multicentred randomized controlled trial of clinical effectiveness and cost-effectiveness”. In: *Health Technology Assessment* 21.67, pp. 1–251.

State of the nation: Stroke statistics (2018). URL: https://www.stroke.org.uk/resources/state-nation-stroke-statistics (visited on 10/16/2019).

Pomeroy, V. M., S. M. Hunter, H. Johansen-Berg, N. S. Ward, N. Kennedy, E. Chandler, C. J. Weir, J. Rothwell, A. Wing, M. Grey, G. Barton, and N. Leavey (2018). "Functional strength training versus movement performance therapy for upper limb motor recovery early after stroke: a RCT". In: *Efficacy and Mechanism Evaluation* 5.3, pp. 1–112.

Koh, C. L., I. P. Hsueh, W. C. Wang, C. F. Sheu, T. Y. Yu, C. H. Wang, and C. L. Hsieh (2006). "Validation of the action research arm test using item response theory in patients after stroke". In: *Journal of Rehabilitation Medicine* 38.6, pp. 375–380.

Donaldson, C., R. Tallis, S. Miller, A. Sunderland, R. Lemon, and V. Pomeroy (2009). "Effects of conventional physical therapy and functional strength training on upper limb motor recovery after stroke: A randomized phase II study". In: *Neurorehabilitation and Neural Repair* 23.4, pp. 389–397.

Littlewood, C., S. May, and S. Walters (2013). "Epidemiology of Rotator Cuff Tendinopathy: A Systematic Review". In: *Shoulder & Elbow* 5.4, pp. 256–265.

Lewis, J. S. (2009). "Rotator cuff tendinopathy". In: *British Journal of Sports Medicine* 43.4, pp. 236–241.

Littlewood, C., M. Bateman, K. Brown, J. Bury, S. Mawson, S. May, and S. J. Walters (2016). "A self-managed single exercise programme versus usual physiotherapy treatment for rotator cuff tendinopathy: A randomised controlled trial (the SELF study)". In: *Clinical Rehabilitation* 30.7, pp. 686–696.

Breckenridge, J. D. and J. H. McAuley (2011). "Shoulder Pain and Disability Index (SPADI)". In: *Journal of Physiotherapy* 57.3, p. 197. URL: http://dx.doi.org/10.1016/S1836-9553(11)70045-5.

Littlewood, C., P. Malliaras, S. Mawson, S. May, and S. J. Walters (2014). "Self-managed loaded exercise versus usual physiotherapy treatment for rotator cuff tendinopathy: A pilot randomised controlled trial". In: *Physiotherapy (United Kingdom)* 100.1, pp. 54–60. URL: http://dx.doi.org/10.1016/j.physio.2013.06.001.

Lewis, H., J. Adamson, K. Atherton, D. Bailey, J. Birtwistle, K. Bosanquet, E. Clare, J. Delgadillo, D. Ekers, D. Foster, R. Gabe, S. Gascoyne, L. Haley, R. Hargate, C. Hewitt, J. Holmes, A. Keding, A. Lilley-Kelly, J. Maya, D. McMillan, S. Meer, J. Meredith, N. Mitchell, S. Nutbrown, K. Overend, M. Pasterfield, D. Richards, K. Spilsbury, D. Torgerson, G. Traviss-Turner, D. Trépel, R. Woodhouse, F. Ziegler, and S. Gilbody (2017). "Collaborative care and active surveillance for screen-positive EldeRs with subthreshold depression (CASPER): A multicentred randomised controlled trial of clinical effectiveness and cost-effectiveness". In: *Health Technology Assessment* 21.8, pp. 1–196.

Volker, D., M. C. Zijlstra-Vlasveld, E. P. Brouwers, W. A. Homans, W. H. Emons, and C. M. van der Feltz-Cornelis (2016). "Validation of the Patient Health Questionnaire-9 for Major Depressive Disorder in the Occupational Health Setting". In: *Journal of Occupational Rehabilitation* 26.2, pp. 237–244.

Thomas, K. J., H. MacPherson, L. Thorpe, J. Brazier, M. Fitter, M. J. Campbell, M. Roman, S. J. Walters, and J. Nicholl (2006). "Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain". In: *British Medical Journal* 333.7569, pp. 623–626.

Ware, J., K. Snow, M. Kosinski, and B. Gandek (1993). *SF-36 Health Survey: Manual and interpretation guide*. Boston: Health Institute, New England Medical Centre.

Fleming, D. M. and A. J. Elliot (2008). "Lessons from 40 years' surveillance of influenza in England and Wales". In: *Epidemiology and Infection* 136.7, pp. 866–875.

Matias, G., R. J. Taylor, F. Haguinet, C. Schuck-Paim, R. L. Lustig, and D. M. Fleming (2016). "Modelling estimates of age-specific influenza-related hospitalisation and mortality in the United Kingdom". In: *BMC Public Health* 16.1, pp. 1–10. URL: `http://dx.doi.org/10.1186/s12889-016-3128-4`.

Dalbem, J., A. Dalbem, H. Siqueira, R. Alvarenga, and M. Andraus (Oct. 2015). "Epilepsy prevalence in children and adolescents aged 0-19 years - a door-to-door survey". In: *Journal of the Neurological Sciences* 357, e195–e196.

Epilepsy Action:Epileptic seizures explained (2020). URL: `www.epilepsy.org.uk` (visited on 04/10/2020).

Meador, K. J., H. Yang, J. E. Piña-Garza, A. Laurenza, D. Kumar, and K. A. Wesnes (2016). "Cognitive effects of adjunctive perampanel for partial-onset seizures: A randomized trial". In: *Epilepsia* 57.2, pp. 243–251.

Goodacre, S., E. Cross, J. Arnold, K. Angelini, S. Capewell, and J. Nicholl (2005). "The health care burden of acute chest pain". In: *Heart* 91.2, pp. 229–230.

Goodacre, S., M. Bradburn, P. Fitzgerald, E. Cross, P. Collinson, A. Gray, and A. S. Hall (2011). "The RATPAC (randomised assessment of treatment using panel assay of cardiac markers) trial: A randomized controlled trial of point-of-care cardiac markers in the emergency department". In: *Health Technology Assessment* 15.23, pp. 1–108.

Fleetcroft, R., A. Martin, E. Coombes, J. Ford, N. Steel, and M. Noble (2016). "Emergency hospital admissions for asthma and access to primary care: Cross-sectional analysis". In: *British Journal of General Practice* 66.650, e640–e666.

NHS (2018). *Arrhythmia*. URL: `https://www.nhs.uk/conditions/arrhythmia/` (visited on 09/06/2019).

American Heart Association (2016). URL: `https://www.heart.org/en/health-topics/arrhythmia/prevention--treatment-of-arrhythmia/ablation-for-arrhythmias` (visited on 09/06/2019).

Sharples, L., C. Everett, J. Singh, C. Mills, T. Spyt, Y. Abu-Omar, S. Fynn, B. Thorpe, V. Stoneman, H. Goddard, J. Fox-Rushby, and S. Nashef (2018). "Amaze: A double-blind, multicentre randomised controlled trial to investigate the clinical effectiveness and cost-effectiveness of adding an ablation device-based maze procedure as an adjunct to routine cardiac surgery for patients with pre-existing atria". In: *Health Technology Assessment* 22.19.

Greiner, A. N., P. W. Hellings, G. Rotiroti, and G. K. Scadding (2011). "Allergic rhinitis". In: *The Lancet* 378.9809, pp. 2112–2122. URL: `http://dx.doi.org/10.1016/S0140-6736(11)60130-X`.

Yanez, A., A. Dimitroff, P. Bremner, C.-S. Rhee, G. Luscombe, B. A. Prillaman, and N. Johnson (2016). "A patient preference study that evaluated fluticasone furoate and mometasone furoate nasal sprays for allergic rhinitis". In: *Allergy & Rhinology* 7.4, pp. 183–192.

Meltzer, E. O., J. E. Stahlman, J. Leflein, S. Meltzer, J. Lim, A. A. Dalal, B. A. Prillaman, and E. E. Philpot (2008). “Preferences of adult patients with allergic rhinitis for the sensory attributes of fluticasone furoate versus fluticasone propionate nasal sprays: A randomized, multicenter, double-blind, single-dose, crossover study”. In: *Clinical Therapeutics* 30.2, pp. 271–279.

Van Deuren, M., P. Brandtzaeg, and J. W. Van Der Meer (2000). “Update on meningococcal disease with emphasis on pathogenesis and clinical management”. In: *Clinical Microbiology Reviews* 13.1, pp. 144–166.

Peate, I. (2011). “Raising awareness of meningitis and septicaemia in children”. In: *British Journal of Healthcare Assistants* 5.7, pp. 320–324.

Foundation, M. R. (2020). *Meningococcal meningitis*. URL: `https://www.meningitis.org/meningitis/causes/meningococcal-meningitis` (visited on 04/22/2020).

A | Search strategy

A.1 Search Strategies

A.1.1 Web of Science:

TS=(((((((((("sample size reestimation") OR "sample size re-estimation") OR "sample size adjustment") OR "sample size readjustment") OR "sample size modification") OR "sample size recalculation") OR "sample size reassessment") OR "adaptive sample size") OR "*crease in sample size") OR "*creased sample size") AND TS=((("promising zone") OR (“promising region”) OR (((("promising") AND "results") AND "interim") AND "conditional power")))

A.1.2 Pubmed:

(((((((((((("sample size reestimation") OR "sample size re-estimation") OR "sample size adjustment") OR "sample size readjustment") OR "sample size modification") OR "sample size recalculation") OR "sample size reassessment") OR "adaptive sample size") OR "increase in sample size") OR "increased sample size" OR "decrease in sample size") OR "decreased sample size")) AND ((("promising zone") OR (“promising region”) OR (((("promising") AND "results") AND "interim") AND "conditional power")))

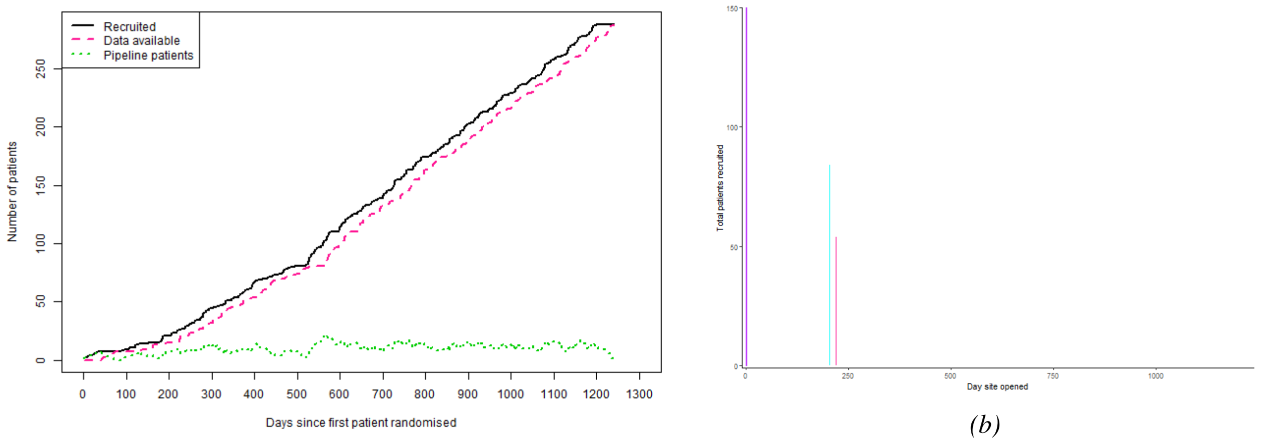
B | Detailed data decription

B.1 FAST INdiCATE RCT

Stroke rehabilitation is vital in order to recover mobility, with almost two thirds of stroke survivors leaving hospital with a disability according to the Stroke Association (*State of the nation: Stroke statistics* 2018). Upper limb recovery is particularly important for performing everyday activities and the ability for living independently (Pomeroy 2018). While it is known that physical therapy aids mobility recovery, not everyone responds in the same way to task specific training, which could be a result of a difference in neural deficits following stroke. FST is a form of physical therapy that is particularly focused on the ability to perform everyday tasks. Alternatively, MPT focuses on the quality of movement when performing these everyday tasks.

The FAST INdiCATE RCT (Functional strength training versus movement performance therapy for upper limb motor recovery early after stroke) was designed to gain a greater understanding of different neural deficits following stroke, and how a patient may respond to different physical therapies in addition to standard care (Conventional Physical Therapy (CPT)) (CPT + FST vs CPT + MPT). The primary outcome was upper limb functionality at 6 weeks, assessed using the Action Research Arm Test (ARAT) score. The ARAT score is a continuous measure, ranging from 0-57, where a higher score indicates a higher level of performance (Koh 2006).

Data from a previous early phase study was used to inform the sample size calculation (Donaldson 2009). To detect a mean difference of 6.2 points in ARAT score with 80% power and 5% two-sided significance level, and allowing for differentSDs (CPT + FST; SD=19.3, CPT + MPT; SD=7.9), 99 patients would be required in each arm. However, due to the expected clustering of patients within a therapist within a treatment arm within site (Intraclass Correlation (ICC)=0.01), and allowing for a 10% loss to follow up rate, 144 patients would be required in each treatment arm.



(b)

(a)

Figure B.1: Recruitment to the FAST INdiCATE trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

A total of 288 patients were randomised from three UK stroke services between October 2012 and January 2016. Figure B.1a shows the recruitment rate to the FAST INdiCATE study (black line), and when primary outcome data becomes available (pink dashed line). Those who have been recruited to the study, but do not yet have data available are referred to as pipeline patients (green dotted line).

Figure B.1b describes the site recruitment, where each bar represent when a site was opened (x-axis), and total recruitment to that site (y-axis).

Total days of study	1241
Days to last patient recruited	1199
Days to 50% patient recruitment	708
Number of sites	3
Days to last site recruited	204
Days to 50% site recruitment	220

Table B.1: Recruitment summary for the FAST INdiCATE study

Primary analysis used ANCOVA to model ARAT score, adjusted for baseline score, centre and time after stroke. All patients had stratification variables available (centre and time after stroke). At 6 weeks, 114/143 (79%) of CPT + MPT patients, and 126/145 (87%) of CPT + FST patients had both baseline and primary outcome measurements available.

Both treatment groups improved in terms of mean ARAT score between baseline and 6 weeks (CPT + FST = 9.70 (SD = 11.72); CPT + MPT = 7.90 (SD = 9.18)). However there was no statistically significant difference between the two treatment groups (Least squares difference = 1.35, 95% CI (-1.20, 3.90), *p*=0.298). Figure B.2 shows a summary of patient flow through the trial and details of missing data.

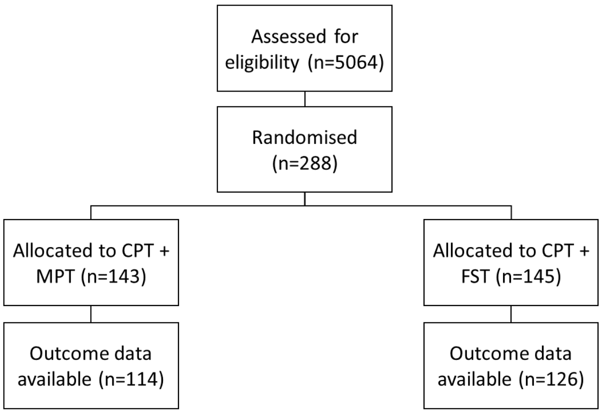

Assessed for eligibility (n=5064)
Randomised (n=288)
Allocated to CPT + MPT (n=143)
Allocated to CPT + FST (n=145)
Outcome data available (n=114)
Outcome data available (n=126)

Figure B.2: Flowchart of participants in the FAST INdiCATE trial

B.2 SELF

Rotator cuff tendinopathy, a condition where pain comes from at least one of the four tendons that make up the rotator cuff, is one of the most commmon causes of shoulder pain (Littlewood 2013). Treatments include conservative methods (such as physiotherapy or medication/injections), or surgical procedures (Lewis 2009). Loaded exercise (against gravity, or with resistance) may have some potential benefits over conventional physiotherapy, but the effects are unknown.

The SELF study aimed to investigate a self managed loaded single exercise program vs usual physiotherapy treatment (Littlewood 2016). The primary outcome was Shoulder Pain and Disability Index (SPADI), a continuous measure consisting of 13 questions (5 to assess pain, and 8 to assess disability), each based on a 0-10 Visual Analogue Scale (VAS) score (Breckenridge 2011). The subscales are converted to a total score of 0-100, where a higher score indicates greater pain or disability. SPADI scores were collected at baseline and at 3 months post-randomisation.

An external pilot study was used to inform the sample size calculation (Littlewood 2014).

To detect a MCID of 10 points in SPADI, with SD=16.8, 80% power and 5% two-sided significance, a correlation of 0.5 between baseline and 3 month SPADI scores and a 15% loss to follow up rate, meant that a total of 78 participants were required (Littlewood 2016).

Between April 2012 and July 2013, a total of 86 patients were randomised to the SELF study (intervention group: n=42, control group: n=44) across three sites. Site data and randomisation dates were not available in the dataset provided for analysis in this thesis. Randomisation dates were modelled according to the Section 4.4.3.3 in the analysis plan. Figure B.1a shows patient recruitment (black line) using these modelled dates, and when primary outcome data becomes available (pink dashed line). Those who have been recruited to the study, but do not yet have data available are referred to as pipeline patients (green dotted line).

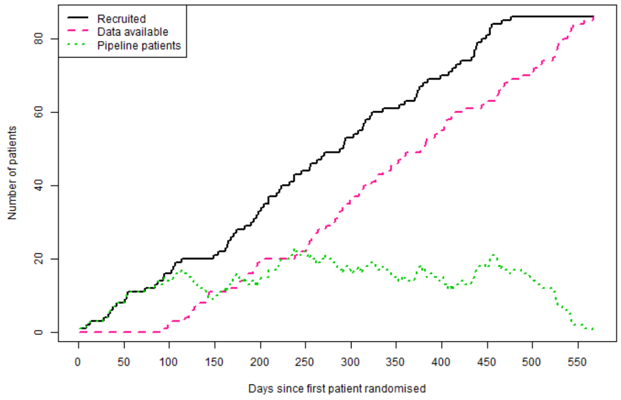


Figure B.3: Time between patients recruitment and primary outcome data became available. Green (dotted) line show pipeline patients who have been randomised, but have no data available yet.

Total days of study	567
Days to last patient recruited	477
Days to 50% patient recruitment	238
Number of sites	3

Table B.2: Recruitment summary for the SELF study

Primary analysis used ANCOVA to model SPADI, adjusted for baseline score. At 3 months, 27/42 (64%) had primary outcome data available intervention group, and 33/44 (75%) had data available in the control group. One patient in the control group had missing baseline data.

Both treatment groups improved in terms of SPADI between baseline and 3 months (self managed exercise by 12.4 points (95% CI 5.4 - 19.5, *p*<0.01), and physiotherapy group by 16.7 points (95% CI 9.6 - 23.7, *p*<0.01). However, there was no statistically significant difference between the groups (Adjusted MD=3.2 points (95% CI -6.0 - 12.4, *p*=0.49).

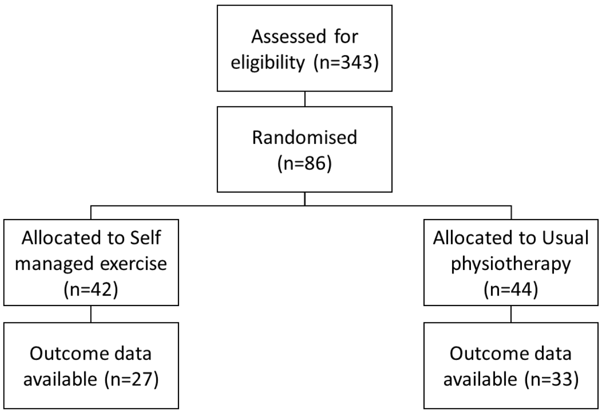

Assessed for eligibility (n=343)
Randomised (n=86)
Allocated to Self managed exercise (n=42)
Allocated to Usual physiotherapy (n=44)
Outcome data available (n=27)
Outcome data available (n=33)

Figure B.4: Flowchart of participants in the SELF trial

Figure B.4 shows a summary of patient flow through the trial and details of missing data (adapted from Littlewood 2016).

B.3 CASPER

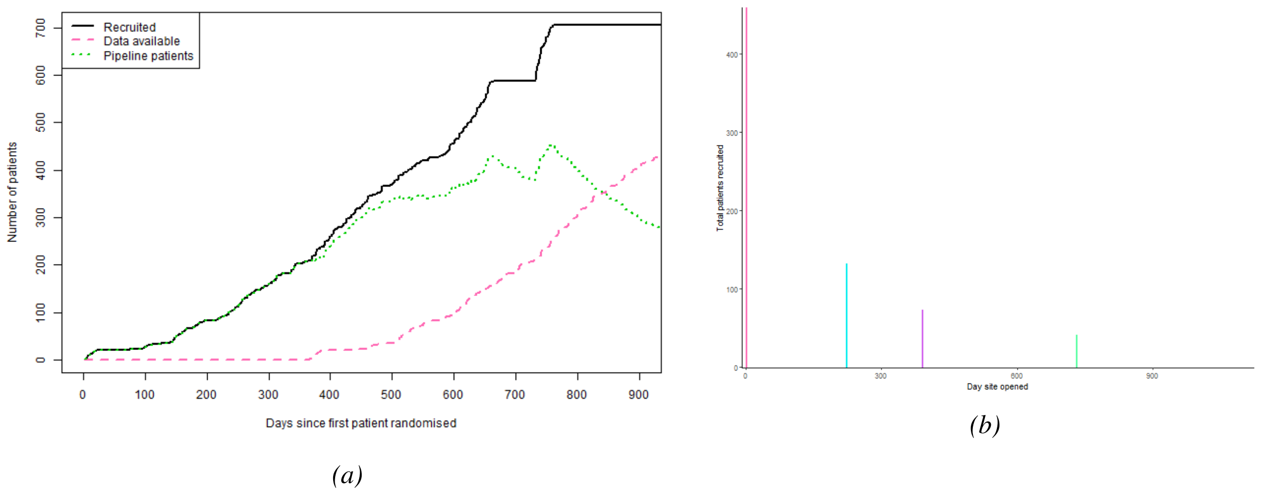
Older adults are at a higher risk of depression, particularly those who are isolated or lonely (Lewis 2017). However, this condition often goes untreated, particularly in mild cases, despite being at risk for progression to more severe depression. It is also thought in milder cases, treatment needs to be psychological and/or social based interventions.

The CASPER (collaborative care and active surveillance for screen-positive elders) trial aimed to assess usual GP care vs collaborative care, where a case manager working with the primary care team delivers low-intensity psychological treatments by telephone, alongside usual GP care.

The primary outcome was depression severity at 4 months, using the Patient Health Questionnaire-9 items (PHQ-9) score, a continuous measure ranging from 0-27, where a higher score indicates a greater level of depression (Volker 2016).

To detect an effect size of 0.3 with 80% power, 5% two-sided significance and allowing for a 26% loss to follow up, a total of 704 patients would be required.

Between June 2011 and July 2013, 705 participants were randomised into the study. Figure B.5a shows recruitment, data availability and pipeline patients during the progression of the CASPER trial. Figure B.5b describes the site recruitment, where each bar represent when a site was opened (x-axis), and total recruitment to that site (y-axis).



(a)

(b)

Figure B.5: Recruitment to the CASPER trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Total days of study	1126
Days to last patient recruited	761
Days to 50% patient recruitment	482
Number of sites	4
Days to last site recruited	732
Days to 50% site recruitment	222

Table B.3: Recruitment summary for the CASPER study

Primary analysis in the original trial used a linear mixed model, using baseline, 4 month and 12 months (outcome) PHQ-9 scores, and adjusted for baseline 12 item Short Form (SF-12) scores. The original analysis found a statistically significant difference in PHQ-9 scores between the two groups at 12 months in favour of the collaborative care intervention arm (MD=1.33, 95% CI 0.55 - 2.1, *p*=0.001).

The data re-analysis in this thesis will use data from the CASPER trial to investigate two different outcome time points. Original 12 month outcome data will be modelled with

ANCOVA, adjusting for baseline PHQ-9 and SF-12 scores. Data collected at 4 months will be re-imagined as a 4-week time point, and will also use ANCOVA adjusted for baseline PHQ-9 and SF-12 scores. To distinguish between the two, analyses using the 12 month outcome data will be referred to as CASPER, and analyses using the 4 month data (re-imagined at 4 weeks) will be referred to as the CASPER MINUS study (see Section B.3.1).

The flow of patients using only 12 month data only is shown in Figure B.6. All patients had baseline PHQ-9 scores, but 12 patients were missing SF-12 scores (7 in the collaborative care group, and 5 in the usual GP care group).

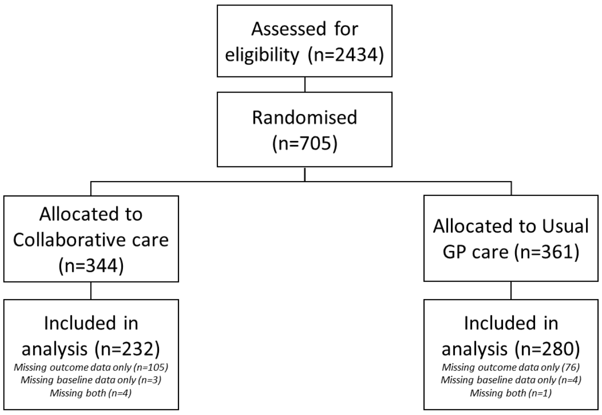

Assessed for eligibility (n=2434)
Randomised (n=705)
Allocated to Collaborative care (n=344)
Allocated to Usual GP care (n=361)
Included in analysis (n=232)
Missing outcome data only (n=105)
Missing baseline data only (n=3)
Missing both (n=4)
Included in analysis (n=280)
Missing outcome data only (76)
Missing baseline data only (n=4)
Missing both (n=1)

Figure B.6: Flowchart of participants in the CASPER trial (12 month data)

B.3.1 Re-imagined time point of CASPER

Figure B.7 shows the pipeline patients using 4 month data re-imagined at 4 weeks as the primary outcome. The total number of study days now becomes 789, as opposed to 1126 in the 12 month study, but all other recruitment metrics remain the same as the CASPER study.

Figure B.8 shows the missing data for the 4 month outcome. Again, no patients were missing baseline PHQ-9 scores and missing baseline data refers to the SF-12 scores.

B.4 CASPER Plus

The CASPER PLUS study aimed to assess collaborative care vs usual GP care in elderly patients with major depression (as opposed to low-level depression investigated in the CASPER

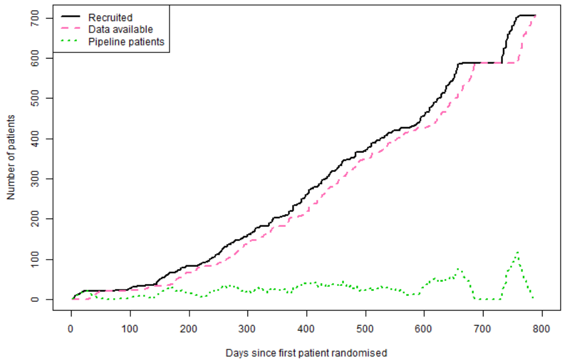


Figure B.7: Time between patients recruitment and primary outcome data became available. Green (dotted) line show pipeline patients who have been randomised, but have no data available yet.

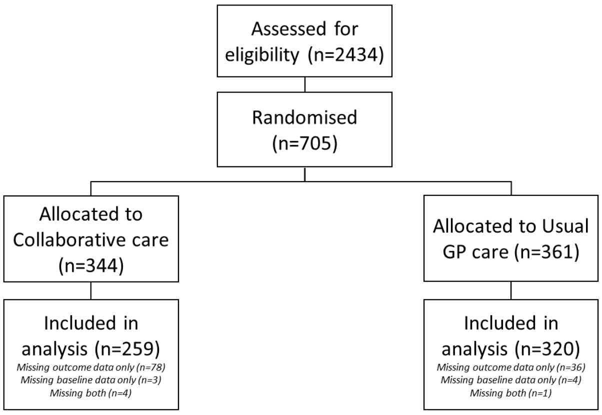
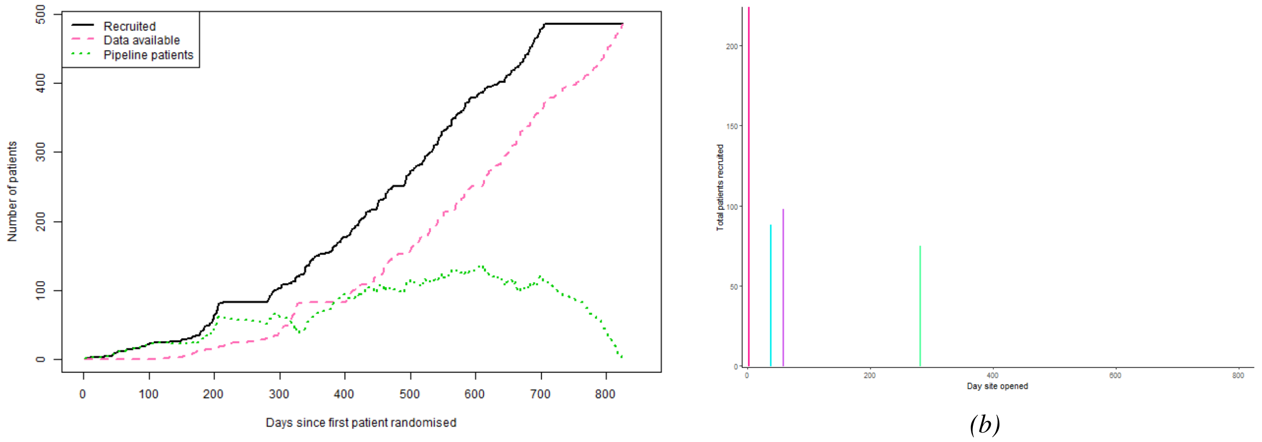

Assessed for eligibility (n=2434)
Randomised (n=705)
Allocated to Collaborative care (n=344)
Allocated to Usual GP care (n=361)
Included in analysis (n=259)
Missing outcome data only (n=78)
Missing baseline data only (n=3)
Missing both (n=4)
Included in analysis (n=320)
Missing outcome data only (n=36)
Missing baseline data only (n=4)
Missing both (n=1)

Figure B.8: Flowchart of participants in the CASPER MINUS trial

study). The CASPER PLUS study started recruiting 15 months after the start of the CASPER trial, but before the trial had finished recruiting.

The primary outcome of the CASPER PLUS study was also depression severity using the PHQ-9 score, but the primary endpoint time was measured at 4 months. It was thought that in the major depression population, it was thought an effect size of 0.35 would be clinically important, in line with the estimate used to inform the CASPER trial sample size. Allowing for a 20% loss to follow up rate, 5% two-sided significance and 80% power and an effect size of 0.35 would require 484 patients, or 284 per group.

Between September 2012 and August 2014, 485 patients were randomised into the CASPER PLUS trial (249 to collaborative care, and 236 to usual GP care). Figure B.9a shows the time between recruitment and primary outcome data becoming available. Figure B.9b describes the site recruitment, where each bar represent when a site was opened (x-axis), and total recruitment to that site (y-axis).



(b)

(a)

Figure B.9: Recruitment to the CASPER PLUS trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Figure B.10 shows the missing data for the 4 month outcome and flow of participants through the study.

In the original analysis of CASPER PLUS, a linear mixed model was used for PHQ-9 scores at 4 ,12 and 18 months and adjusting for baseline PHQ-9 and SF-12 scores. At 4 months (primary outcome time point), the study showed a mean difference of 1.92 points

Total days of study	825
Days to last patient recruited	705
Days to 50% patient recruitment	465
Number of sites	4
Days to last site recruited	281
Days to 50% site recruitment	37

Table B.4: Recruitment summary for the CASPER PLUS study

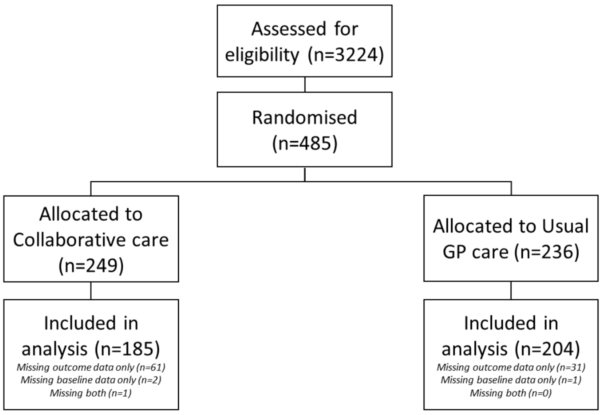

Assessed for eligibility (n=3224)
Randomised (n=485)
Allocated to Collaborative care (n=249)
Allocated to Usual GP care (n=236)
Included in analysis (n=185)
Missing outcome data only (n=61)
Missing baseline data only (n=2)
Missing both (n=1)
Included in analysis (n=204)
Missing outcome data only (n=31)
Missing baseline data only (n=1)
Missing both (n=0)

Figure B.10: Flowchart of participants in the CASPER PLUS trial

(95% CI 0.85-2.99, p<0.001) in favour of the collaborative care group.

B.5 Acupuncture Study

Low back pain is common in the adult population, with around 16% of adults in the UK visiting their GP each year for treatment (Thomas 2006). Whilst some evidence suggests acupuncture can provide short term pain relief, the long term effects are unknown, and is not commonly used in the NHS setting (Thomas 2006). The Acupuncture trial aimed to investigate the effect acupuncture provided in a non-NHS setting vs usual primary care treatment in patients with persistent non-specific low back pain.

The primary outcome was pain score, assessed using the 36 item Short Form (SF-36) form. The form consists of 8 subscales, which can be converted to obtain two overall summary measures; one of which is the pain functioning score which was used in this study as the primary outcome. The score ranges from 0 to 100, with a higher score indicating a better health state, i.e. less pain (Ware 1993). Pain scores were collected at baseline and at

12 months post-randomisation.

The sample size calculation was informed by a pilot study. To detect a mean difference of 10 points with a SD of 19.3, with 90% power, 2-sided significance of 5% and a loss to follow up rate of 10-15%, 100 patients would be needed per arm. It was then decided to use a 2:1 randomisation ratio for the intervention arm in order to tests the effects between acupuncturists. Keeping all other parameters the same, the sample size was increased to 240 patients in total.

Between August 1999 and January 2001, 241 patients were recruited to the trial (161 to the acupuncture treatment arm, and 81 to usual care group). Figure B.11 shows patient recruitment and information on when data would become available in the Acupuncture trial.

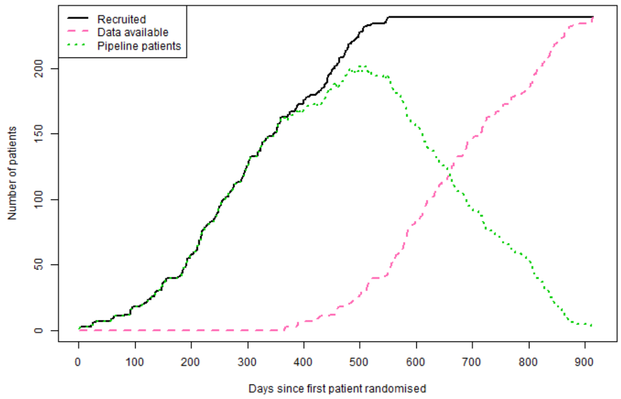


Figure B.11: Time between patients recruitment and primary outcome data became available. Green (dotted) line show pipeline patients who have been randomised, but have no data available yet.

The trial involved 39 practitioners from 16 centres, but no site information was available for the data re-analysis. A summary of patient recruitment is shown in Table B.5

Total days of study	914
Days to last patient recruited	549
Days to 50% patient recruitment	295
Number of sites	16

Table B.5: Recruitment summary for the Acupuncture study

Figure B.12 shows the missing data for the 12 month outcome and flow of participants through the study, adapted from Thomas 2006.

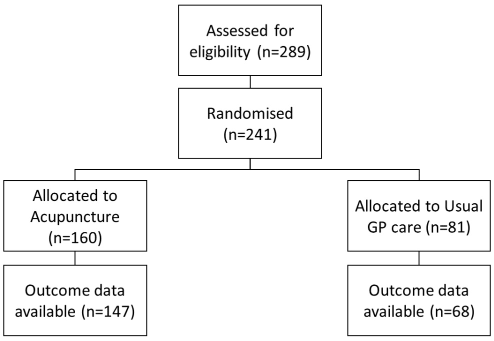

Assessed for eligibility (n=289)
Randomised (n=241)
Allocated to Acupuncture (n=160)
Allocated to Usual GP care (n=81)
Outcome data available (n=147)
Outcome data available (n=68)

Figure B.12: Flowchart of participants in the Acupuncture trial

Primary analysis used ANCOVA, adjusting for baseline pain scores. The original analysis found a mean difference of 5.6 points in favour of the acupuncture group, but this was not statistically significant (95% CI (-0.2, 11.4) p=0.06).

B.6 Flu Vaccine - SANOFI

Flu is a common respiratory illness, and can affect people of all ages (Fleming 2008). Whilst symptoms are often managed through self-care, more severe cases require hospital admission and can be fatal. A UK study (SANOFI-QID01) looking at data between 1997 and 2009 estimated that in an average season, 28,517 hospital admissions and 7163 deaths could be attributable to influenza (Matias 2016). Two trivalent (TIV) flu vaccines have been available previously, containing three inactivated strains of the flu virus: A/H1N1, A/H3N2, and one of the primary B lineages (Yamagata (TIV-1) or Victoria strains(TIV-2)), referred to in this thesis as strains B1 and B2 respectively. However, it is often difficult to predict what the dominating B-strain will be in each season and therefore which vaccine should be given (Matias 2016). Alternatively, a quadrivalent (QIV) containing A/H1N1, A/H3N2, and both B lineage strains vaccination has been developed. Sanofi Pasteur conducted a RCT in the 2012-2013 flu season, to consider the non-inferiority of the QIV vaccine, to the two TIV vaccines in terms of the immunogenicity of the four flu strains mentioned above.

Patients were randomised in a 2:1:1 ratio to QIV, TIV-1 and TIV-2 vaccines respectively. For comparisons of both A strains (H1N1 and H3N2), QIV was compared to the pooled TIV groups. For B strain comparisons, the QIV group was compared with the corresponding

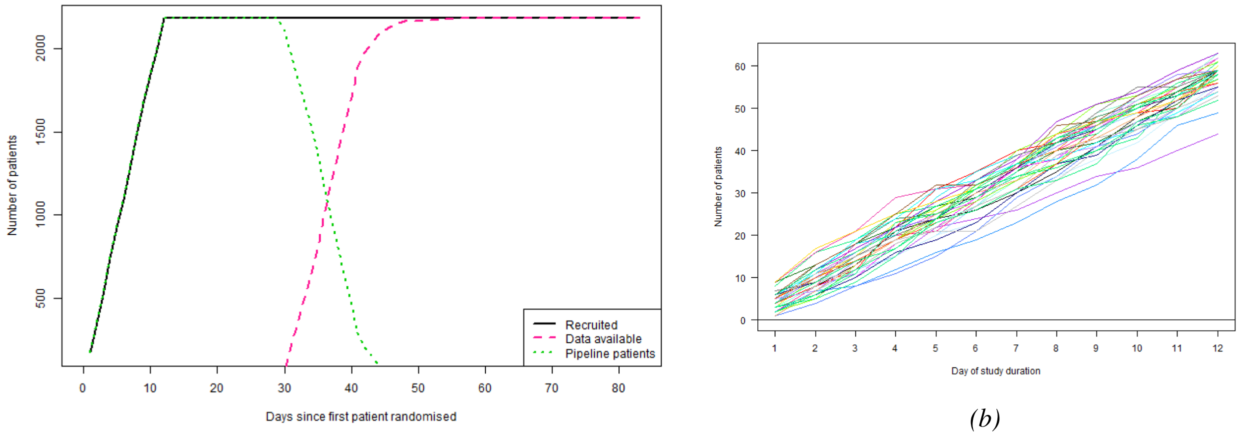
TIV group. The trial had two primary outcomes: geometric mean titer ratio (continuous) and seroconversion rates (binary). Both outcomes are used in this thesis due to the limited industry data available.

Only very limited sample size information was available due to a redacted protocol and analysis plan. With at least 1119 patients in the QIV and pooled TIV groups, 92% power, 5% significance level and a non-inferiority limit of 2/3 for the geometric mean titer ratio between QIV and TIV groups, resulted in a hypothesised effect size of 0.142. Keeping the effect size the same, power was adjusted for the smaller sample size in B strain comparisons (QIV compared to only one TIV group), and resulted in 77% power. Individual seroconversion rates were not available. Given a non-inferiority limit of 0.1, 92% power, and 5% significance, sample size was calculated for varying combinations of pi_A and pi_B between 0 and 1. Event rates of 0.55 in each arm would require at least 1115 per arm, the closest to the 1119 patients with data available. Again, adjusting the power for the B strain comparisons gave 77% power.

Recruitment began in October 2012 and 2249 patients were randomised in 38 sites over a 12 day period (QIV n=1119, TIV-1 n=560, TIV-2 n=570). No randomisation dates were available and were therefore modelled according to Section 4.4.3.3. However, the number of days between randomisation and primary outcome data collection was given for each patient, and this has been used for working out when data became available. The planned outcome measure was at 28 days, but it should be noted that some patients exceeded this time point. Figure B.13 shows patient and accrual by site. Due to the short time frame of recruitment, and large number of recruiting centres, the recruitment phase (12 days) is presented with a line for each site recruitment.

A summary of patient recruitment is shown in Table B.6

In the original analysis, the geometric mean titer ratio between QIV and TIV groups and the corresponding 95% CI was calculated using the normal approximation of log-transformed titers. This resulted in a ratio of 589/680 = 0.866, 95% CI (0.777, 0.966) in strain A/H1N1; 368/430=0.857, 95% CI (0.770, 0.955) in strain A1/H3N2; 105/93.5=1.13, 95% CI (1.02, 1.25) in strain B1; and 136/130=1.05, 95% CI (0.939, 1.16) in strain B2.



(b)

(a)

Figure B.13: Recruitment to the Flu vaccine trial. (a) shows rate of recruitment and time to data availability. (b) shows recruitment by site

Total days of study	83
Days to last patient recruited	12
Days to 50% patient recruitment	6
Number of sites	38
Days to last site recruited	1
Days to 50% site recruitment	1

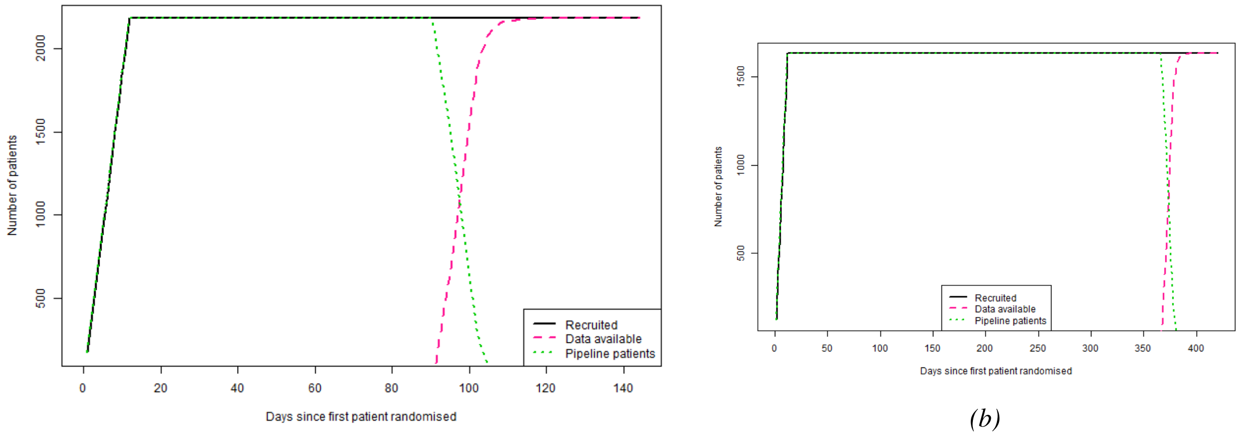
Table B.6: Recruitment summary for the Flu vaccine study

Therefore, the non-inferiority limits were met for all four strains for the continuous outcome. The difference between proportions of patients reaching seroconversion (QIV - TIV) and 95% CIs were calculated using the Wilson score method. This results in a difference of -2.72 (-6.90, 1.47) for strain A1/H1N1; -1.30 (-5.48, 2.89) for strain A1/H3N2; 8.78 (2.58, 13.9) for strain B1, and 6.34 (1.13, 11.5) for strain B2. Again, all four strains met the non-inferiority limits in the binary outcome.

B.6.1 Re-imagined time points of the Flu vaccine trial

The flu vaccine trial will be re-imagined at two further time points for both continuous and binary endpoints. Because of this, each imagined time point will use data from a different strain, instead of repeating each strain 3 times. The original trial, with a 28 day outcome will be classed as the “short-term” time point, and will use data from the A/H1N1 strain only. An additional 61 days (∼2 months) will be added to the time to primary outcome to form the “medium” time point, which will use data from the A/H3N2 strain only. Finally, the “long-term” endpoint will be 1 year from randomisation, and will use data from the B1 strain only.

Figure B.14 shows the pipeline patients for the 3-month (a) and 1-year (b) outcomes. The total study time increases to 144 and 420 days respectively.



(b)

(a)

Figure B.14: Recruitment to the Flu vaccine trial re-imagined at 3 months (a) and 1 year (b)

The overall number of patients screened for eligibility was not available. Both groups had 97% of data available for strains A/H1N1 and A/H3N2 (QIV: 1085/1119, TIV:1095/1130). For strain B1, data availability was 96% (1086/1119) and 98% (547/560) for QIV and TIV-1 groups respectively. Data availability was the same in the binary endpoints as it was in the continuous endpoints.

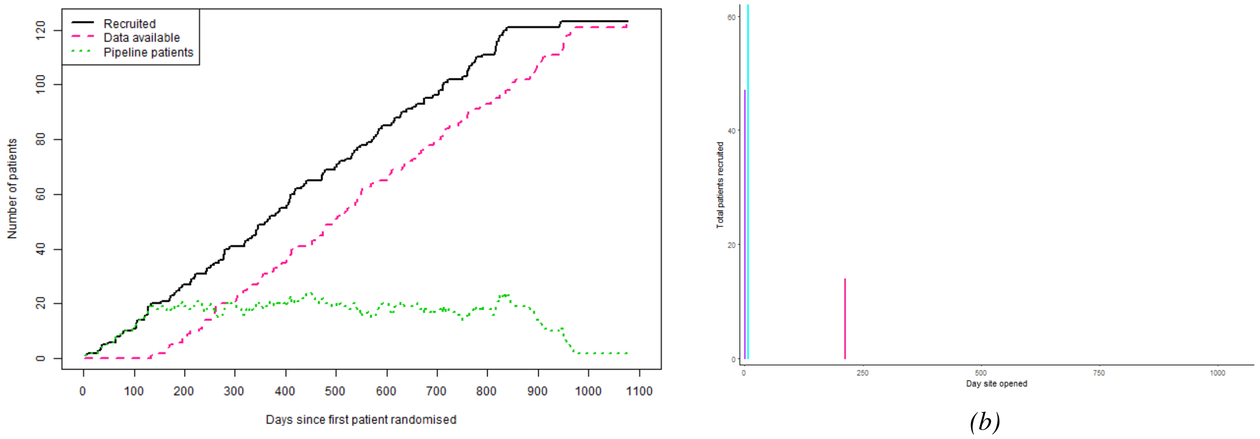
B.7 Epilepsy in adolescents - EISAI

Epilepsy is a chronic condition involving a sudden burst of electrical activity in the brain, and affects 50 million people worldwide (Dalbem 2015). Partial-onset seizures (or focal seizures) start off in just one side of the brain (*Epilepsy Action:Epileptic seizures explained* 2020). A common treatment in epileptic patients are the use of one or more Anti-Epileptic Drugs (AEDs). However, some are thought to impair cognitive function and have behavioural side effects (Meador 2016). Perampanel is a type of AED, but no previous trials have assessed its cognitive effects (Meador 2016). Therefore, Eisai Inc. conducted a phase II RCT (EISAI-E2007-G000-235) in 2015 to assess cognitive function in adolescents aged 12-17 taking between 1-3 AEDs, randomised to an adjunctive therapy of either perampanel or placebo for the purposes of the study.

Cognitive function (the primary outcome) was assessed using the Cognitive Drug Research (CDR) system, consisting of 5 core sub-scales assessing various aspects of attention and memory. Each of the 5 sub-scores are compared to the normative population mean, and then standardised to create a score between 0-100. Each score was added together to create a global score, with a greater score indicates a greater level of cognitive function. The change in global CDR score between baseline and 19 weeks was used as the primary outcome measure.

To detect a clinically meaningful difference in global CDR score of 5 points, with two-sided significance of 5%, 80% power, SD=9.10, and a randomisation ratio of 2:1, 117 evaluable patients would be required (n_A=39, n_B=78). To allow for a drop-out rate of 10%, the trial aimed to recruit at least 130 patients in total.

Recruitment took place between September 2010 and January 2013, randomising 133 patients in total (48 to placebo and 85 to Parampanel). Specific randomisation dates were not available for the study, and so dates have been modelled according to Section 4.4.3.3. Figure B.15b shows approximate patient and site recruitment. "Site" here is actually Region, as individual site details were not available.



(a)

(b)

Figure B.15: Recruitment to the Epilepsy trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Figure B.15b describes the site recruitment, where each bar represent when a site was opened (x-axis), and total recruitment to that site (y-axis).

Total days of study	1078
Days to last patient recruited	945
Days to 50% patient recruitment	419
Number of sites	3
Days to last site recruited	212
Days to 50% site recruitment	7

Table B.7: Recruitment summary for the Epilepsy study

The primary analysis was ANCOVA, modelling change in global CDR score, adjusted for baseline score, gender, age, region and treatment group. In the placebo group, patients had on average higher scores at 19 weeks than at baseline (Baseline score 41.2 points (SD=10.7), 19 weeks 42.2 points (SD=11.8), overall change +1.6 points (SD=1.3)). Conversely, the Parampanel group actually decreased on global CDR score (Baseline score 40.8 points (SD=13.0), 19 weeks 39.7 points (SD=13.5), overall change -0.6 points (SD=1.0)). However, this difference was not statistically significant (mean difference -2.2 points, 95%

CI (-5.2, 0.8), *p*=0.145). Figure B.16 shows patient flow through the trial.

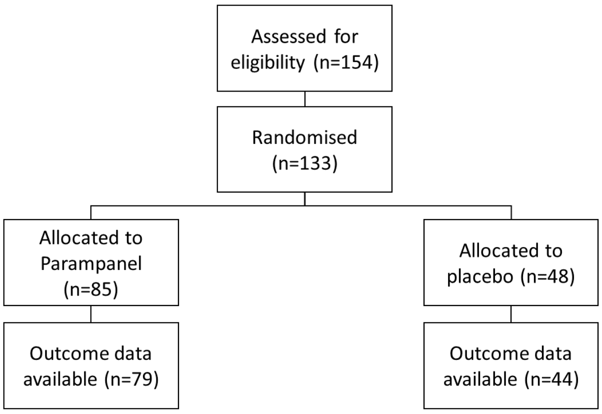
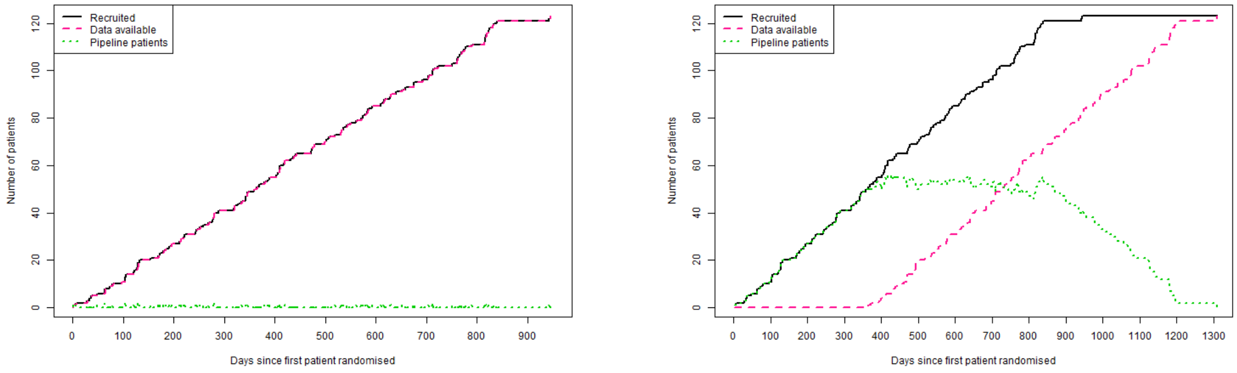

Assessed for eligibility (n=154)
Randomised (n=133)
Allocated to Parampanel (n=85)
Allocated to placebo (n=48)
Outcome data available (n=79)
Outcome data available (n=44)

Figure B.16: Flowchart of participants in the Epilepsy trial

B.7.1 Re-imagined time points of the Epilepsy trial

Due to a lack of datasets being available, the Epilepsy trial is also imagined with a much shorter time to primary outcome (1 day), and a much longer time to primary outcome (1 year).

Figure B.17 shows the pipeline patients using 19 weeks data re-imagined at 1 day (Figure (a)), and at 1 year (Figure (b)). The total number of study days now becomes 946 for the 1 day time point, and 1310 for the 1 year time point, as opposed to 1078 in the 12 month study, but all other recruitment metrics remain the same as the original Epilepsy study.



(a) *(b)*

Figure B.17: Recruitment to the re-imagined time points of the Epilepsy trial. (a) shows data availability for the 1 day outcome. (b) shows data availability for the 1 year outcome

B.8 IMPROVE

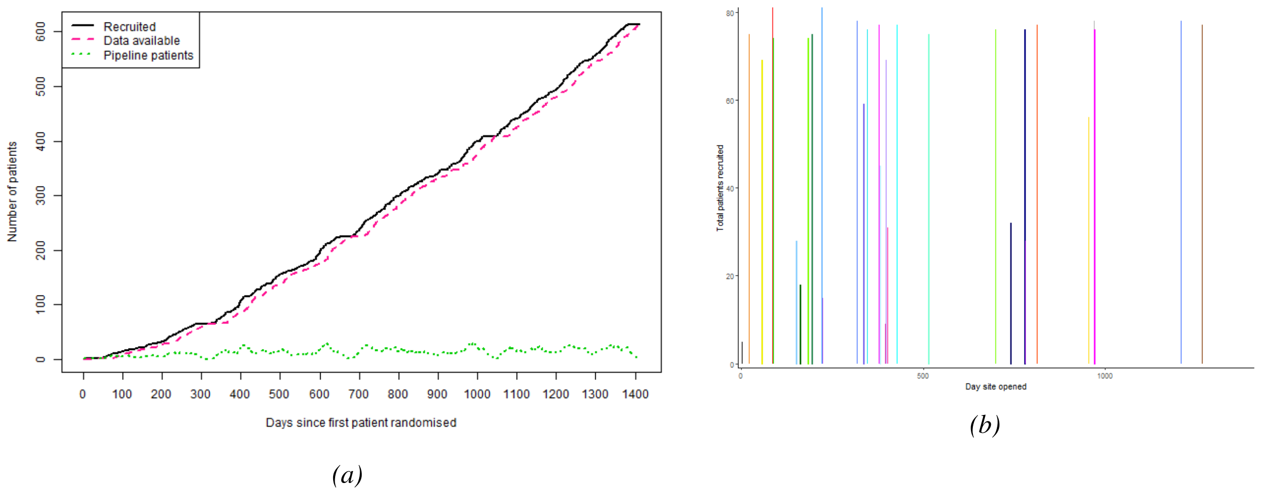
The aorta is the main blood vessel carrying oxygenated blood to the rest of the body. An AAA is a swelling of the aorta in the abdomen, and, left untreated, can enlarge and eventually rupture (Ulug 2018). A ruptured AAA is a common cause of death, with many not reaching hospital in time, and even with surgery, only about half make it out of hospital alive. Surgical intervention is typically an open repair, and those that survive have a lengthy recuperation time. It is thought that a keyhole surgery intervention, endovascular repair, may shorten recovery time and result in a lower 30 day mortality rate compared to open repair.

The IMPROVE trial (The Immediate Management of the Patient with Rupture: Open Versus Endovascular repair trial) aimed to compare these two surgical techniques. The primary outcome was mortality at 30 days. To detect a risk difference of 14%, assuming a mortality rate of 44.7% in the open repair group, and 30.4% in the endovascular repair group, with two sided 5% significance, 94% power, and 5% loss to follow up, required a total of 600 participants. The treatment effect was converted to an OR for data re-analysis, resulting in an odds ratio of mortality in the intervention group compared to the control group of 0.539, and SE of 0.171.

A total of 613 patients (open repair: n=297, endovascular repair: n=316) were randomised to the IMPROVE study from 31 centres in the UK and Canada. Figure B.18a shows recruitment rate, and number of patients with data available, or unavailable (pipeline patients). Figure B.18b shows when sites were opened, and the total number recruited to each. Table B.8 shows the time to 50% and 100% recruitment of both patients and sites.

Total days of study	1410
Days to last patient recruited	1380
Days to 50% patient recruitment	811
Number of sites	31
Days to last site recruited	1266
Days to 50% site recruitment	381

Table B.8: Recruitment summary for the IMPROVE study



(a)

(b)

Figure B.18: Recruitment to the IMPROVE trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Figure B.19 shows the missing data for the 30 day outcome and flow of participants through the study. No patients were missing outcome data, age or sex (baseline covariates). Missing baseline data came from the Hardman index covariate data.

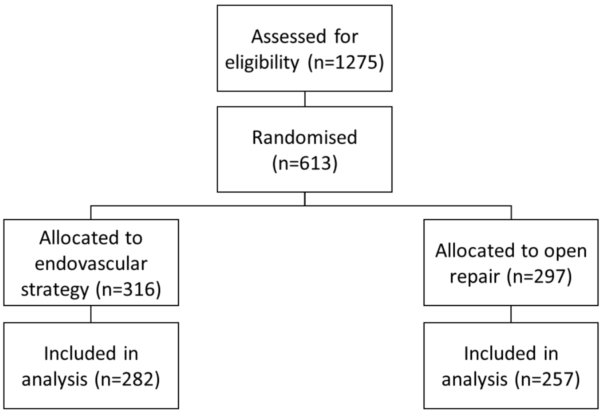

Assessed for eligibility (n=1275)
Randomised (n=613)
Allocated to endovascular strategy (n=316)
Allocated to open repair (n=297)
Included in analysis (n=282)
Included in analysis (n=257)

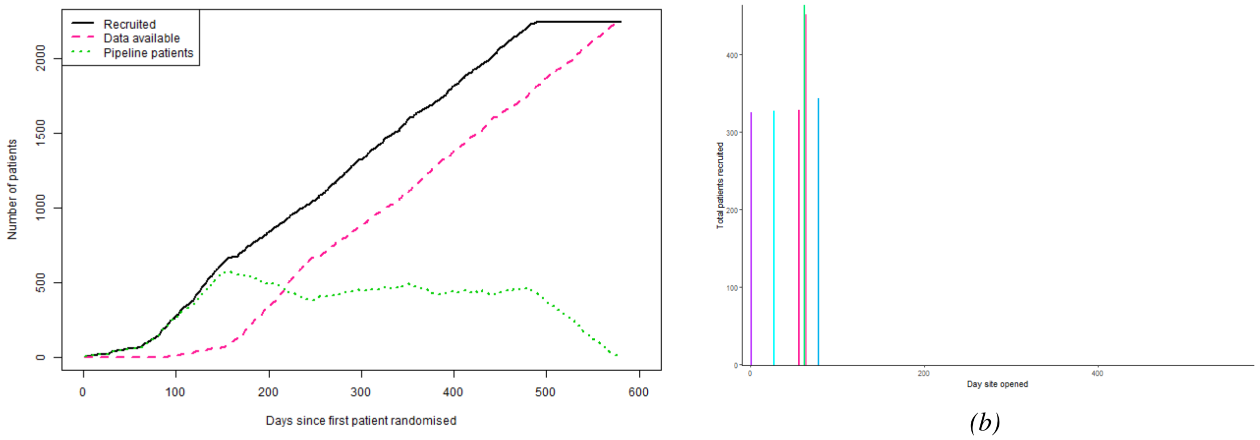
Figure B.19: Flowchart of participants in the IMPROVE trial

Primary analysis used a Pearson Chi squared test to asses proportions of patients surviving to 30 days in each group. A logistic regression was also used to provide an adjusted odds ratio, adjusting for baseline covariates age, sex and Hardman Index. The endovascular repair group resulted in slightly lower odds of death compared to the open repair group, but this was not statistically significant (OR=0.92, 95% CI (0.66, 1.28), *p*=0.62).

B.9 RATPAC

National guidelines recommend that anyone experiencing chest pain seek emergency medical help rather than contacting a GP. Therefore chest pain results a common cause of patient attendances in emergency departments in England and Wales, and appoximately 25% of hospital admissions (Goodacre 2005). Acute Myocardial Infarction (AMI) is typically diagnosed with troponin measurements in the blood. Patients with suspected but not proven AMI are recommened to have a troponin measurement taken 12 hours after onset of chest pain, due to the delay in troponin levels reaching optimal levels. However, this results in potentially unnecessary hospital admission, with most patients presenting with suspected AMI not actually having AMI (Goodacre 2011). Point-of-care testing is a quicker alternative approach, assessing additional markers including creatinine kinase MB and myoglobin at baseline and again at 90 minutes, as a potential alternative indication of AMI to troponin.

The RATPAC (Randomised Assessment of Treatment using Panel Assay of Cardiac markers) study aimed to assess the two diagnostic strategies. The primary outcome was successful discharge from hospital, defined as both (i) a decision to be discharged made within 4 hours from initial presentation, and (ii) no major adverse event within 3 months. To detect an absolute risk difference of 5% of successful discharge between the two groups,



(a)

(b)

Figure B.20: Recruitment to the RATPAC trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

assuming a successful discharge rate of 55% in the point-of-care group, and 50% in the usual care group, with 80% power and two sided 5% significance, a total of 3130 patients would be required. The RATPAC terminated early, due to slow recruitment, and the results of a CP analysis to determine a funding extension showing clear efficacy for the successful discharge primary outcome. In light of this, the power has been adjusted for the re-analysis. With at least 1118 evaluable patients per arm, the RATPAC study had 66% power to determine the same absolute risk difference. This is equivalent to an OR of 1.22 of successful discharge in the intervention group compared to the control group, with SE of 0.072

Between January 2008 and August 2009, 2263 patients were randomised to the RATPAC trial from 6 UK NHS centres (intervention group n=1132, control group n=1131). All 6 sites had been recruited by day 78 of a total of 490 days (all patients having primary outcome data). Figures B.20a and B.20b show patient recruitment and site recruitment to the RATPAC study.

Total days of study	490
Days to last patient recruited	490
Days to 50% patient recruitment	264
Number of sites	6
Days to last site recruited	78
Days to 50% site recruitment	56

Table B.9: Recruitment summary for the RATPAC study

Those that did not complete initial follow up were excluded from the analysis (intervention group n=7, control group n=13). Figure B.21 shows flow of participants and missing data for the RATPAC trial.

The primary analysis used logistic regression, adjusting for site, age, gender and past history of CHD. The original analysis showed patients in the point-of-care group were more likely to be successfully discharged home (OR=3.81, 95% CI (3.01,4.82, p<0.001). Age, gender and history of CHD data was unavailable for data re-analysis and was therefore excluded from the model.

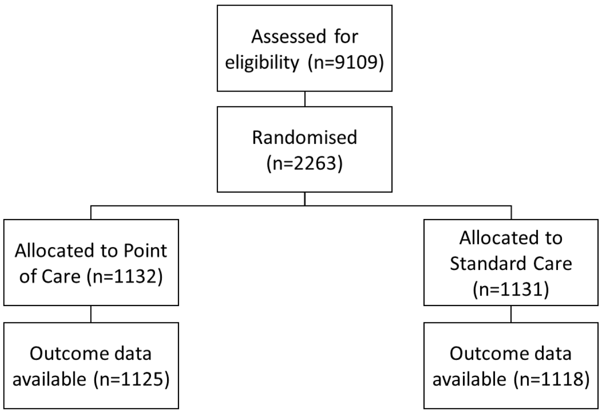

Assessed for eligibility (n=9109)
Randomised (n=2263)
Allocated to Point of Care (n=1132)
Allocated to Standard Care (n=1131)
Outcome data available (n=1125)
Outcome data available (n=1118)

Figure B.21: Flowchart of participants in the RATPAC trial

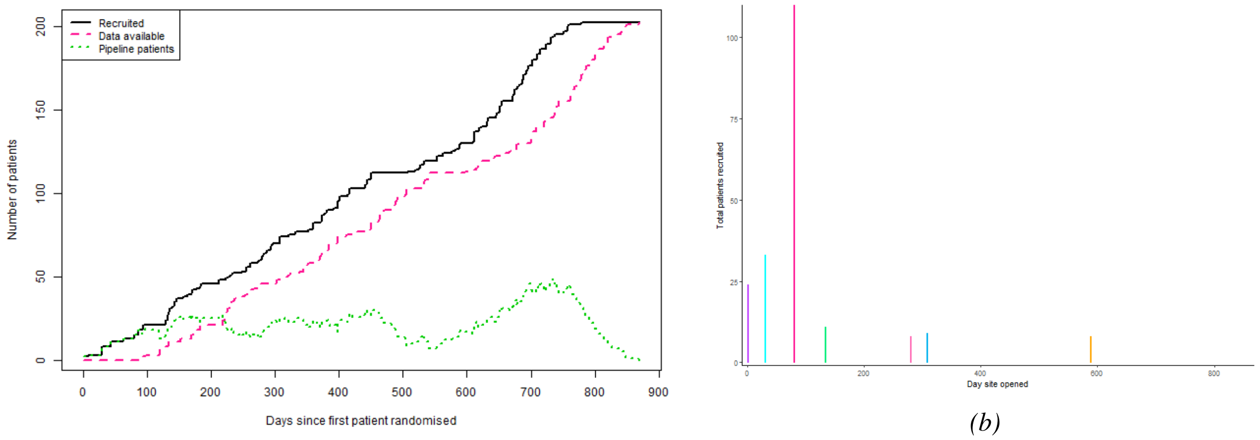
B.10 Corn plasters

Pain arising from foot problems is one of the main reasons for visiting a podiatrist. A large proportion of these patients complaints is due to corns. Even though a podiatrists workload is largely made up of treating corns, there is some uncertainty as to the best treatment to use. Typically, corns are removed with a scalpel, which needs to be repeated as the corn returns. However, plasters containing salicylic acid have been found to be effective at removing corns. Despite the evidence, podiatrists believe they could lead to complications.

The corn plaster study aimed to investigate usual care and salicylic plasters for corn removal. The primary outcome was the proportion of patients with a resolved corn at 3 months post randomisation. If more than one corn was present, patients were asked at baseline to choose one corn to be the ‘index’ corn to be monitored.

Assuming a recurrence rate of 40% in the plaster group and 60% in the scalpel group, with 80% power and two-sided 5% significance, 100 patients per arm would be required to detect an absolute risk reduction of 20%, or corresponding OR of 2.25. Between September 2009 and October 2011, 202 patients were randomised to the trial (n=101 in each arm). Figure B.22 shows patient and site recruitment to the Corn plasters trial.

The primary analysis was logistic regression, adjusted for site and size of corn at baseline. In the intervention group, 34% of corns had resolved, compared to 21% in the usual care group. The analysis found a statistically significant difference in favour of the treatment group (OR=2.00, 95% CI (1.02, 3.93), p=0.044). Figure B.23 shows the flow of patients in



(b)

(a)

Figure B.22: Recruitment to the Corn plasters trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Total days of study	869
Days to last patient recruited	779
Days to 50% patient recruitment	416
Number of sites	7
Days to last site recruited	588
Days to 50% site recruitment	134

Table B.10: Recruitment summary for the Corn plasters study

the corn plaster trial.

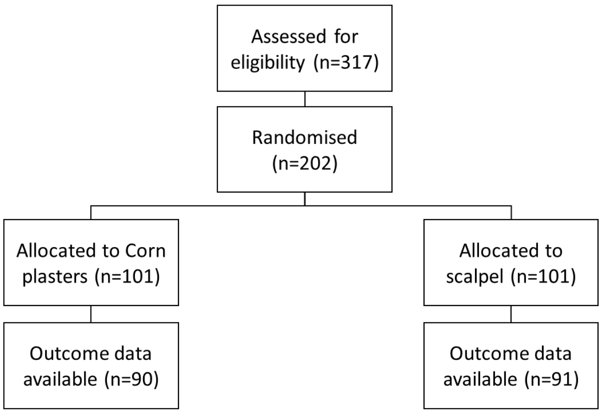

Assessed for eligibility (n=317)
Randomised (n=202)
Allocated to Corn plasters (n=101)
Allocated to scalpel (n=101)
Outcome data available (n=90)
Outcome data available (n=91)

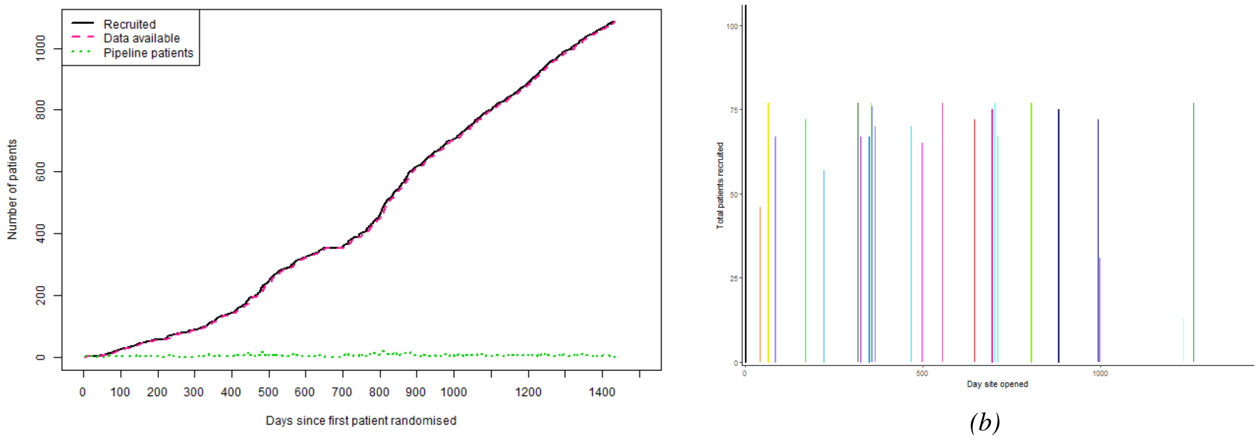
Figure B.23: Flowchart of participants in the Corn plasters trial

B.11 3MG

Acute asthma exacerbations, or asthma attacks, causes breathlessness,wheezing and chest tightness. Whilst largely managed in the primary care settings, more severe asthma attacks could require emergency treatment, and potential subsequent hospital admission (Fleetcroft 2016). Treatment prior to hospital, or within the emergency department may include steroids and/or bronchodilators. If a patient has a life threatening episode, or if symptoms persist, the patient is admitted to hospital. It is thought that the use of magnesium sulphate, given either intravenously or inhaled (nebuliser route) may benefit patients due to the anti-inflammatory action and muscle relaxation. However the true effect, and the optimal dose is unknown.

The 3MG trial was set up to investigate the effect of either intravenous or nebulised magnesium sulphate compared to a placebo group in adults with severe asthma. Two co-primary outcomes included the proportion of patients admitted to hospital within 1 week of presentation at the emergency department, and a VAS score for breathlessness assessed at 2 hours following initial treatment. Only the hospital admission outcome will be analysed in this re-analysis.

In order to detect a absolute risk reduction of 10% of hospital admissions, assuming 80% rate in the placebo group, 70% rate in treatment group, 90% power and 5% two-sided significance, 400 patients would be required in each group, or 1200 in total. The two treatment groups (nebulised and IV routes) were combined for a comparison of active vs placebo as a primary analysis. For pairwise comparisons between the three groups, Simes method was used to adjust for multiplicity. The trial was terminated early due to slow recruitment, and only 1109 patients were randomised between July 2008 and June 2012 from 25 sites (nebulised route n=339, IV route n=406, placebo n=364). In light of this, the trial had 87% power to detect the same absolute difference. Figure B.24 shows patient and site recruitment during the 3MG trial.



(a)

(b)

Figure B.24: Recruitment to the 3MG trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Total days of study	1434
Days to last patient recruited	1427
Days to 50% patient recruitment	847
Number of sites	25
Days to last site recruited	1262
Days to 50% site recruitment	467

Table B.11: Recruitment summary for the 3MG study

The primary outcome was analysed using logistic regression adjusted for centre. In total, 25 patients were excluded from the analysis (22 did not receive treatment and 3 for protocol violations discovered after receiving treatment). Figure B.25 shows patient flow through the 3MG trial. Similar admission rates were found between the placebo and nebulised magnesium sulfate group (78% and 79% respectively), but a lower rate (72%) in the IV group. The active group slightly decreased odds of a hospital admission compared to placebo but this was not statistically significant (OR=0.85, 95% CI (0.61,1.19, *p*=0.348).

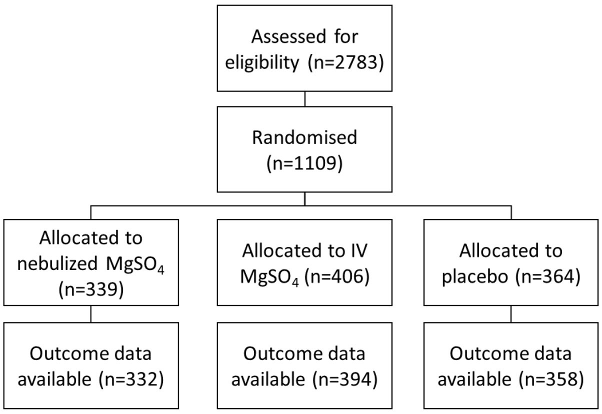

Assessed for eligibility (n=2783)
Randomised (n=1109)
Allocated to nebulized $MgSO_4$ (n=339)
Allocated to IV $MgSO_4$ (n=406)
Allocated to placebo (n=364)
Outcome data available (n=332)
Outcome data available (n=394)
Outcome data available (n=358)

Figure B.25: Flowchart of participants in the 3MG trial

B.11.1 Re-imagined time point of 3MG

The 3MG trial data will be used twice in the re-analysis in this thesis. The same primary outcome data will be used, but imagined at a 12 month time point, instead of the original 7 days. Figure B.26 shows the data availability and pipeline patients using this time point. Recruitment rate remains the same as the original trial, but now has a total trial duration of 1792 days.

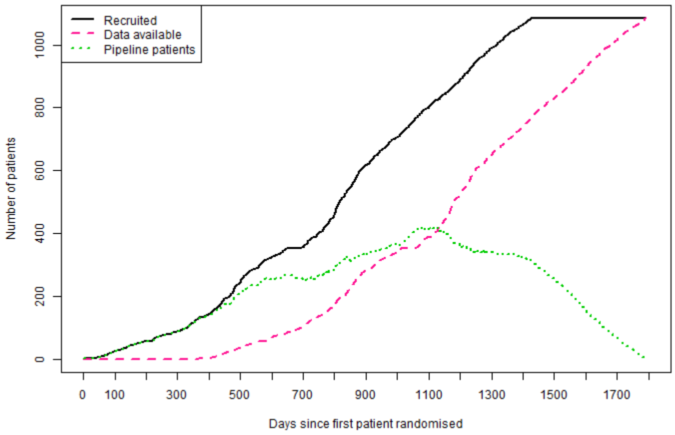


Figure B.26: Time between patients recruitment and primary outcome data became available. Green (dotted) line show pipeline patients who have been randomised, but have no data available yet.

B.12 AMAZE

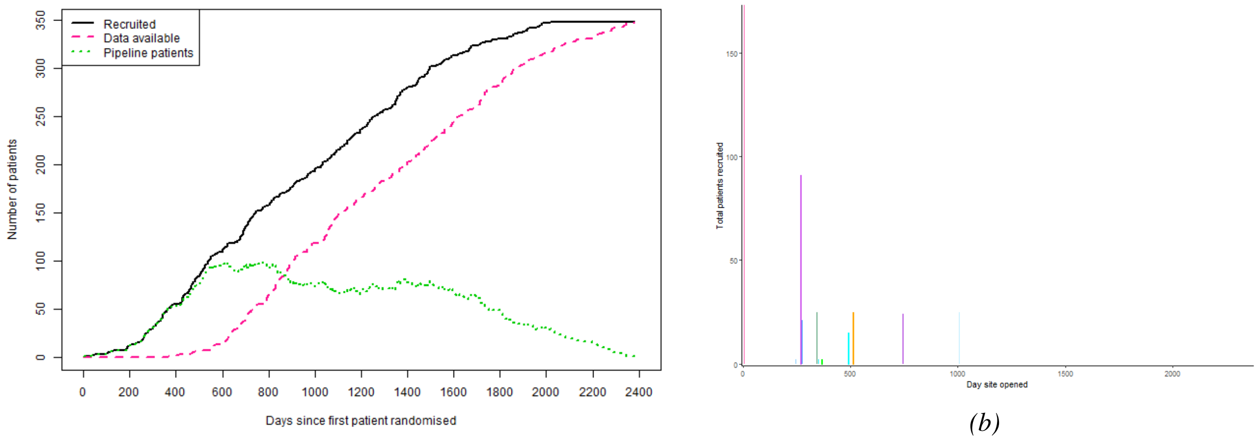
Arrhythmia, or irregular heartbeat, affects approximately 2 million people in the UK per year (NHS 2018); the most common of which is Atrial Fibrillation (AF). As well as causing chest pain, dizziness and other side effects, AF can also increase the risk of clot formation which, if dislodged, can result in stroke. Anticoagulants can be given to reduce the risk, but these drugs can increase the risk of bleeding. During cardiac surgery, a maze procedure can be performed to help the heart beat regularly again. The procedure involves complex techniques to remove the heart tissue that is causing the irregular heart beats (*American Heart Association* 2016), using devices such as radiofrequency, microwave energy or 'cut and sew' methods (Sharples 2018).

The AMAZE RCT was a NIHR funded Phase III, parallel arm RCT that recruited 352 patients undergoing cardiac surgery across 11 NHS specialist cardiac centres between February 2009 and March 2014 (Sharples 2018). Patients were randomised in a 1:1 ratio to either receive (i) planned cardiac surgery alone (control arm); or (ii) maze procedure and planned cardiac surgery (intervention arm).

Joint primary outcomes included return to sinus rhythm at 12 months after surgery, and quality-adjusted survival over two years. Only one primary outcome had to be significant for the maze procedure to be considered effective. Only the return to sinus rhythm outcome is considered for re-analysis in this thesis.

The recruitment target was set as 200 patients per arm to detect a target difference of 15% in the return to sinus rhythm between the two groups with 80% power and two-sided 5% significance level, assuming a 30% rate in the control group, and 45% rate in the intervention group. This allowed for up to 15% loss to follow up or death. Due to slow recruitment however, the trial was terminated early, after recruiting only 352 randomised patients (176 patients per treatment arm) between February 2009 and March 2014 from 11 UK NHS centres.

Figure B.27b describes the site recruitment, where each bar represent when a site was opened (x-axis), and total recruitment to that site (y-axis).



(a) *(b)*

Figure B.27: Recruitment to the AMAZE trial. Figure B.27a shows rate of recruitment and time to data availability. Figure B.27b shows site recruitment (day opened, and total patients recruited)

Total days of study	2381
Days to last patient recruited	2016
Days to 50% patient recruitment	887
Number of sites	311
Days to last site recruited	1006
Days to 50% site recruitment	349

Table B.12: Recruitment summary for the AMAZE study

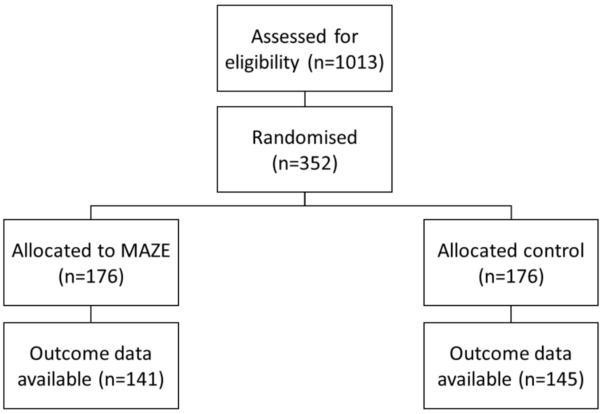

Assessed for eligibility (n=1013)
Randomised (n=352)
Allocated to MAZE (n=176)
Allocated control (n=176)
Outcome data available (n=141)
Outcome data available (n=145)

Figure B.28: Flowchart of participants in the AMAZE trial

Figure B.28 shows a summary of patient flow through the trial and details of missing data. Outcome data was available for 286 patients (intervention group n=141, control group, n=145).

Return to sinus rhythm was analysed using binary logistic regression, adjusted for baseline heart rhythm and planned surgical procedure (fixed effects), and surgeon (random effect). The OR for return to sinus rhythm in the maze group compared to the control group was 2.06 (95% CI (1.20, 3.54); p=0.0091) in the ITT analysis.

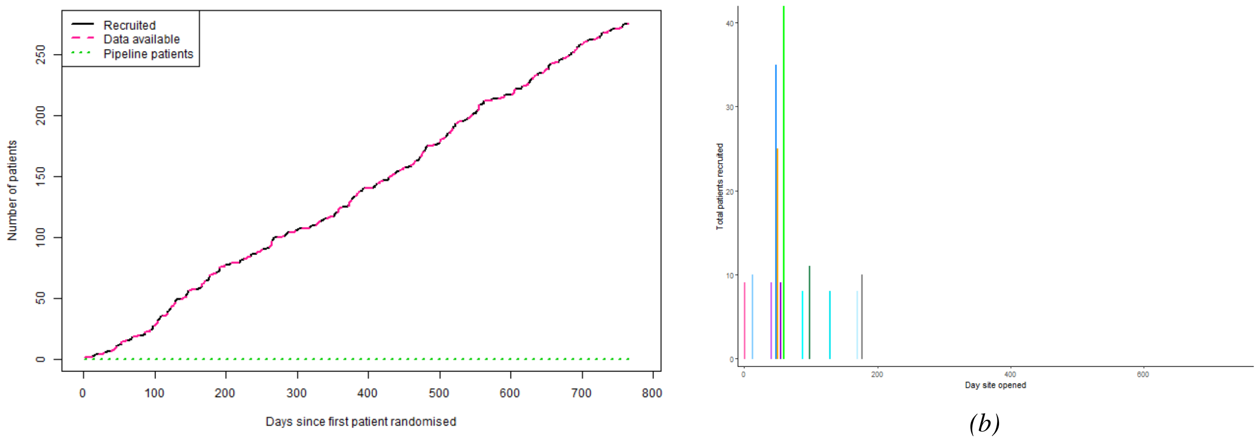
B.13 Nasal sprays - GSK

Allergic Rhinitis is an inflammatory disorder affecting the nose following exposure to an allergen such as pollen, dust or mould, affecting approximately 400 million people worldwide (Greiner 2011). Even though many may only suffer seasonally, symptoms can be persistent, and have a negative impact on quality of life. Current treatments can include intranasal corticosteroids, administered through a nasal spray. However, some patients have reported that the odour or after taste of some nasal sprays can decrease compliance to treatments (Yanez 2016).

A crossover trial (GSK-201474) was set up to determine patient preference between two types of nasal sprays: Fluticasone Furoate Nasal Spray (FFNS) vs Mometasone Furoate Nasal Spray (MFNS). Patients received both treatments, with a washout period of 30 minutes, but were randomly allocated which treatment they received first. The primary outcome was the overall preference of nasal spray. After receiving both treatments, patients were asked their overall preference, in which they could answer treatment 1, treatment 2, or no preference.

A previous study (GSK FFU108556), investigating FFNS and FP nasal sprays, was used to estimate the preference rate of the FFNS nasal spray (Meltzer 2008). Sample size calculations based on a one-sample chi square test, two-sided significance level of 5%, 90% power and a 60% preference rate for FFNS nasal spray (vs. 40% preferring MFNS or stating no preference), a total of 263 patients would be required. To ensure 263 evaluable subjects, the study aimed to recruit 300 patients.

The study met their target, recruiting 300 patients between April 2014 and May 2016 from 12 sites in 4 countries. Out of these 300, 276 were included in the primary outcome analysis. Figure B.29a shows the recruitment rate and data availability. Patient randomisation, treatment and primary outcome questionnaire all happened in the same day, and therefore this study had no pipeline patients.



(a)

(b)

Figure B.29: Recruitment to the nasal spray trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Figure B.29b describes the site recruitment, where each bar represents when a site was opened (x-axis), and total recruitment to that site (y-axis).

Total days of study	765
Days to last patient recruited	765
Days to 50% patient recruitment	387
Number of sites	12
Days to last site recruited	176
Days to 50% site recruitment	54

Table B.13: Recruitment summary for the nasal spray study

Primary analysis used a Cochran-Mantel-Haenszel test. The study found a statistically significant result (p<0.001), favouring the FFNS (56%), compared to MFNS (32%) or no preference between the two treatments (12%).

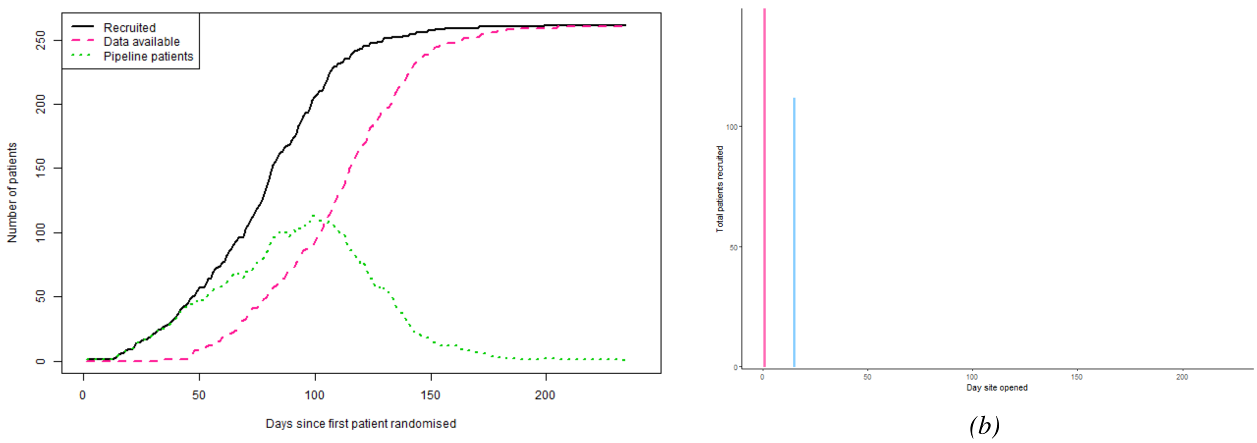
B.14 Meningitis vaccine - GSK

Meningococcal bacteria, found in the nose and throat, is generally harmless, with approximately 10% of the global population carrying the disease without any symptoms (Van Deuren 2000). However, the bacteria could enter the bloodstream (septicaemia), move to the meninges or outer lining of the brain/spinal chord (meningitis), or both (meningococcal disease) (Peate 2011). There are a number of strains of meningococcal bacteria, namely A, B, C, W, X and Y (Foundation 2020). Mencevax ACWY is a vaccine for meningococcal bacteria strains A and C developed by GlaxoSmithKline (GSK), who conducted a phase III booster vaccination study in 2006 (GSK-105239). This study followed on from a previous trial, where babies were randomised to receive in a 2:1 ratio either DTPw-HBV/Hib-MenAC vaccine ("AC primed" group), or DTPw-HBV/Hib vaccine ("AC unprimed" group) at 6, 10 and 14 weeks of age. Between 24 and 30 months of age, the child was asked to participate in the booster trial, where all patients received a full dose of Mencevax ACWY. One month following this vaccination, blood samples were taken, and Serum Bactericidal Antibody (SBA)s were compared between the primed and unprimed groups in the original study. A titer cut-off value of $\geq$1:128 meant a positive outcome.

A total of 517 patients were included in the primary study. Assuming 50% of patients would continue to the booster study, approximately 255 patients would be evaluable ($\sim$ 170 in the primed group, and $\sim$ 85 in the unprimed group). With an anticipated reference rate of 98%, a non-inferiority limit of 0.1 and 2.5% significance, the study had 99% power for each primary outcome (SBA-MenA and SBA-MenC).

Between May and November 2006, 261 patients were enrolled from 2 sites to the booster stage of the study and were included in the immunogenicity analysis. The time to primary outcome was approximately 1 month. Figure B.30 shows patients and site recruitment and time to data availability, and Table B.14 summarises key recruitment milestones.

All patients had a SBA titers $\geq$128 at one month following the full dose of Mencevax, in both the primed and unprimed groups. Because of this, data from one month earlier (at the time of full dose vaccination visit) will be used as though it is the primary outcome data.



(a)

(b)

Figure B.30: Recruitment to the Mencevax trial. (a) shows rate of recruitment and time to data availability. (b) shows site recruitment (day opened, and total patients recruited)

Total days of study	234
Days to last patient recruited	199
Days to 50% patient recruitment	79
Number of sites	2
Days to last site recruited	15
Days to 50% site recruitment	1

Table B.14: Recruitment summary for the Mencevax study

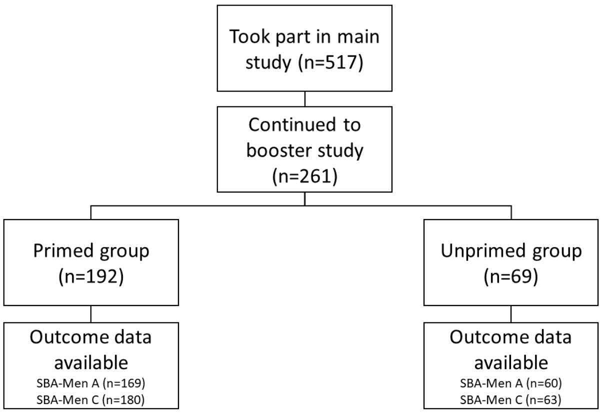


Figure B.31: Flowchart of participants in the Mencevax trial

Only the A strain data will be re-analysed at the 1 month analysis. SBA-Men C data will be used for a re-imagined time point at 1 year instead of the original 1 month. Figure B.32 shows the data availability for the re-imagined time point.

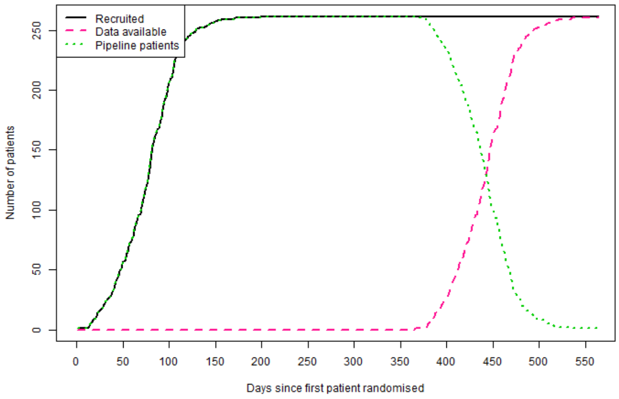


Figure B.32: Time between patients recruitment and primary outcome data became available. Green (dotted) line show pipeline patients who have been randomised, but have no data available yet.

Original analysis calculated asymptotic 95% CI for the difference in proportions. For the SBA-Men A outcome, the primed group had 157/168 (92.9%) with a titer $\geq$ 128 (95% CI [87.9, 96.3]) and compared to 55/60 (91.7%) in the unprimed group (95% CI [81.6, 97.2]). For the SBA-Men C outcome, the primed group had 98/180 (54.4%) with a titer $\geq$128 (95% CI [46.9,61.9]) compared to 16/63 (25.4%) 95% CI in the unprimed group. Figure B.31 shows patient flow through the booster trial and details of missing data.

C | Data re-analysis results

This appendix includes the SSR design comparison and CP plots from the results of the retrospective data analysis section (Chapter 5) for all trials including design comparisons and estimate stability results.

In the cases where trial data is used at one or more time points, CP values remain the same, and therefore promising zone and stepwise designs yield the same graphs. Therefore, only one n^* graph is presented, with any differences in combination test design plotted in purple and/or orange (see figure legends for details).

C.1 FAST INdiCATE

Figure C.1 shows CP calculated after every patient from patient 10 onwards assuming the four different treatment effects. All four lines have very high CP values in the very early stages of the trial. The current trend decreases very early, remaining in the unfavourable zone from patient 34 onwards, with the exception of one small peak in the promising zone at patient 119. Conversely, the hypothesised effect assumption means CP stays in the favourable zone until patient 235 (with a small exception for patient 228 and 229). By 280 patients, all four lines have decreased to 0% CP. The current trend and optimistic confidence limits fluctuate substantially throughout the trial, indicating random highs and lows of the treatment estimate. The alternative hypothesis assumption consistently yields higher CP values throughout the trial, while the current trend assumption is consistently lower.

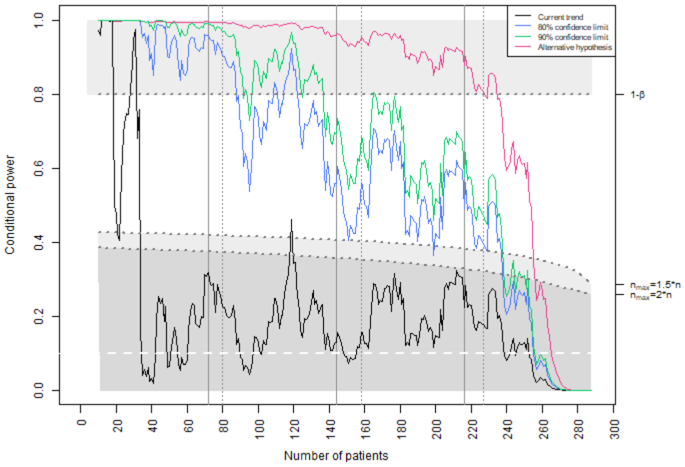


Figure C.1: Conditional power calculated after every patient in the FAST INdiCATE trial

Table C.2 presents results for three designs (promising zone, combination test and stepwise) for the FAST INdiCATE trial, based on CP values at three time points (25, 50 and 75% of data becoming available). Alternatively, results for the same time points but for percentage of patients recruited as opposed to data being available are presented in Table C.2. The table presents the new sample size required for each design, the zone the CP value falls in for the promising zone design, and the minimum CP value for a sample size increase for the promising zone and stepwise designs.

The promising zone and stepwise designs remain at 288 patients using the current trend and hypothesised treatment effect. An increase in sample size in both designs occur at the 50 and 75% time point using either optimistic limit. The largest increase in promising zone design happens at 50% available and the 80% limit, with an increase of 42%, and in the stepwise design, a 2-fold increase is seen with 75% data available using the 90% limit. The combination test design new sample size ranges from 158, a reduction of 130 patients (45% of the original sample size), observed using the current trend, to increasing to 471 patients (a 64% increase in sample size), observed using the 80% optimistic limit.

Whilst no time point would have reached the 10% futility bound investigated and therefore would not have stopped early for futility, the current trend line does have three main dips below this value (total of 33 times) between patient 35 and 160. All treatment assumptions drop below this futility boundary by patient 265. Table C.1 summarises the total number of times CP values fall in each zone.

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	13	223	88	124
	Promising	10	22	139	104
	Unfavourable	182	10	18	19
	Futility	74	24	34	32
n_{max}=2	Favourable	13	223	88	124
	Promising	13	23	140	112
	Unfavourable	179	9	17	11
	Futility	74	24	34	32

Table C.1: Number of times CP values fall in each zone for the promising zone design for the FAST INdiCATE trial. For a design where no futility boundary is considered, these values become unfavourable instead.

A graphical representation of the new sample size calculated after every patient is seen in Figure C.2. n^* reaches n_{max}=2 for the combination test in all assumptions except current trend, where only n_{max}=1.5 is reached. The stepwise and promising zone designs increase sample size very early on in the trial using current trend, and again at around 120 patients, with CP briefly falling into the promising zone at this point. Conversely, sample size increases later on through the trial (from around patient 100) for either optimistic limit, and again much later (from around patient 220) using the hypothesised treatment effect.

			Promising zone			Combination test	Stepwise design	
	Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND								
25% Available: 72	n_{max}=432	0.318	0.419	Unfavourable	288	293	0.466	288
Recruited:80	n_{max}=576	0.318	0.374	Unfavourable	288	293	0.440	288
50% Available: 144	n_{max}=432	0.128	0.406	Unfavourable	288	158	0.460	288
Recruited: 158	n_{max}=576	0.128	0.357	Unfavourable	288	158	0.430	288
75% Available: 216	n_{max}=432	0.224	0.382	Unfavourable	288	380	0.447	288
Recruited: 227	n_{max}=576	0.224	0.328	Unfavourable	288	380	0.410	288
HYPOTHESISED EFFECT								
25% Available: 72	n_{max}=432	0.996	0.419	Favourable	288	210	0.466	288
Recruited:80	n_{max}=576	0.996	0.374	Favourable	288	210	0.440	288
50% Available: 144	n_{max}=432	0.960	0.406	Favourable	288	290	0.460	288
Recruited: 158	n_{max}=576	0.960	0.357	Favourable	288	290	0.430	288
75% Available: 216	n_{max}=432	0.877	0.382	Favourable	288	332	0.447	288
Recruited: 227	n_{max}=576	0.877	0.328	Favourable	288	332	0.410	288
80% OPTIMISTIC LIMIT								
25% Available: 72	n_{max}=432	0.960	0.419	Favourable	288	265	0.466	288
Recruited:80	n_{max}=576	0.960	0.374	Favourable	288	265	0.440	288
50% Available: 144	n_{max}=432	0.558	0.406	Promising	409	432	0.460	384
Recruited: 158	n_{max}=576	0.558	0.357	Promising	409	462	0.430	480
75% Available: 216	n_{max}=432	0.493	0.382	Promising	410	432	0.447	336
Recruited: 227	n_{max}=576	0.493	0.328	Promising	410	471	0.410	384
90% OPTIMISTIC LIMIT								
25% Available: 72	n_{max}=432	0.991	0.419	Favourable	288	225	0.466	288
Recruited:80	n_{max}=576	0.991	0.374	Favourable	288	225	0.440	288
50% Available: 144	n_{max}=432	0.695	0.406	Promising	333	406	0.460	432
Recruited: 158	n_{max}=576	0.695	0.357	Promising	333	406	0.430	576
75% Available: 216	n_{max}=432	0.576	0.382	Promising	361	432	0.447	384
Recruited: 227	n_{max}=576	0.576	0.328	Promising	361	435	0.410	480

Table C.2: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

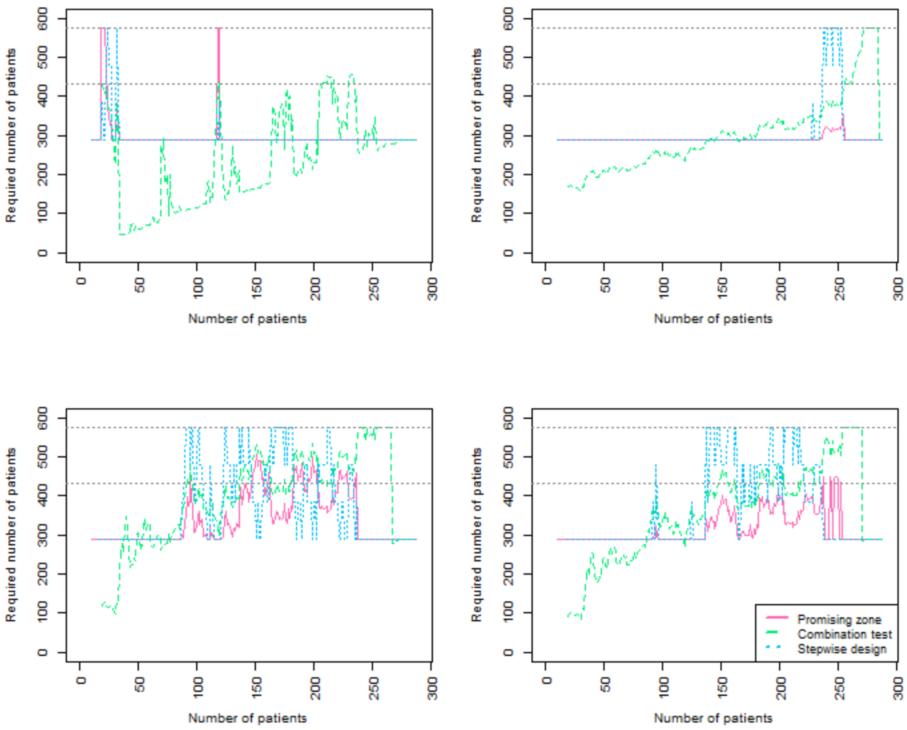
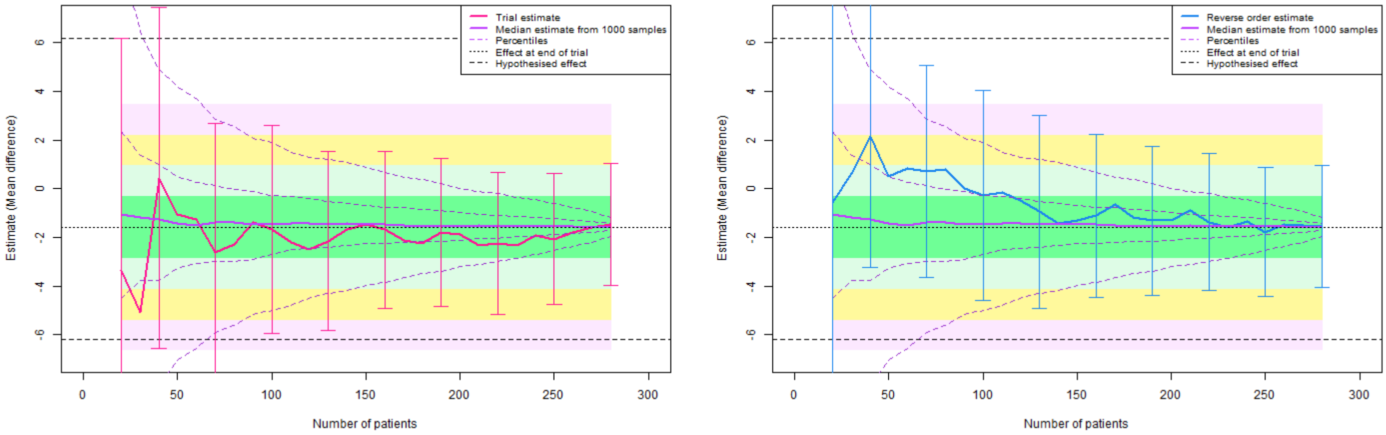


Figure C.2: Comparison of three SSR designs for the FAST INdiCATE trial data

Figure C.3 shows the treatment effect (mean difference) in both original sequential order (a) and reverse order (b). As neither treatment arm is a control group, there are two lines to represent the hypothesised treatment effect, ±6.2.

The original sequential order starts by under-estimating the treatment effect, but lies within the ± 1*SE boundary by patient 50 (17% through the trial), and remains there for the remainder of the trial. The sequential estimate briefly falls into the ±3*SE boundary at 30 patients, but remains within the ±2*SE limit at all other values of n. Comparatively, the reverse order estimate spends a little longer in the ±3*SE boundary, over-estimating the treatment effect to start. However, it slowly decreases, entering the ±1*SE boundary at 120 patients, and remaining there from that point forwards.



(a) *(b)*

Figure C.3: Stability of the estimate in the FAST INdiCATE trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.2 Acupuncture trial

Figure C.4 shows CP calculated after every patient from patient 10 onwards assuming the four different treatment effects. The current trend line is most frequently below CP_{min} values, consistently below this line between patients 53 and 181/189 for n_{max}=1.5/2 respectively. CP largely becomes promising after this point, but falls back into the unfavourable zone twice more. The hypothesised effect line (red) mainly stays in the favourable zone ($> 1-\beta$), but intermittently dropping below the favourable line 30 times between patient 132-178. All four lines drop to zero by the very end of the trial. Current trend and optimistic limit lines fluctuate throughout the duration of the study, indicating random highs and lows. Table C.3 summarises the number of times each line falls into the four zones (with futility zone being included with unfavourable if no stopping boundary is being used).

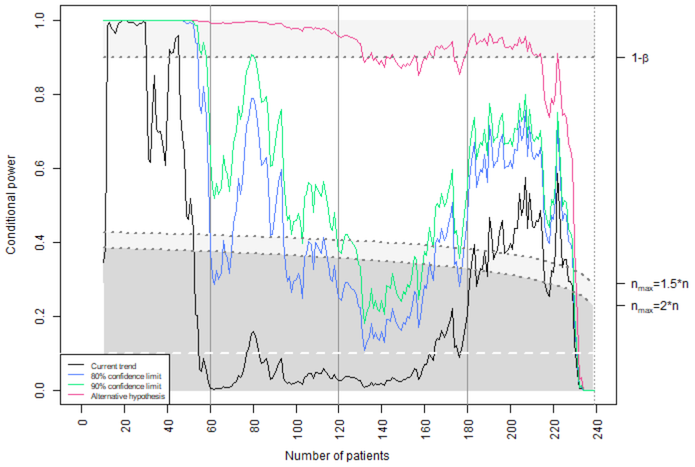


Figure C.4: Conditional power calculated after every patient in the Acupuncture trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	24	176	45	51
	Promising	46	45	89	127
	Unfavourable	45	1	87	43
	(Futility)	115	8	9	9
n_{max}=2	Favourable	24	176	45	51
	Promising	59	45	105	140
	Unfavourable	32	1	71	30
	(Futility)	115	8	9	9

Table C.3: Number of times CP values fall in each zone for the promising zone design for the Acupuncture trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table C.4 summarises the decisions for all three designs at the three specified interim analysis time points. Under the current trend assumption, the decision at either interim analysis 1 or 2 would be to stop the trial for futility. By interim analysis 3 (75% data available) the CP is now above the futility bound, but still lies in the unfavourable zone. Using either optimistic limit, there would be some increases in sample size in all three designs. The largest increase would have been the maximum of 478, twice the original sample size. This can be seen in the promising zone and combination test designs at 25% (80% limit), and 50% with the 90% limit (477 patients for the combination test (99.6% increase)); and at 25% for the 90% limit using the stepwise design. The combination test has no decreases in patients at any of the interim time points

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 60	n_{max}=359	0.006	0.419	Unfavourable	239	239	0.466	239
	Recruited: 239	n_{max}=478	0.006	0.374	Unfavourable	239	239	0.440	239
50%	Available: 120	n_{max}=359	0.026	0.406	Unfavourable	239	239	0.460	239
	Recruited: 239	n_{max}=478	0.026	0.357	Unfavourable	239	239	0.430	239
75%	Available: 180	n_{max}=359	0.208	0.382	Unfavourable	239	359	0.446	239
	Recruited: 239	n_{max}=478	0.208	0.327	Unfavourable	239	398	0.409	239
HYPOTHESISED EFFECT									
25%	Available: 60	n_{max}=359	0.995	0.419	Favourable	239	239	0.466	239
	Recruited:239	n_{max}=478	0.995	0.374	Favourable	239	239	0.440	239
50%	Available: 120	n_{max}=359	0.956	0.406	Favourable	239	252	0.460	239
	Recruited: 239	n_{max}=478	0.956	0.357	Favourable	239	252	0.430	239
75%	Available: 180	n_{max}=359	0.921	0.382	Favourable	239	268	0.446	239
	Recruited: 239	n_{max}=478	0.921	0.327	Favourable	239	268	0.409	239
80% OPTIMISTIC LIMIT									
25%	Available: 60	n_{max}=359	0.376	0.419	Unfavourable	239	359	0.466	239
	Recruited:239	n_{max}=478	0.376	0.374	Promising	478	478	0.440	239
50%	Available: 120	n_{max}=359	0.250	0.406	Unfavourable	239	359	0.460	239
	Recruited: 239	n_{max}=478	0.250	0.357	Unfavourable	239	478	0.430	239
75%	Available: 180	n_{max}=359	0.469	0.382	Promising	359	359	0.446	279
	Recruited: 239	n_{max}=478	0.469	0.327	Promising	423	419	0.409	319
90% OPTIMISTIC LIMIT									
25%	Available: 60	n_{max}=359	0.622	0.419	Promising	359	359	0.466	359
	Recruited:239	n_{max}=478	0.622	0.374	Promising	394	388	0.440	478
50%	Available: 120	n_{max}=359	0.378	0.406	Unfavourable	239	359	0.460	239
	Recruited: 239	n_{max}=478	0.378	0.357	Promising	478	477	0.430	239
75%	Available: 180	n_{max}=359	0.552	0.382	Promising	359	359	0.446	319
	Recruited: 239	n_{max}=478	0.552	0.327	Promising	362	382	0.409	399

Table C.4: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

Figure C.5 shows the new total sample size required calculated after every patient. A small decrease in sample size using the combination test can be seen very early on using any treatment effect assumption. Current trend requires an increase if looking at the beginning or end of the trial, but remains at the original sample size between ≈ 60 - 160 patients. The hypothesised effect sees much smaller increases in sample size, until near the end of the trial. Sample size n^* fluctuates much more using either of the optimistic confidence limits.

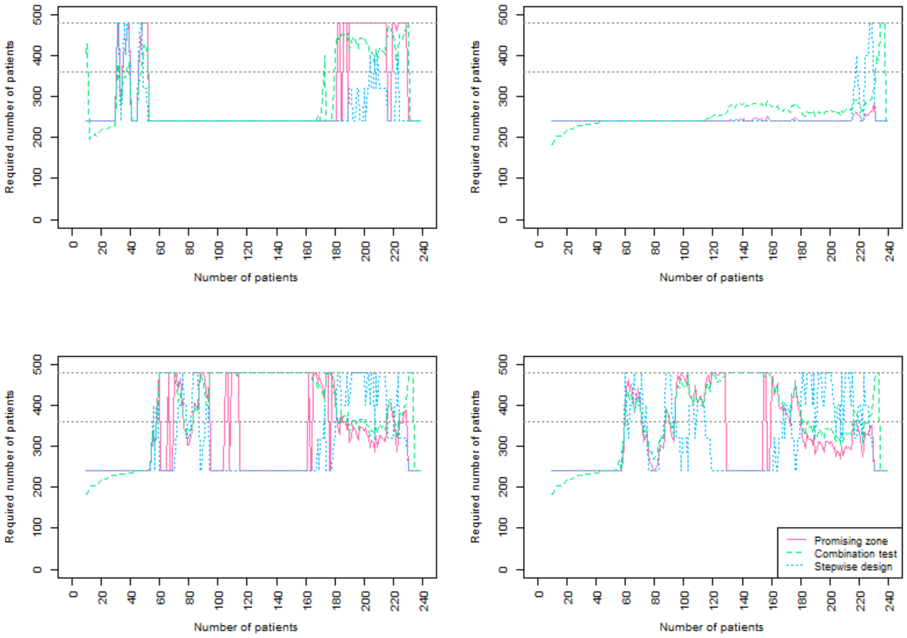
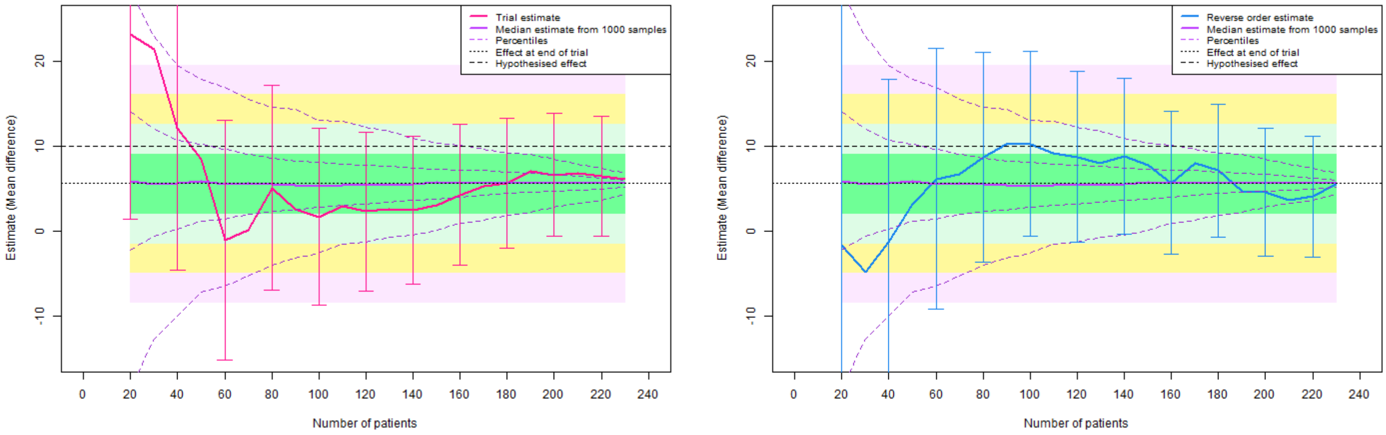


Figure C.5: Comparison of three SSR designs for the Acupuncture trial data

Figure C.6 shows the mean difference from the Acupuncture trial, calculated after every 10 patients from patient 20, in original (a) and reversed order (b). The original order starts by over-estimating the treatment effect seen at the end of the trial, falling just outside the largest boundary investigated (±4*SE). The estimate decreases, reaching the lower -2*SE boundary by patient 60. The estimate first enters the ±1*SE boundary at 50 patients (21% through the trial). By 110 patients (46% through the trial) the estimate re-enters this boundary, and remains there until the end of the trial. The reverse order starts by under-estimating the treatment effect, but always remains within the ±3*SE (yellow) boundary. Again, the estimate first reaches the ±1*SE boundary by patient 50 (21%), but only remains within this boundary from patient 120 onwards (50%).



(a)

(b)

Figure C.6: Stability of the estimate in the Acupuncture trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.3 SELF trial

Figure C.7 shows CP for the SELF trial. The current trend assumption remains consistently low throughout the trial duration, which coincides with no sample size increase for promising zone or stepwise designs (Figure C.8), and a decrease in sample size with the combination test design. The decision at any interim analysis would have been to stop the trial for futility, if this bound was used. Using the hypothesised effect assumption, CP starts in the favourable zone and gradually drops to zero by patient 65. The optimisitc 80% limit line fluctuates between unfavourable and promising up until patient 29, falling to zero by patient 66. The 90% limit sees a similar pattern, but reaches the favourable zone twice before gradually falling to zero by patient 66 also. For both limits, no increase in sample size is seen above patient 40 with the promising zone design, and only reaches the maximum of twice the sample size using the 80% limit using the promising zone design. Combination test design sees the biggest number of times increasing the trial for both limits, and for the hypothesised effect. Stepwise and promising zone designs are broadly similar for the optimistic limit plots.

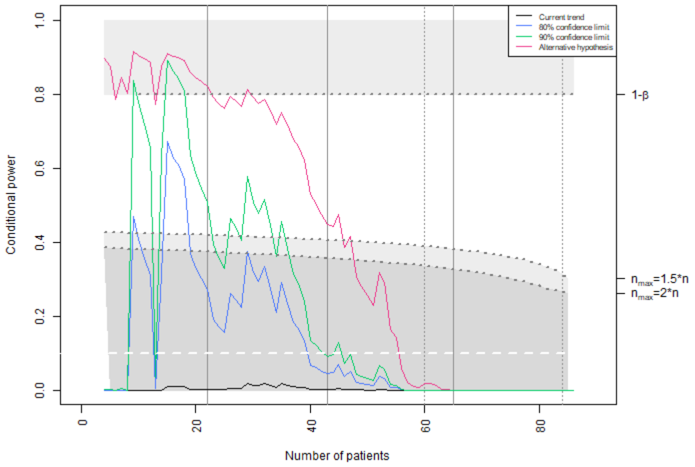
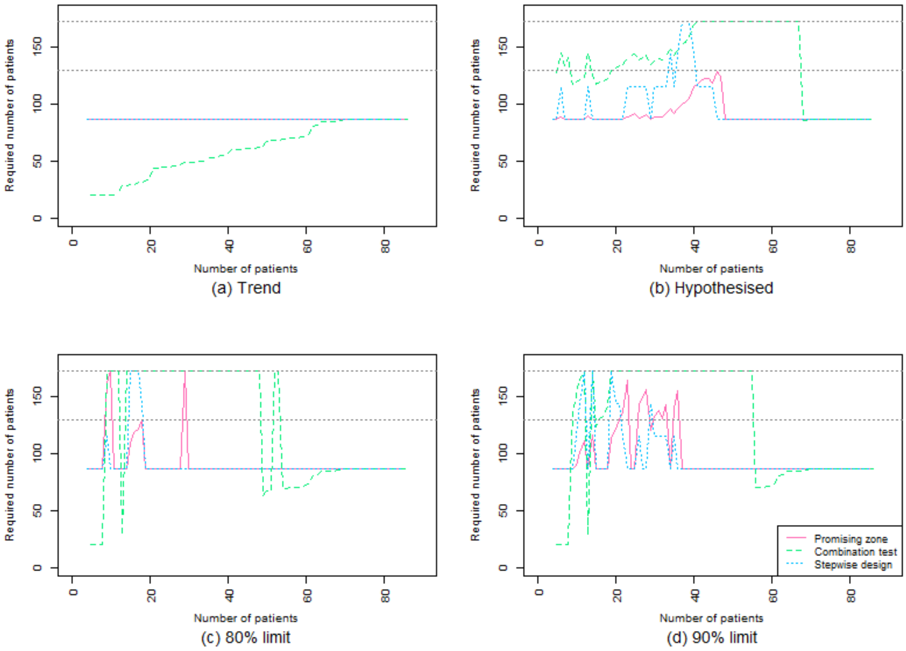


Figure C.7: Conditional power calculated after every patient in the SELF trial



(a) Trend

(b) Hypothesised

(c) 80% limit

(d) 90% limit

Figure C.8: Comparison of three SSR designs for the SELF trial data

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	0	18	0	5
	Promising	0	25	5	16
	Unfavourable	0	9	25	13
	(Futility)	83	31	53	49
n_{max}=2	Favourable	0	18	0	5
	Promising	0	26	7	19
	Unfavourable	0	8	23	10
	(Futility)	83	31	53	49

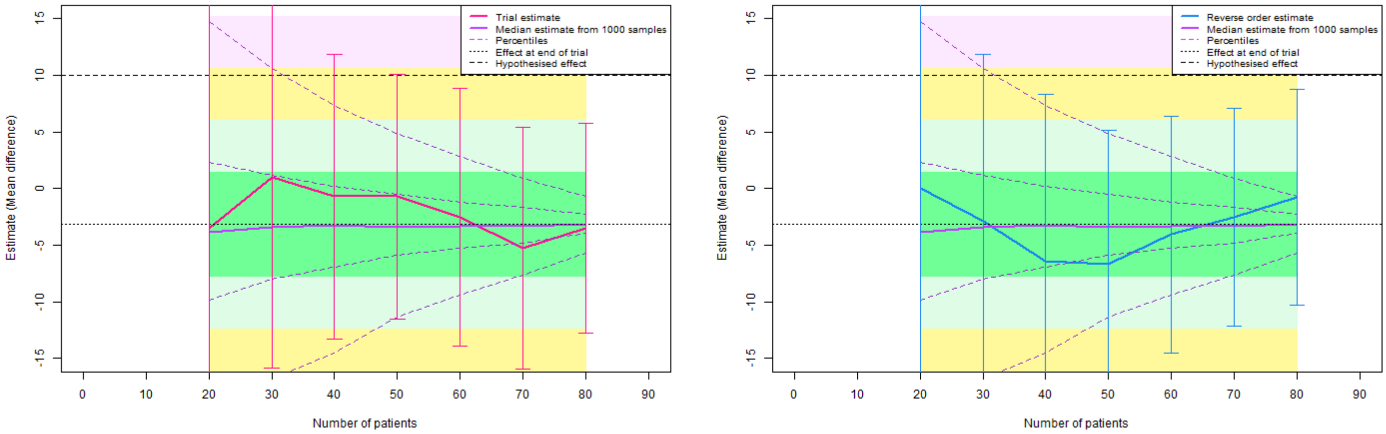
Table C.5: Number of times CP values fall in each zone for the promising zone design for the SELF trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table C.5 summarises the number of times each line falls into each of the four zones, and Table C.6 summarises the decisions for the three designs at the three specified interim analyses. The stepwise design increases the sample size by a maximum of 34% (25% time point using hypothesised effect, and 25% time point using 90% optimistic limit). At the same time points, promising zone also increases, with the maximum being 59% increase. The combination test ranges from a maximum decrease in sample size of 50% using the current trend assumption, to an increase of twice the sample size using either optimisitc limit or hypothesised effect assumptions.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 22	n_{max}=129	0.003	0.419	Unfavourable	86	43	0.465	86
	Recruited: 43	n_{max}=172	0.003	0.374	Unfavourable	86	43	0.440	86
50%	Available: 43	n_{max}=129	0.001	0.406	Unfavourable	86	60	0.458	86
	Recruited: 60	n_{max}=172	0.001	0.357	Unfavourable	86	60	0.429	86
75%	Available: 65	n_{max}=129	0.000	0.382	Unfavourable	86	84	0.444	86
	Recruited: 84	n_{max}=172	0.000	0.327	Unfavourable	86	84	0.409	86
HYPOTHESISED EFFECT									
25%	Available: 22	n_{max}=129	0.823	0.419	Favourable	86	129	0.465	86
	Recruited: 43	n_{max}=172	0.823	0.374	Favourable	86	135	0.440	86
50%	Available: 43	n_{max}=129	0.449	0.406	Promising	122	129	0.458	86
	Recruited: 60	n_{max}=172	0.449	0.357	Promising	122	172	0.429	115
75%	Available: 65	n_{max}=129	0.000	0.382	Unfavourable	86	84	0.444	86
	Recruited: 84	n_{max}=172	0.000	0.327	Unfavourable	86	172	0.409	86
80% OPTIMISTIC LIMIT									
25%	Available: 22	n_{max}=129	0.273	0.419	Unfavourable	86	129	0.465	86
	Recruited: 43	n_{max}=172	0.273	0.374	Unfavourable	86	172	0.440	86
50%	Available: 43	n_{max}=129	0.045	0.406	Unfavourable	86	129	0.458	86
	Recruited: 60	n_{max}=172	0.045	0.357	Unfavourable	86	172	0.429	86
75%	Available: 65	n_{max}=129	0.000	0.382	Unfavourable	86	84	0.444	86
	Recruited: 84	n_{max}=172	0.000	0.327	Unfavourable	86	84	0.409	86
90% OPTIMISTIC LIMIT									
25%	Available: 22	n_{max}=129	0.507	0.419	Promising	129	129	0.465	101
	Recruited: 43	n_{max}=172	0.507	0.374	Promising	137	172	0.440	115
50%	Available: 43	n_{max}=129	0.092	0.406	Unfavourable	86	129	0.458	86
	Recruited: 60	n_{max}=172	0.092	0.357	Unfavourable	86	172	0.429	86
75%	Available: 65	n_{max}=129	0.000	0.382	Unfavourable	86	84	0.444	86
	Recruited: 84	n_{max}=172	0.000	0.327	Unfavourable	86	84	0.409	86

Table C.6: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

The SELF trial sampling estimate from patient 20 onwards is shown in Figure C.9. The estimate remains inside the ±1*SE boundary throughout the trial duration, starting almost exactly at the end treatment effect. The reverse order also remains within ±1*SE of the assumed true treatment effect, starting off with a slight overestimate. All estimates are well below the originally assumed mean difference of +10.



(a)

(b)

Figure C.9: Stability of the estimate in the SELF trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.4 CASPER MINUS trial

Figure C.10 shows CP after every patient in the CASPER MINUS trial (re-imagined time point of the CASPER trial). Both optimistic confidence limits and hypothesised treatment effect assumption lines are (≈) 1 throughout the trial duration. Under the current trend, the CP line starts low (zero) and gradually increases to 1 by patient 301, fluctuating first between unfavourable and promising zones before patient 121, and between promising and favourable zones until patient 300. Table C.7 summarises the number of times CP falls into each zone between patient 12 and 705.

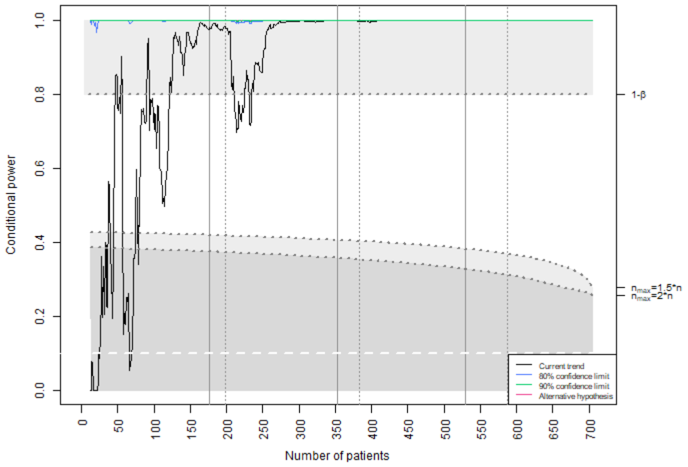


Figure C.10: Conditional power calculated after every patient in the CASPER MINUS trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	572	694	694	694
	Promising	74	0	0	0
	Unfavourable	31	0	0	0
	(Futility)	17	0	0	0
n_{max}=2	Favourable	572	694	694	694
	Promising	76	0	0	0
	Unfavourable	29	0	0	0
	(Futility)	17	0	0	0

Table C.7: Number of times CP values fall in each zone for the promising zone design for the CASPER MINUS trial. For a design where no futility boundary is considered, these values become unfavourable instead.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 177	n_{max}=1058	0.98	0.419	Favourable	705	400	0.466	705
	Recruited:198	n_{max}=1410	0.98	0.374	Favourable	705	400	0.440	705
50%	Available: 353	n_{max}=1058	1	0.406	Favourable	705	395	0.460	705
	Recruited: 383	n_{max}=1410	1	0.357	Favourable	705	395	0.430	705
75%	Available: 529	n_{max}=1058	1	0.382	Favourable	705	587	0.447	705
	Recruited: 587	n_{max}=1410	1	0.328	Favourable	705	587	0.410	705
HYPOTHESISED EFFECT									
25%	Available: 177	n_{max}=1058	1	0.419	Favourable	705	314	0.466	705
	Recruited:198	n_{max}=1410	1	0.374	Favourable	705	314	0.440	705
50%	Available: 353	n_{max}=1058	1	0.406	Favourable	705	385	0.460	705
	Recruited: 383	n_{max}=1410	1	0.357	Favourable	705	385	0.430	705
75%	Available: 529	n_{max}=1058	1	0.382	Favourable	705	587	0.447	705
	Recruited: 587	n_{max}=1410	1	0.328	Favourable	705	587	0.410	705
80% OPTIMISTIC LIMIT									
25%	Available: 177	n_{max}=1058	1	0.419	Favourable	705	305	0.466	705
	Recruited:198	n_{max}=1410	1	0.374	Favourable	705	305	0.440	705
50%	Available: 353	n_{max}=1058	1	0.406	Favourable	705	386	0.460	705
	Recruited: 383	n_{max}=1410	1	0.357	Favourable	705	386	0.430	705
75%	Available: 529	n_{max}=1058	1	0.382	Favourable	705	587	0.447	705
	Recruited: 587	n_{max}=1410	1	0.328	Favourable	705	587	0.410	705
90% OPTIMISTIC LIMIT									
25%	Available: 177	n_{max}=1058	1	0.419	Favourable	705	288	0.466	705
	Recruited:198	n_{max}=1410	1	0.374	Favourable	705	288	0.440	705
50%	Available: 353	n_{max}=1058	1	0.406	Favourable	705	384	0.460	705
	Recruited: 383	n_{max}=1410	1	0.357	Favourable	705	384	0.430	705
75%	Available: 529	n_{max}=1058	1	0.382	Favourable	705	587	0.447	705
	Recruited: 587	n_{max}=1410	1	0.328	Favourable	705	587	0.410	705

Table C.8: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

Table C.8 reports interim decisions under each assumption. By the time 25% of the data is available, CP has reached $\approx$ 1 for all assumptions. Therefore, even under the current trend, the promising zone and stepwise designs would continue to the original sample size planned (705). In all cases, the combination test would have decreased the sample size, from 41% (90% limit assumption) to 83% (all assumptions) of the original sample size. Figure C.11 expands on this, calculating a new n^* after every patient with data available. The combination test design (green) is always $\leq$705, generally consistently increasing using all assumptions, with some larger fluctuations towards the beginning of the trial using the current trend assumption. Only the current trend plot increases sample size for promising zone and stepwise designs, but no increase is seen past the first 250 patients.

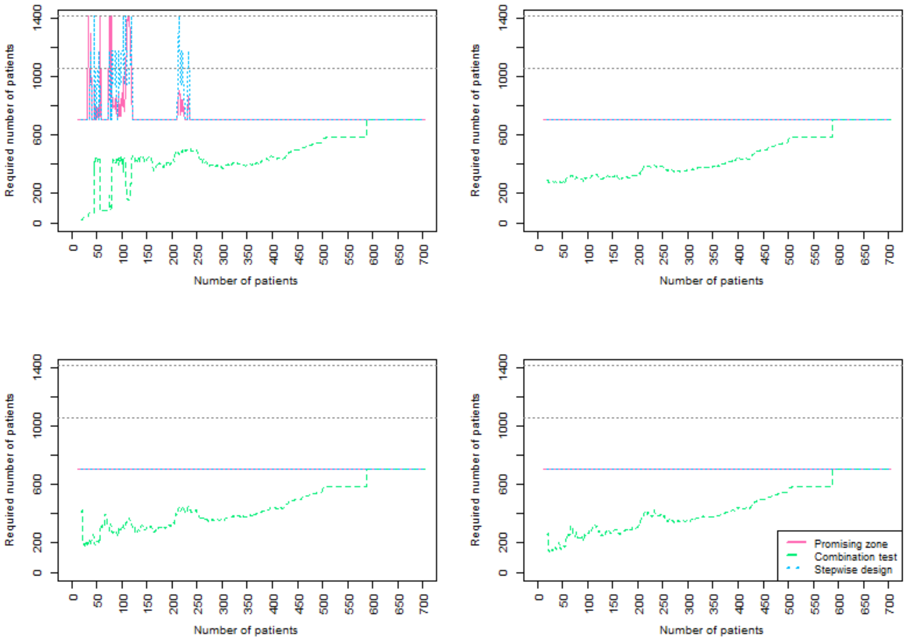
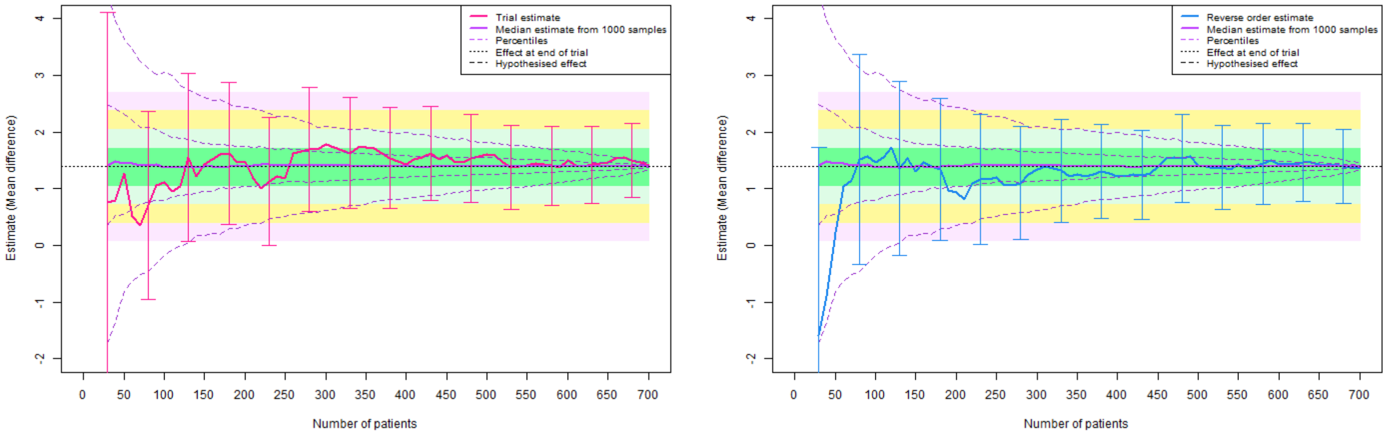


Figure C.11: Comparison of three SSR designs for the CASPER MINUS trial data

The CASPER MINUS trial starts with an under-estimate of the treatment effect for both the original and reverse order (Figure C.12), particularly emphasised in the reverse order estimate. The original order stays within the ±4*SE boundary, as opposed to the reverse order which starts outside the largest boundary considered. The original sequential estimate first enters the ±1*SE boundary at patient 50, just 7% through the trial duration. However, it does not remain there until patient 360 (51% through the trial). Despite a similar first instance in the ±0.05*SD boundary for the reverse order (70 patient, 9% through the trial), the reverse order estimate remains in this boundary sooner than the original order; from patient 280 onwards (40% through the trial).



(a)

(b)

Figure C.12: Stability of the estimate in the CASPER MINUS trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.5 CASPER trial

Figure C.13 is similar to the CASPER MINUS trial, but takes longer for the current trend line to reach 1 (∼600 patients). The hypothesised effect and 90% optimistic limit lines again are consistently ≈1 throughout the trial duration. The 80% limit assumption is predominantly>0.98, with a small dip between patients 57-79, which still remains in the promising zone. The current trend line is highly variable at first, and then gradually increases from 0 to 1, being in the favourable zone consistently from patient 341 onwards. Table C.9 summarises the number of times the CP falls in each zone.

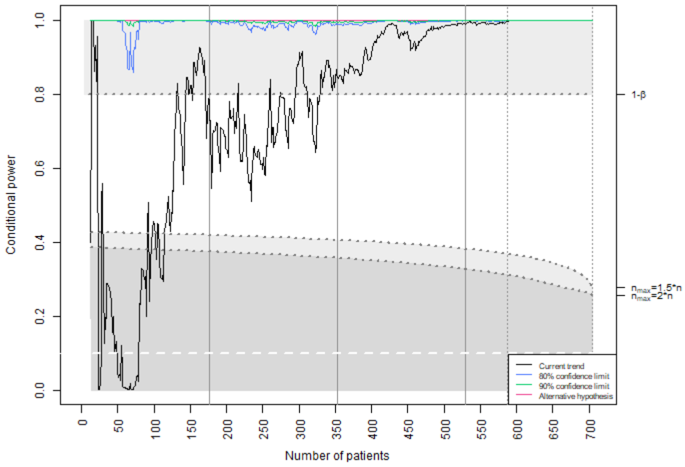


Figure C.13: Conditional power calculated after every patient in the CASPER trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	435	694	694	694
	Promising	177	0	0	0
	Unfavourable	48	0	0	0
	(Futility)	34	0	0	0
n_{max}=2	Favourable	435	694	694	694
	Promising	182	0	0	0
	Unfavourable	43	0	0	0
	(Futility)	34	0	0	0

Table C.9: Number of times CP values fall in each zone for the promising zone design for the CASPER trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table C.10 presents interim analysis decisions for the chosen time points. Promising zone and stepwise designs would only increase if the current trend assumption were used, and a 25% data available interim analysis (18% increase for either n_{max} value with the promising zone design, and 33% and 66% for n_{max} values of 1.5 and 2 respectively with the stepwise design. The combination test design would have decreased to 83% of the original sample size for any assumption at the 25% interim time point, and remained at the original 705 patients for the other time points.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 177	n_{max}=1058	0.731	0.419	Promising	835	587	0.466	940
	Recruited:587	n_{max}=1410	0.731	0.374	Promising	835	587	0.440	1175
50%	Available: 353	n_{max}=1058	0.845	0.406	Favourable	705	705	0.460	705
	Recruited: 705	n_{max}=1410	0.845	0.357	Favourable	705	705	0.430	705
75%	Available: 529	n_{max}=1058	0.990	0.382	Favourable	705	705	0.447	705
	Recruited: 705	n_{max}=1410	0.990	0.328	Favourable	705	705	0.410	705
HYPOTHESISED EFFECT									
25%	Available: 177	n_{max}=1058	1.000	0.419	Favourable	705	587	0.466	705
	Recruited:587	n_{max}=1410	1.000	0.374	Favourable	705	587	0.440	705
50%	Available: 353	n_{max}=1058	1.000	0.406	Favourable	705	705	0.460	705
	Recruited: 705	n_{max}=1410	1.000	0.357	Favourable	705	705	0.430	705
75%	Available: 529	n_{max}=1058	1.000	0.382	Favourable	705	705	0.447	705
	Recruited: 705	n_{max}=1410	1.000	0.328	Favourable	705	705	0.410	705
80% OPTIMISTIC LIMIT									
25%	Available: 177	n_{max}=1058	0.998	0.419	Favourable	705	587	0.466	705
	Recruited:587	n_{max}=1410	0.998	0.374	Favourable	705	587	0.440	705
50%	Available: 353	n_{max}=1058	0.989	0.406	Favourable	705	705	0.460	705
	Recruited: 705	n_{max}=1410	0.989	0.357	Favourable	705	705	0.430	705
75%	Available: 529	n_{max}=1058	0.999	0.382	Favourable	705	705	0.447	705
	Recruited: 705	n_{max}=1410	0.999	0.328	Favourable	705	705	0.410	705
90% OPTIMISTIC LIMIT									
25%	Available: 177	n_{max}=1058	1.000	0.419	Favourable	705	587	0.466	705
	Recruited:587	n_{max}=1410	1.000	0.374	Favourable	705	587	0.440	705
50%	Available: 353	n_{max}=1058	0.996	0.406	Favourable	705	705	0.460	705
	Recruited: 705	n_{max}=1410	0.996	0.357	Favourable	705	705	0.430	705
75%	Available: 529	n_{max}=1058	0.999	0.382	Favourable	705	705	0.447	705
	Recruited: 705	n_{max}=1410	0.999	0.328	Favourable	705	705	0.410	705

Table C.10: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

Figure C.14 shows n^* calculated after every patient. In all cases, the combination test design would be $\leq$705. Due to the longer time until primary outcome data is available, the combination test (green) reaches the original planned sample size much quicker than CASPER MINUS (by about patient 250 onwards). For either optimistic limit or hypothesised effect, no increase in sample size is seen. Both promising zone and stepwise designs see a large fluctuation of sample size if an interim analysis were to be carried out before 350 patients had data available, using the current trend assumption, with both designs reaching the maximum value of 2 times the original sample size n.

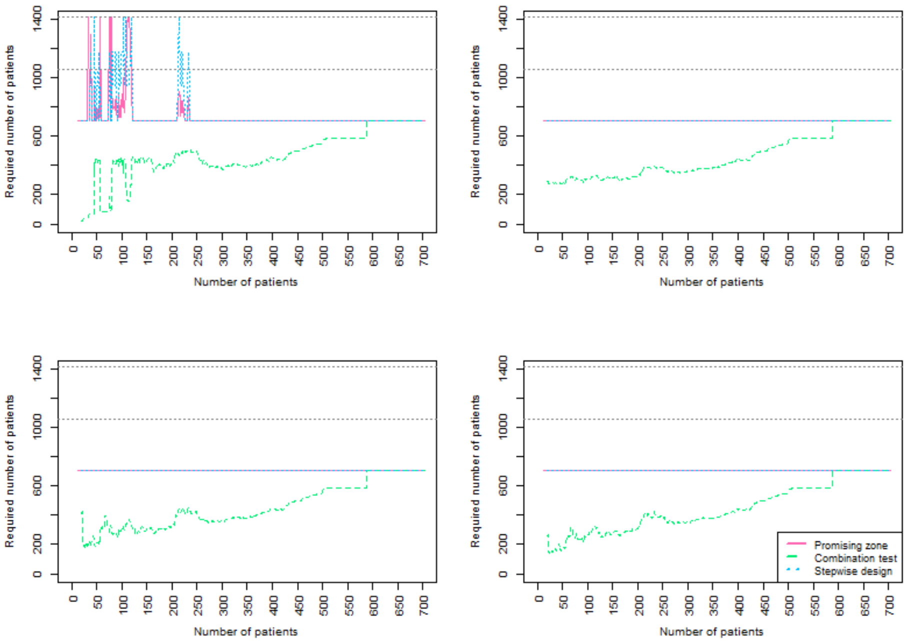
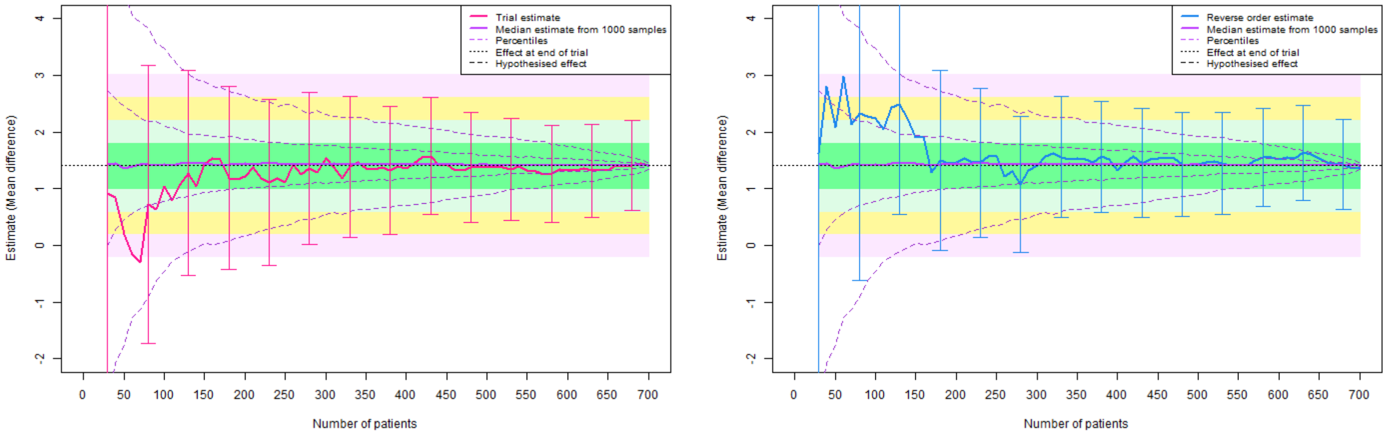


Figure C.14: Comparison of three SSR designs for the CASPER trial data

The sequential order estimate (calculated from patient 30 onwards) from the CASPER trial (Figure C.15)starts in the ±2*SE boundary at 30 patients but drops through two boundaries shortly after. However, at 120 patients the estimate is back in the 1*SE boundary, and remains there until the end, just 17% through the trial. Whilst the estimate fluctuates slightly between over and under estimating the true effect going forward, it remains stable from this point. A similar pattern is observed in the reverse order, starting with an over-estimate, but dropping to within ±1*SE by 170 patients (24% through the trial), where it then remains.



(a) *(b)*

Figure C.15: Stability of the estimate in the CASPER trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.6 CASPER PLUS trial

Figure C.16 shows CP values for the CASPER PLUS trial. The hypothesised effect line remains consistently high (≈1) throughout the trial. The other three lines start at zero, and gradually increase to 1, fluctuating between zones before this point (∼430 patients). The current trend line remains below the futility boundary (0.1) until patient 125, unfavourable until 176, varying between all three zones until patient 350, where it remains in the favourable zone. The optimistic limits rapidly alter between all zones for the first 120 patients. After this point, the 90% limit remains in the favourable zone, and the 80% limit dips twice into the promising zone. Table C.16 summarises the number of times each line falls into each zone.

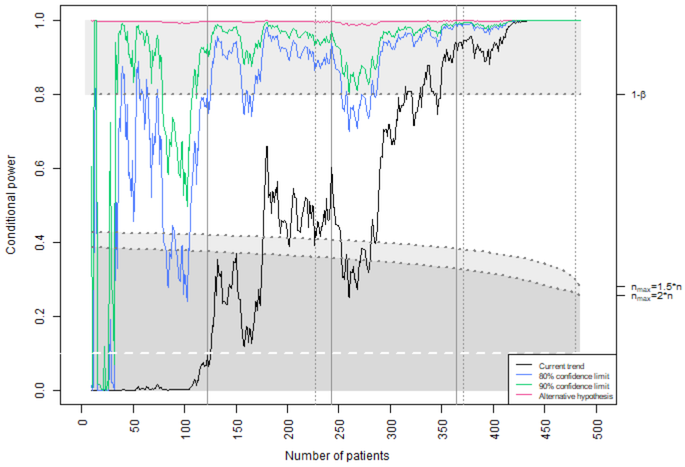


Figure C.16: Conditional power calculated after every patient in the CASPER PLUS trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	155	476	334	425
	Promising	119	0	97	34
	Unfavourable	86	0	25	4
	(Futility)	116	0	20	13
n_{max}=2	Favourable	155	476	334	425
	Promising	132	0	104	34
	Unfavourable	73	0	18	4
	(Futility)	116	0	20	13

Table C.11: Number of times CP values fall in each zone for the promising zone design for the CASPER PLUS trial. For a design where no futility boundary is considered, these values become unfavourable instead.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 121	n_{max}=728	0.093	0.419	Unfavourable	485	227	0.466	485
	Recruited:227	n_{max}=970	0.093	0.374	Unfavourable	485	227	0.440	485
50%	Available: 243	n_{max}=728	0.602	0.406	Promising	728	485	0.460	728
	Recruited: 371	n_{max}=970	0.602	0.357	Promising	763	485	0.430	970
75%	Available: 364	n_{max}=728	0.938	0.382	Favourable	485	480	0.447	485
	Recruited: 480	n_{max}=970	0.938	0.328	Favourable	485	480	0.410	485
HYPOTHESISED EFFECT									
25%	Available: 121	n_{max}=728	0.997	0.419	Favourable	485	326	0.466	485
	Recruited:227	n_{max}=970	0.997	0.374	Favourable	485	326	0.440	485
50%	Available: 243	n_{max}=728	0.998	0.406	Favourable	485	371	0.460	485
	Recruited: 371	n_{max}=970	0.998	0.357	Favourable	485	371	0.430	485
75%	Available: 364	n_{max}=728	0.999	0.382	Favourable	485	480	0.447	485
	Recruited: 480	n_{max}=970	0.999	0.328	Favourable	485	480	0.410	485
80% OPTIMISTIC LIMIT									
25%	Available: 121	n_{max}=728	0.812	0.419	Favourable	485	496	0.466	485
	Recruited:227	n_{max}=970	0.812	0.374	Favourable	485	496	0.440	485
50%	Available: 243	n_{max}=728	0.938	0.406	Favourable	485	441	0.460	485
	Recruited: 371	n_{max}=970	0.938	0.357	Favourable	485	441	0.430	485
75%	Available: 364	n_{max}=728	0.989	0.382	Favourable	485	480	0.447	485
	Recruited: 480	n_{max}=970	0.989	0.328	Favourable	485	480	0.410	485
90% OPTIMISTIC LIMIT									
25%	Available: 121	n_{max}=728	0.935	0.419	Favourable	485	425	0.466	485
	Recruited:227	n_{max}=970	0.935	0.374	Favourable	485	425	0.440	485
50%	Available: 243	n_{max}=728	0.971	0.406	Favourable	485	415	0.460	485
	Recruited: 371	n_{max}=970	0.971	0.357	Favourable	485	415	0.430	485
75%	Available: 364	n_{max}=728	0.994	0.382	Favourable	485	480	0.447	485
	Recruited: 480	n_{max}=970	0.994	0.328	Favourable	485	480	0.410	485

Table C.12: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

Table C.12 presents decisions for the three specified interim analyses. Promising zone and stepwise designs only increase sample size at 50% data available, using the current trend assumption. With n_{max}=1.5, both designs reach the maximum increase of 50%, and with n_{max}=2, promising zone increases by 78%, compared to the maximum of 100% increase for the stepwise design. The combination test would have increased in just one scenario (80% optimistic limit, 25% data available) by just 11 patients (2%). The largest decrease occurs at 25% data available under the current trend, to 47% of the original sample size planned.

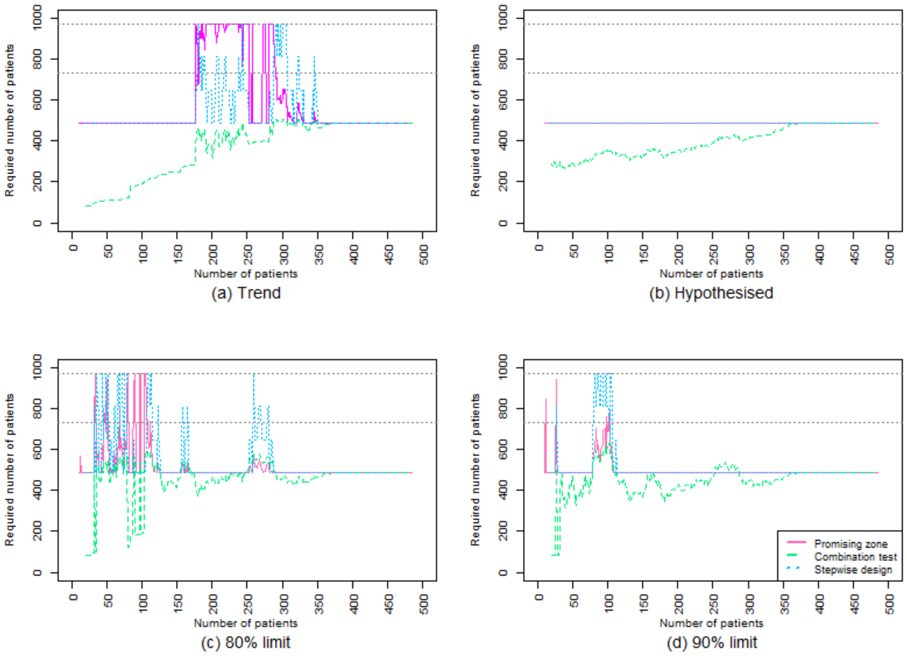
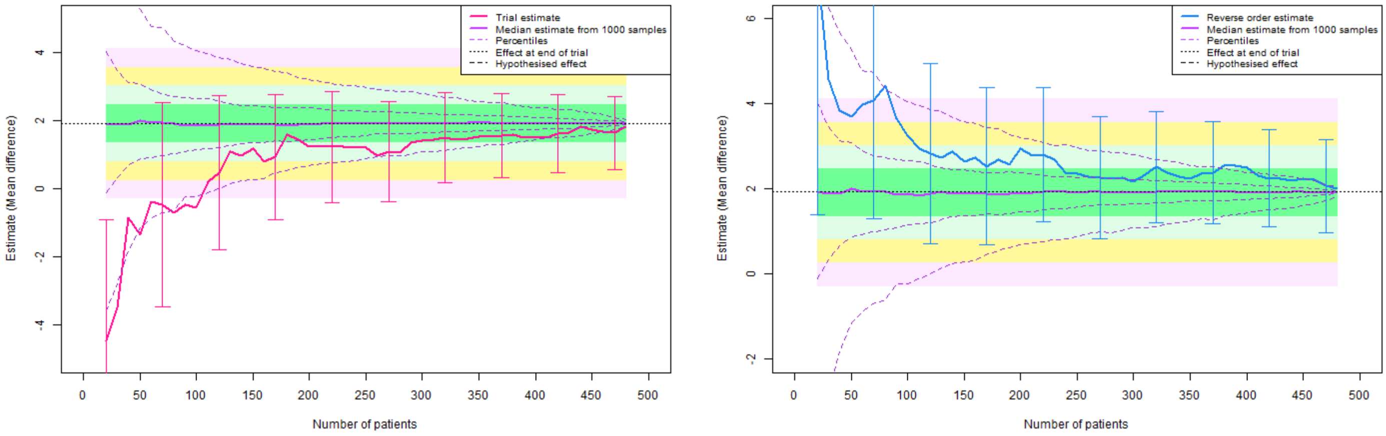


Figure C.17: Comparison of three SSR designs for the CASPER PLUS trial data

Figure C.17 compares n^* under each assumption. No increase is seen at any point using the hypothesised effect assumption. The combination test design would decrease sample size at any interim point before 370 patients, after which the sample size would remain at the originally planned sample size. Large fluctuations in sample size can be seen in the other three assumptions: predominantly early on in the trial for the optimistic limits, and mainly in the middle part of the trial for the current trend. An additional spike in sample size can be seen using the 80% limit between 250 and 300 patients, with the largest rise in sample size being seen for the stepwise design (blue). Figure C.18 shows the sequential and reverse order estimates calculated from patient 20 onwards, after every 10 patients. The original order estimate here starts far below the original analysis treatment effect. with even the upper boundary of the 95% CI not reaching the largest boundary investigated, ±4*SE. However, from this point forward it increases, reaching the ±4*SE, ±3*SE ,±2*SE, and ±1*SE boundaries by 110, 120, 130 and 180 patients respectively. From 300 patients onwards (62% through the trial), the estimate is within ±1*SE of the end estimate, and remains there. The reverse order however, takes longer to first reach the ±1*SE boundary (240 patients), and to remain in this boundary (410 patients onwards, 85% through the trial).



(a) *(b)*

Figure C.18: Stability of the estimate in the CASPER PLUS trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.7 Epilepsy trial

Figure C.19 shows CP lines for the four treatment effect assumptions in the Epilepsy trial. Whilst the trial data is re-imagined at two additional time points (1 day and 1 year), as well as the original 19 weeks, CP values are the same for all time points, as there is no change in sequential order. However, the combination test design takes into account the number of patients already recruited, and therefore Figure C.20 showing n^* also includes two additional lines: one for 1 day data (orange) and 1 year (purple).

CP starts very low low (close to zero) and remains low throughout the trial duration. Optimistic limit and hypothesised effect lines start close to 1 and gradually decrease to zero; the optimistic limits having a sharper decrease (reaching zero by 50 patients), compared to a more gradual decline (by patient 75). This corresponds to the n^* plot (Figure C.20), with the promising zone only increasing using the optimistic limits or hypothesised effect cases, matching up with the time spent in the promising zone for each. The current trend assumption does not result in an increase at all, with a decrease being seen in all three combination test designs prior to around 80 patients (2/3 through the trial). More increases in sample size can be seen in the hypothesised effect assumption, predominantly in the combination test design case, before establishing a decrease instead later on in the trial. Optimistic values show a similar situation, but increases occur in a much smaller percentage of values of n_1 than in the hypothesised effect.

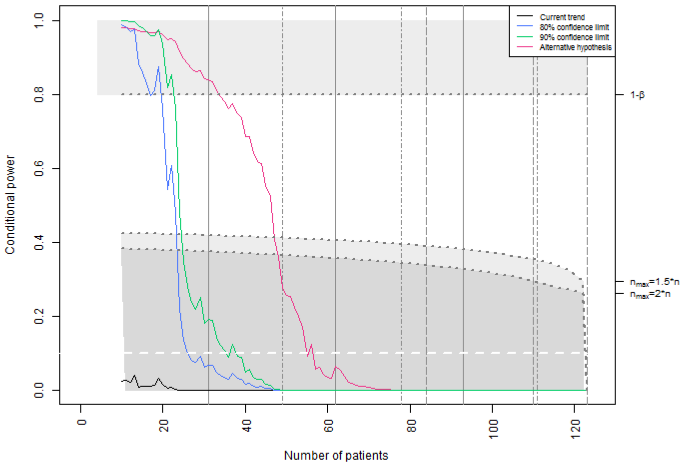
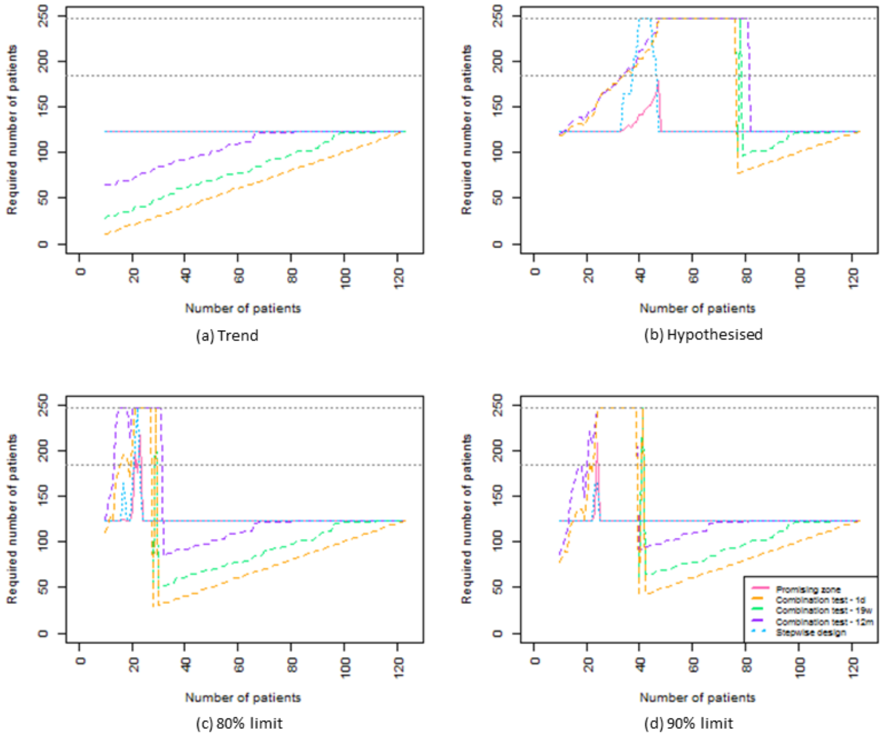


Figure C.19: Conditional power calculated after every patient in the Epilepsy trial



(a) Trend

(b) Hypothesised

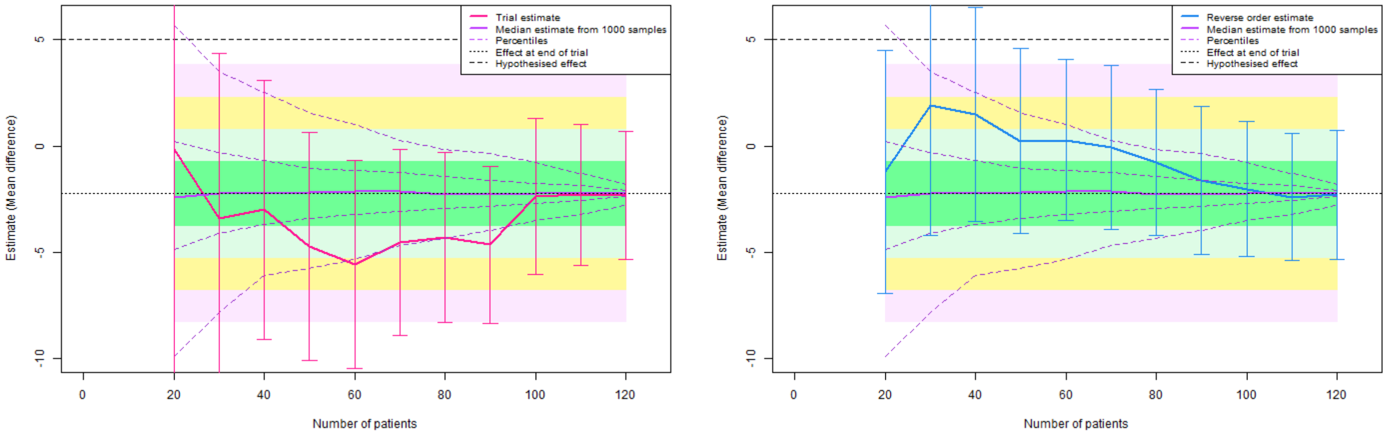
(c) 80% limit

(d) 90% limit

Figure C.20: Comparison of three SSR designs for the Epilepsy trial data

In all cases, promising zone and stepwise designs remain at the originally planned 123 patients, with CP values <0.2 for all cases except at 25% data available using the hypothesised effect, with a CP value of 0.84. All three combination test designs see a maximum increase of doubling the sample size to 246 patients. One year results sees the most number of increases in sample size, which corresponds to having a higher number recruited at the time of the interim analysis. The largest sample size increase here is 32%, compared to 60% decrease for the 19 week endpoint, or 75% decrease in the 1 day endpoint.

The original order estimate first reaches the ±1*SE boundary at 30 patients (24% through the trial) (Figure C.21). However, the estimate then drops to the ±2*SD, and even ±3*SD boundary, until returning to within 1*SE from 100 patients (81% through the trial) onwards. Similarly, the reverse order always falls within the ±3*SE boundaries, remaining within the ±1*SE boundary slightly earlier (80 patient, 65% through the trial). This suggests that later patients have outcomes closer to the hypothesised value of a 5 unit increase.



(a) *(b)*

Figure C.21: Stability of the estimate in the Epilepsy trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.8 FLU A/H1N1 trial

CP (Figure C.22) under either optimistic limit or the hypothesised treatment effect remain at 1 throughout the duration of the trial. On the other hand, the current trend assumption results in a highly fluctuating CP line for the first 25% data available, reaching both very high (close to 1) and very low (close to zero) values, before eventually reaching the favourable zone (>0.92) and mostly remaining there from patient 602 onwards. These rapid changes in CP can also be seen on the corresponding n^* graphs (Figure C.23) for the stepwise and promising zone designs under the current trend assumption, also varying between the original sample size and n_{max}. No sample size can be seen beyond n_1=400 for the stepwise design, and n_1=700 for the promising zone design. The combination test design remains consistent at the original sample size for all four assumptions. Promising zone and stepwise design also see no increases in sample size using the optimistic limits or the hypothesised effect assumptions. It should be noted for all three Flu vaccine trials, that all patients have been recruited by the time any endpoint is collected due to the 12 day recruitment phase (shortest outcome is 28 days). For this reason, the smallest n^* can be using the combined objective function, is the number recruited, which is the full original sample size n.

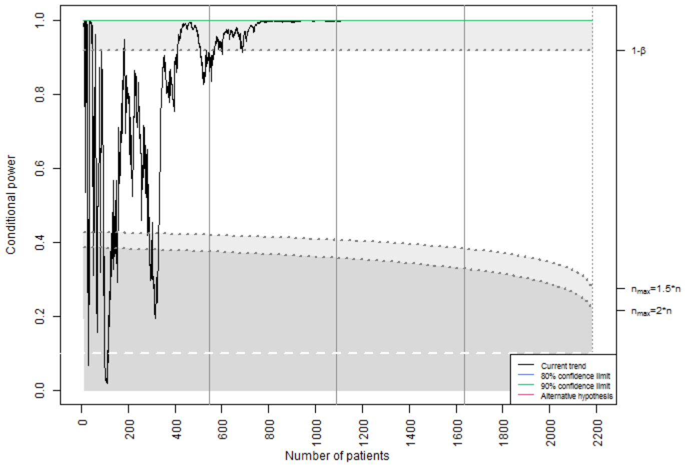


Figure C.22: Conditional power calculated after every patient in the Flu A/H1N1 trial

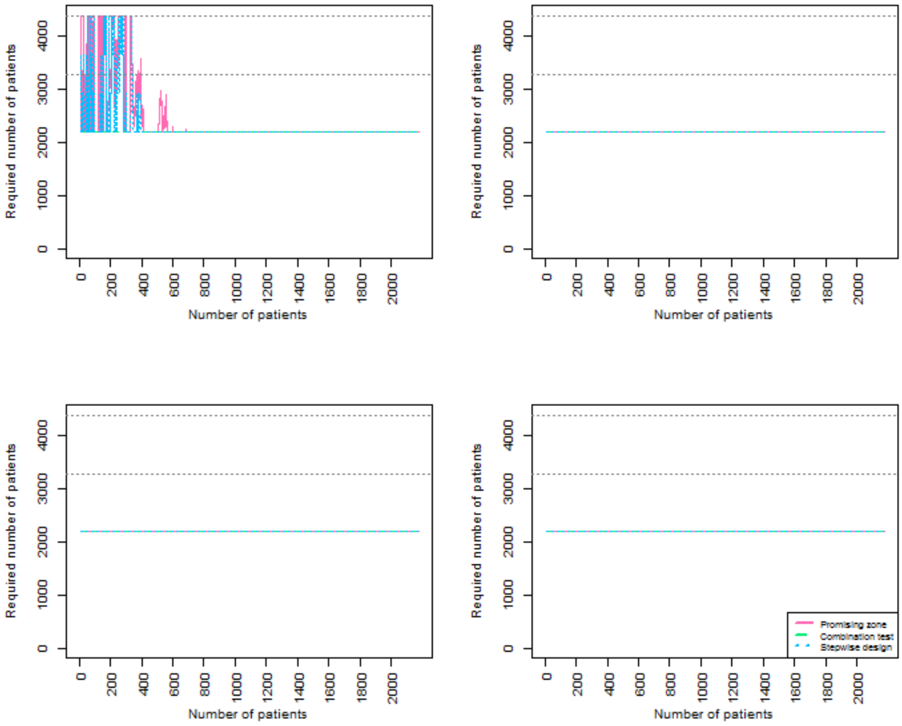
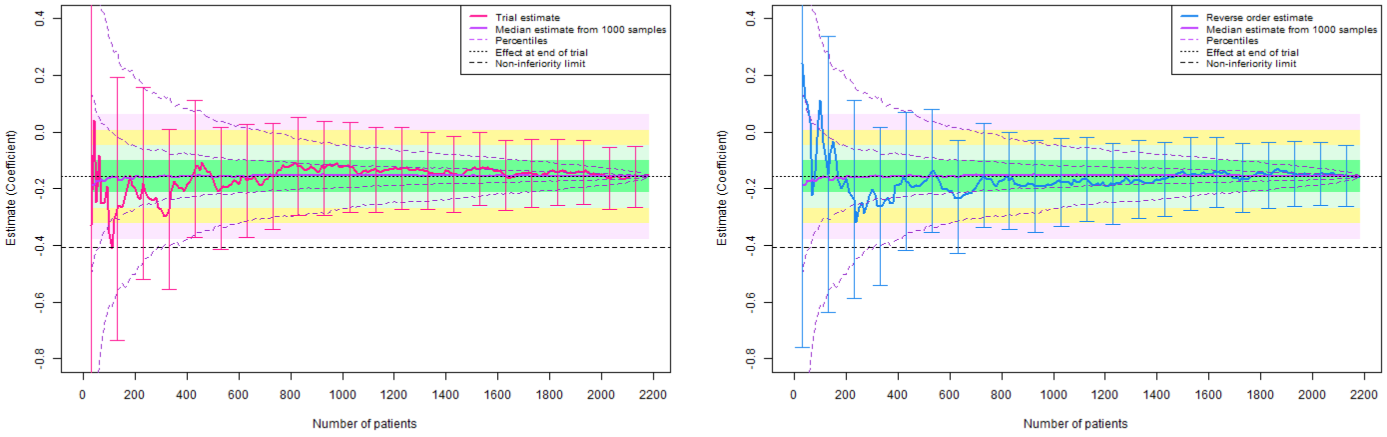


Figure C.23: Comparison of three SSR designs for the Flu A/H1N1 trial data

The stepwise and combination test designs would have remained constant at the original 2182 patients at any interim point investigated here. The promising zone only sees one increase (by 15%), observed at 25% data available using the assumption of the current trend. Figure C.24 shows the sequential and reverse order estimates through the trial duration for the first strain in the flu vaccine trial (A/H1N1). Other than one brief drop beyond the ±4*SE zone at 110 patients, the estimate mainly stays within ±3*SE, and always within ±4*SE of the actual treatment effect from the original analysis. The estimate lies in the darker green boundary from 340 patients (16% through the trial). The estimate from the reverse order also varies greatest near the beginning of the trial, but remains within the ±4*SE boundary by patient 110 (5% through the trial). The reverse order estimate remains within the ±1SE boundary from patient 690 (32%).



(a) *(b)*

Figure C.24: Stability of the estimate in the Flu A/H1N1 vaccine trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.9 FLU A/H3N2 trial

Figure C.25 shows CP calculated after every patient in the Flu vaccine A/H3N2 trial. The hypothesised effect line remains at 1 at all n_1 values. Current trend starts at zero, and gradually increases despite a couple of spikes in CP, reaching the favourable zone by patient 654, before increasing to 1. Both optimistic limits have very rapid fluctuations in CP, changing from ≈1 to ≈0 within <100 patients. Both lines settle at 1 by patient 400, and no more CP spikes can be seen after this point. Looking at the corresponding n^* plots (Figure C.26), no increase can be seen in any design using the hypothesised effect. Some early spikes occur in the other three lines, but settle down before patient 400 (80% limit), 300 (90% limit) or 700 (current trend) for promising zone and stepwise designs. Again, no decrease is seen in the combination test due to the short recruitment phase.

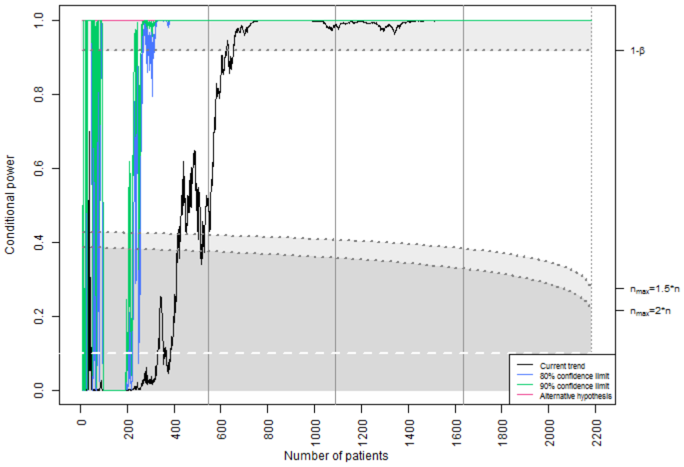


Figure C.25: Conditional power calculated after every patient in the Flu A/H3N2 trial

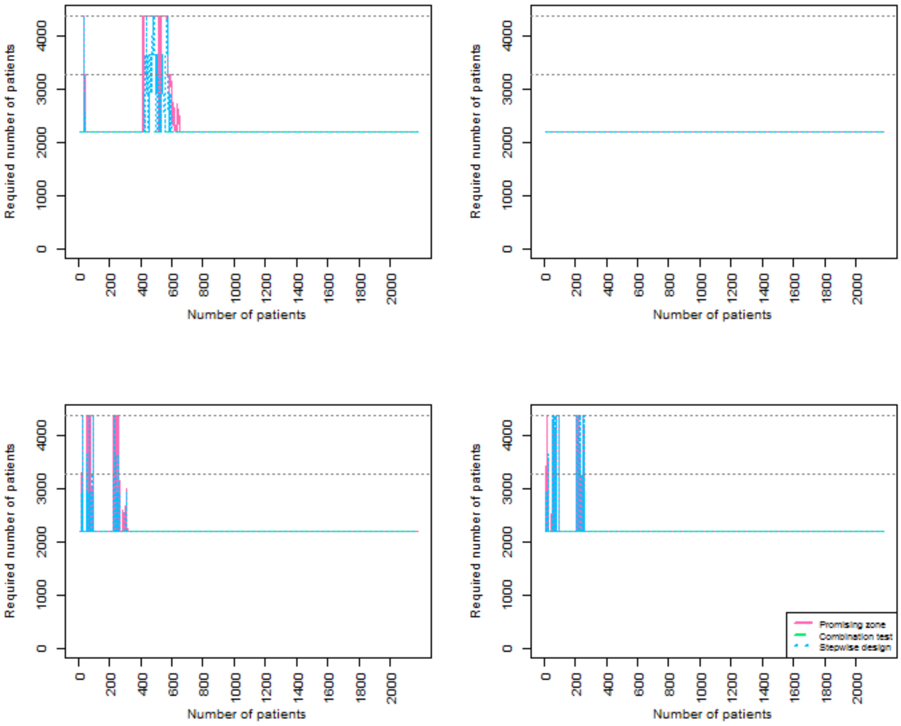
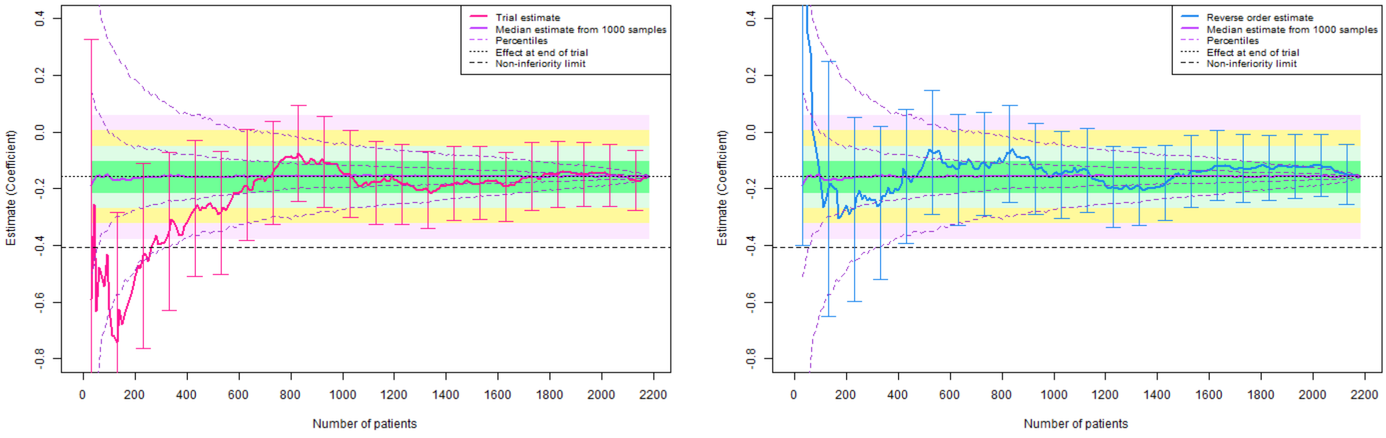


Figure C.26: Comparison of three SSR designs for the Flu A/H3N2 trial data

Similarly to the Flu A/H1N1 trial, only one instance of sample size increase is observed when limited to only the three specified interim analyses. Again, this is seen for the promising zone design at 25% data available, reaching the full maximum all owed sample size (50% increase for n_{max}=1.5, and 100% for n_{max}=2).

Strain A/H3N2 has the lowest treatment estimate out of the three flu vaccine trial continuous outcomes, falling below even the non-inferiority limit before 270 patients (except on brief spike at patient 40). From 330 patients, the estimate lies within 4*SEs, and from 560 patients (26%), lies within 2*SEs of the final treatment estimate. The inner boundary of ±1*SE is not maintained until 1350 patients onwards (62% through the trial). The reverse order first enters the ±1*SE boundary much earlier than the original order (90 patients compared to 600). Additionally, the reverse order estimate remains in this boundary from 910 patients (42% through the trial, compared to 62% in the original sequential order).



(a) *(b)*

Figure C.27: Stability of the estimate in the Flu A/H3N2 vaccine trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.10 FLU B1 trial

CP is much quicker to converge to 1 for the strain B1 in the flu vaccine trial, with lines reaching 1 by patient 162 for the current trend, 38 for the 80% interval, and 16 for the 90% interval. The hypothesised effect again remains at 1 for the entirety of the trial. Unsurprisingly, this corresponds to a very early peak in new total sample size n^* (Figure C.29) for the trend and optimistic limit assumptions, but flatten out to maintain the original sample size at all other values of n_1. Additionally, no sample size increase is observed at any of the three interim time points, for any design, under any assumption.

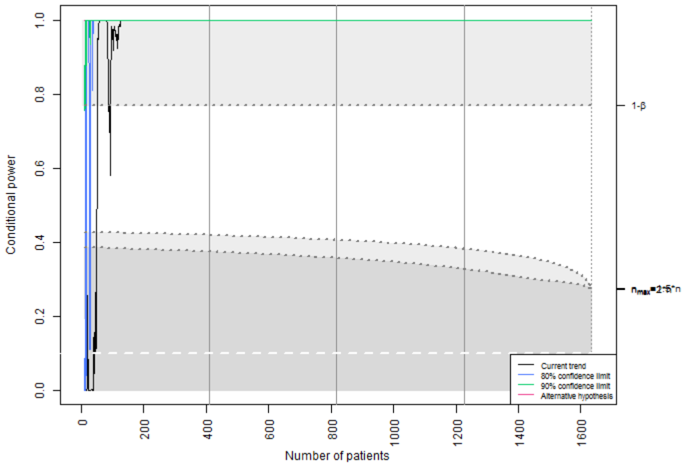


Figure C.28: Conditional power calculated after every patient in the Flu B1 trial

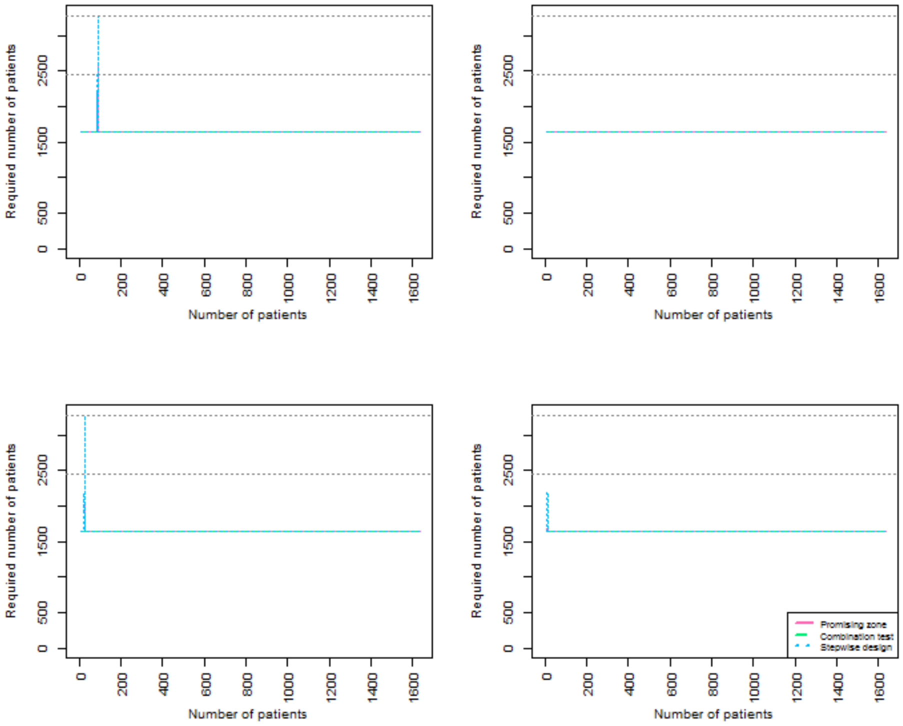
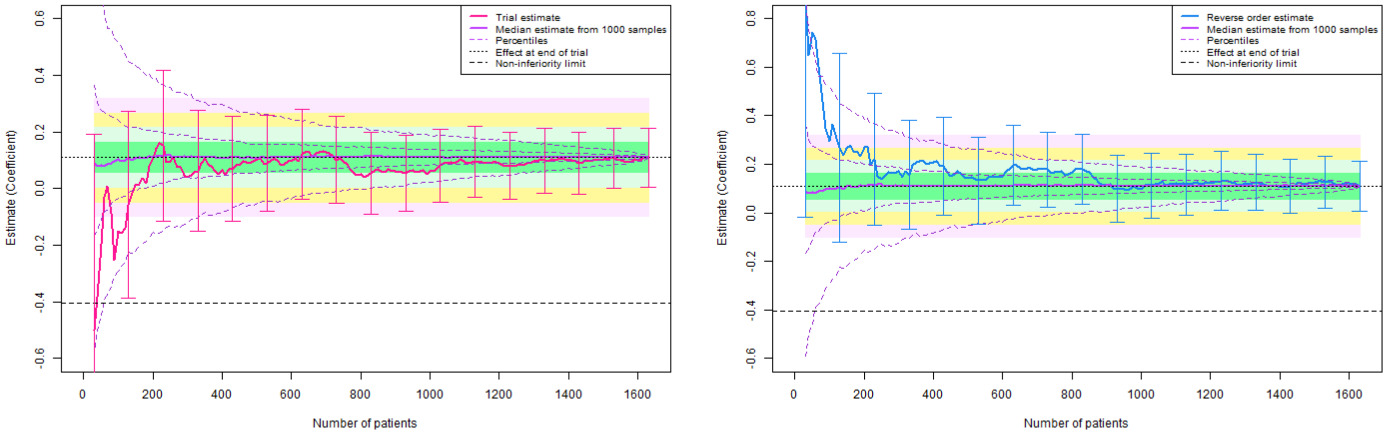


Figure C.29: Comparison of three SSR designs for the Flu BN1 trial data

Whilst the estimate starts below the lower ±4*SE boundary, (and below the non-inferiority limit), it quickly increases, and lies within ±3*SE boundary from patient 140 (150 for the

±2*SD bound). However, the reverse order estimate starts well above the highest boundary (+4*SE) and is slower to reach the ±1*SE limits. Both limits are slow to reach the inner limits of ±1*SE and remain there; 1000 patients (61%) for the original order, and 860 (53%) for the reverse order.



(a) *(b)*

Figure C.30: Stability of the estimate in the Flu B1 vaccine trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.11 IMPROVE trial

Figure C.31 shows CP calculated after every patient assuming four different treatment effects. The hypothesised treatment effect assumption line starts at 1 and gradually decreases, leaving the favourable zone by patient 167, the promising zone by 257 (n_{max}=1.5*n) or 269 (n_{max}=2*n). This line (red) has the highest CP values at any value of n_1 (x-axis). The current trend line starts very low, and remains always below the 10% futility bound, except for one instance at 20 patients, which reaches the promising zone if n_{max}=2 (otherwise is classed as the unfavourable zone). The optimistic limit assumptions result in large fluctuations early in the trial, but settles to almost zero by the second interim time point (50% data available). Both limits have a spike around 110-200 patients, but the 80% line stays within the unfavourable zone for wither value of n_{max}. The 90% optimistic limit line lies in the promising zone a total of 35 times between patient 125 and 177 for n_{max}=1.5, and 45 times in the same interval for n_{max}=2. Table C.3 summarises the number of times each line falls into the four zones (with futility zone being included with unfavourable if no stopping boundary is being used).

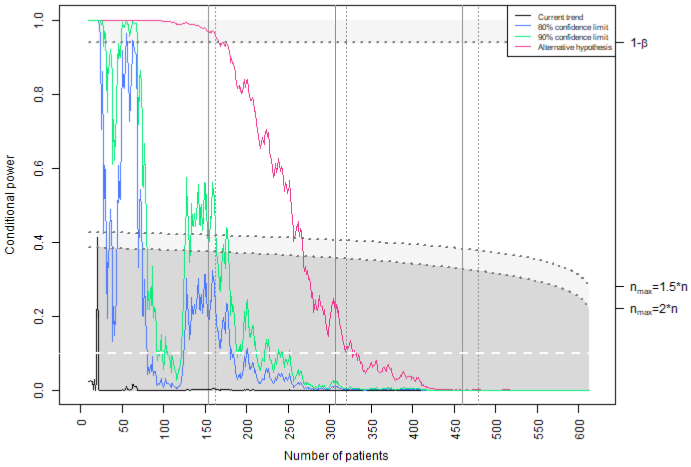


Figure C.31: Conditional power calculated after every patient in the IMPROVE trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	0	159	18	40
	Promising	0	97	32	65
	Unfavourable	1	63	79	85
	(Futility)	603	285	475	414
n_{max}=2	Favourable	0	159	18	40
	Promising	1	100	35	74
	Unfavourable	0	60	76	76
	(Futility)	603	285	475	414

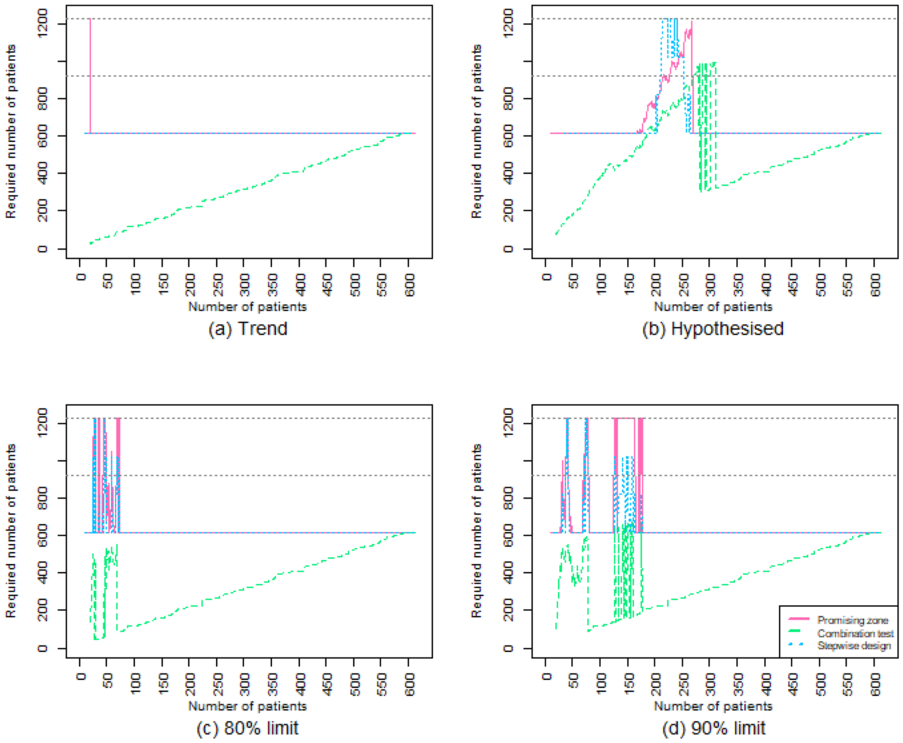
Table C.13: Number of times CP values fall in each zone for the promising zone design for the IMPROVE trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table C.14 presents three specified interim analysis time points (25, 50 and 75% patients with data available), comparing decisions made and new total sample size from the three SSR designs. No sample size increase would have been seen for any time point using the stepwise design, and all values of $n^* = n = 613$, the original planned sample size. One increase (n^*=1226) can be seen using the promising zone design, assuming a 90% optimistic limit and maximum increase of twice the original sample size and 25% data available interim time point. The new sample size from the combination test ranges from a decrease in sample size of 74% (seen at 25% data available assuming current trend or either optimistic limit), to an increase in sample size of 62%, seen at 50% data available under the hypothesised effect assumption and a n_{max}=2, following a CP value of 21%.

Figure C.32 compares n^* values for all n_1 values, for three SSR designs. Other than one sharp peak of n^* for the promising zone at the 20 patient point discussed previously, no increase in sample size can be seen under the current trend assumption. The combination test would have decreased the sample size for every n_1 value using the current trend or optimistic 80% interval, and almost every n_1 value under the other two assumptions. The largest increase in sample size from the combination test design is from the hypothesised treatment effect assumption, from $\approx$ 200-320 patients. The promising zone and stepwise designs also see an increase in this range under the same assumption, but return to no increase in sample size sooner.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 154	n_{max}=920	0.001	0.419	Unfavourable	613	162	0.466	613
	Recruited:162	n_{max}=1226	0.001	0.374	Unfavourable	613	162	0.440	613
50%	Available: 307	n_{max}=920	0.000	0.406	Unfavourable	613	320	0.460	613
	Recruited: 320	n_{max}=1226	0.000	0.357	Unfavourable	613	320	0.430	613
75%	Available: 460	n_{max}=920	0.000	0.382	Unfavourable	613	479	0.446	613
	Recruited: 479	n_{max}=1226	0.000	0.328	Unfavourable	613	479	0.410	613
HYPOTHESISED EFFECT									
25%	Available: 154	n_{max}=920	0.970	0.419	Favourable	613	494	0.466	613
	Recruited:162	n_{max}=1226	0.970	0.374	Favourable	613	494	0.440	613
50%	Available: 307	n_{max}=920	0.211	0.406	Unfavourable	613	320	0.460	613
	Recruited: 320	n_{max}=1226	0.211	0.357	Unfavourable	613	991	0.430	613
75%	Available: 460	n_{max}=920	0.000	0.382	Unfavourable	613	479	0.446	613
	Recruited: 479	n_{max}=1226	0.000	0.328	Unfavourable	613	479	0.410	613
80% OPTIMISTIC LIMIT									
25%	Available: 154	n_{max}=920	0.195	0.419	Unfavourable	613	162	0.466	613
	Recruited:162	n_{max}=1226	0.195	0.374	Unfavourable	613	162	0.440	613
50%	Available: 307	n_{max}=920	0.007	0.406	Unfavourable	613	320	0.460	613
	Recruited: 320	n_{max}=1226	0.007	0.357	Unfavourable	613	320	0.430	613
75%	Available: 460	n_{max}=920	0.000	0.382	Unfavourable	613	479	0.446	613
	Recruited: 479	n_{max}=1226	0.000	0.328	Unfavourable	613	479	0.410	613
90% OPTIMISTIC LIMIT									
25%	Available: 154	n_{max}=920	0.409	0.419	Unfavourable	613	162	0.466	613
	Recruited:162	n_{max}=1226	0.409	0.374	Promising	1226	162	0.440	613
50%	Available: 307	n_{max}=920	0.017	0.406	Unfavourable	613	320	0.460	613
	Recruited: 320	n_{max}=1226	0.017	0.357	Unfavourable	613	320	0.430	613
75%	Available: 460	n_{max}=920	0.000	0.382	Unfavourable	613	479	0.446	613
	Recruited: 479	n_{max}=1226	0.000	0.328	Unfavourable	613	479	0.410	613

Table C.14: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.



(a) Trend

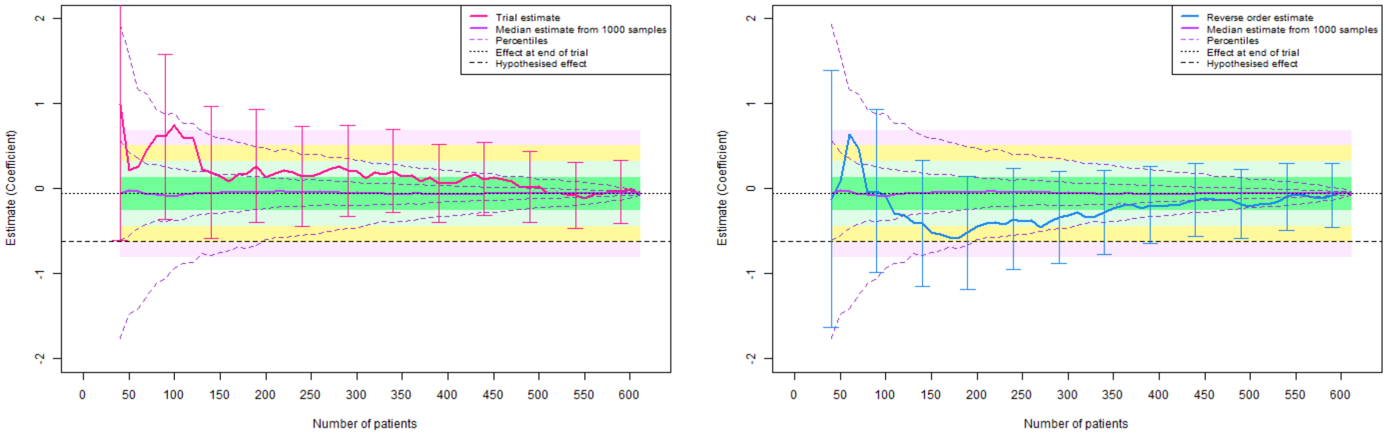
(b) Hypothesised

(c) 80% limit

(d) 90% limit

Figure C.32: Comparison of three SSR designs for the IMPROVE trial data

Figure C.33 shows the estimate calculated after every 10 patients in the original sequential order and reverse order of the trial dataset from patient 40 onwards, with 4 investigated boundaries for stability definition. The reverse order estimate always lies within the ±4*SEs of the assumed true treatment effect (black dotted line). However, the original order estimate does not, until 110 patients (18% through the trial). The original order estimate first reaches the ±1*SE boundaries later than the reverse order (160, 26% through the trial compared to 40, 7%). However, both then subsequently leave this boundary, and do not remain there until patient 350 (57%) for the reverse order estimate, compared with 460 (75%) for the original order estimate. Note that while the estimate remains above the hypothesised effect line, the model coefficients are on a log scale, and therefore a coefficient of 0 implies a OR of 1, meaning no treatment difference.



(a)

(b)

Figure C.33: Stability of the estimate in the IMPROVE trial. A comparison of sequential order (a), reverse order (b), and the median of 1000 random orders (a) and (b)

C.12 Corn plasters trial

Figure C.34 shows CP for the four future treatment effect assumptions, and Table C.15 summarises the time each CP line spends in each zone. The hypothesised treatment effect, and both optimistic limit lines start and end at 1, varying between favourable and promising zones in between. The hypothesised effect line starts smoother than the confidence limit assumption lines, but eventually follows the same pattern of spikes by 80 patients. On the other hand, the current trend line starts low (close to zero) and ends close to 1. However, the line fluctuates greatly in CP values, shifting between zones multiple times throughout the trial duration. This corresponds to the plot of n^* values under the current trend (Figure C.35, where n^* fluctuates a great deal in all three designs, throughout the trial. The combination test results in a decrease in sample size early on in the trial under any of the four assumptions, and continues this pattern in the optimstic limit and hypothesised effect assumptions. The stepwise design sees the highest n^* values, reaching n_{max}=1.5 in all cases, and n_{max}=2 in all but the 90% limit assumption. It should be noted that under the current assumption, the trial would have stopped for futility 22 times before patient 40, regardless of SSR design used, as CP values fall below the 10% boundary (if using).

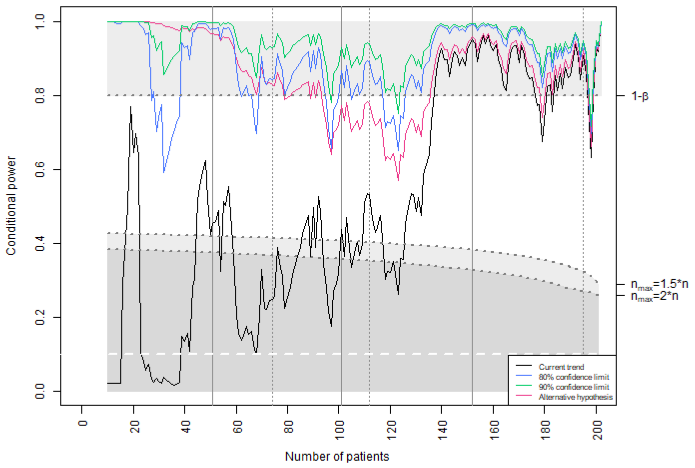


Figure C.34: Conditional power calculated after every patient in the Corn plasters trial

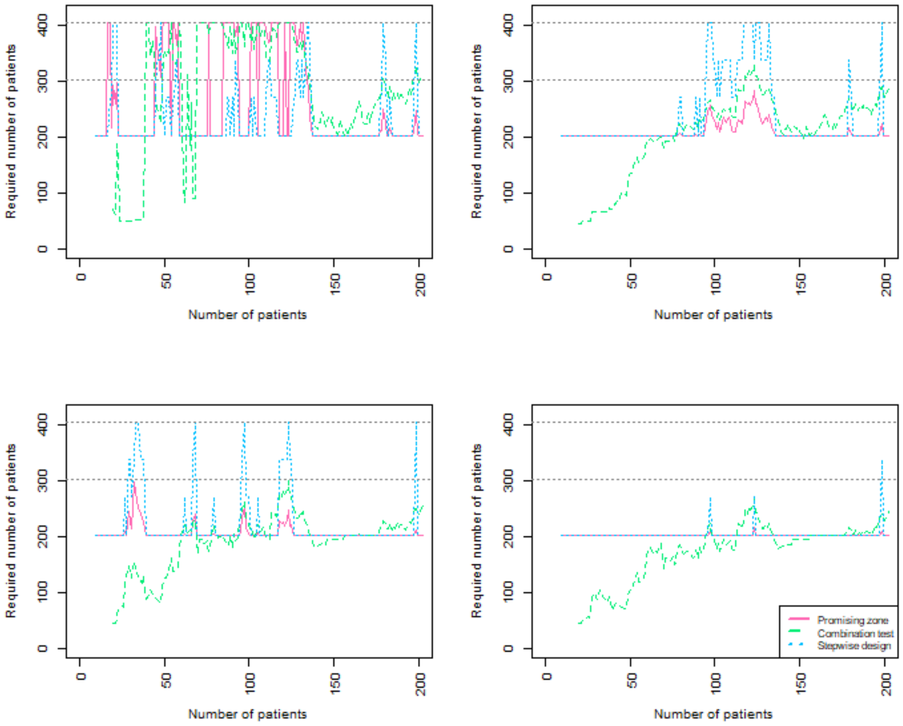


Figure C.35: Comparison of three SSR designs for the Corn plasters trial data

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	58	143	162	190
	Promising	54	50	31	3
	Unfavourable	59	0	0	0
	(Futility)	22	0	0	0
n_{max}=2	Favourable	58	143	162	190
	Promising	67	50	31	3
	Unfavourable	46	0	0	0
	(Futility)	22	0	0	0

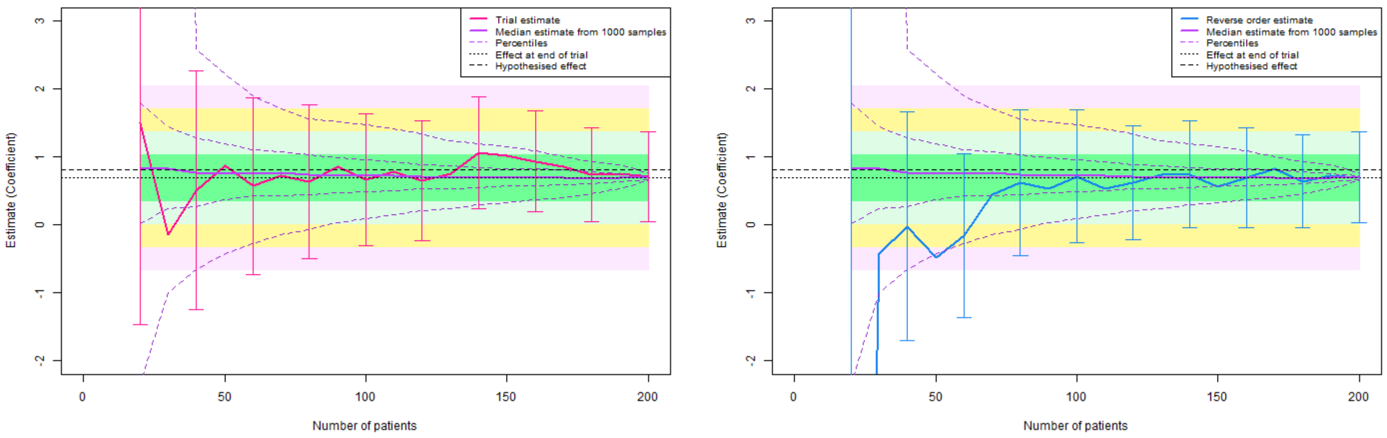
Table C.15: Number of times CP values fall in each zone for the promising zone design for the Corn plasters trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table C.16 gives full details of the three interim time points investigated, for the three designs under comparison. Under the current trend, the promising zone would have increased to the full n_{max} value for both 25% and 50% data available time points. Additionally, a small increase of 5% can be seen under the hypothesised effect assumption at the 50% time point for the promising zone design. The stepwise design sees three increases in sample size, all at the 50% data available time point: once under the current trend with n_{max}=2 (34% increase), and both values of n_{max} under the hypothesised effect (17% and 34% increases respectively). The combination test only sees increases under the current trend, from just 1 patient (at 75% data available), to an increase of 84%. The smallest observed decrease in sample size is 52% of the original n.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 51	n_{max}=303	0.454	0.419	Promising	303	303	0.466	202
	Recruited:74	n_{max}=404	0.454	0.374	Promising	404	344	0.440	202
50%	Available: 101	n_{max}=303	0.436	0.406	Promising	303	303	0.460	202
	Recruited: 112	n_{max}=404	0.436	0.357	Promising	404	372	0.429	270
75%	Available: 152	n_{max}=303	0.953	0.382	Favourable	202	203	0.447	202
	Recruited: 195	n_{max}=404	0.953	0.327	Favourable	202	203	0.410	202
HYPOTHESISED EFFECT									
25%	Available: 51	n_{max}=303	0.968	0.419	Favourable	202	136	0.466	202
	Recruited:74	n_{max}=404	0.968	0.374	Favourable	202	136	0.440	202
50%	Available: 101	n_{max}=303	0.772	0.406	Promising	213	235	0.460	236
	Recruited: 112	n_{max}=404	0.772	0.357	Promising	213	235	0.429	270
75%	Available: 152	n_{max}=303	0.960	0.382	Favourable	202	198	0.447	202
	Recruited: 195	n_{max}=404	0.960	0.327	Favourable	202	198	0.410	202
80% OPTIMISTIC LIMIT									
25%	Available: 51	n_{max}=303	0.982	0.419	Favourable	202	126	0.466	202
	Recruited:74	n_{max}=404	0.982	0.374	Favourable	202	126	0.440	202
50%	Available: 101	n_{max}=303	0.869	0.406	Favourable	202	202	0.460	202
	Recruited: 112	n_{max}=404	0.869	0.357	Favourable	202	202	0.429	202
75%	Available: 152	n_{max}=303	0.992	0.382	Favourable	202	195	0.447	202
	Recruited: 195	n_{max}=404	0.992	0.327	Favourable	202	195	0.410	202
90% OPTIMISTIC LIMIT									
25%	Available: 51	n_{max}=303	0.997	0.419	Favourable	202	105	0.466	202
	Recruited:74	n_{max}=404	0.997	0.374	Favourable	202	105	0.440	202
50%	Available: 101	n_{max}=303	0.931	0.406	Favourable	202	179	0.460	202
	Recruited: 112	n_{max}=404	0.931	0.357	Favourable	202	179	0.429	202
75%	Available: 152	n_{max}=303	0.996	0.382	Favourable	202	195	0.447	202
	Recruited: 195	n_{max}=404	0.996	0.327	Favourable	202	195	0.410	202

Table C.16: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

The original order treatment estimate (Figure C.36) starts by moving from the upper to lower 3*SE boundaries, before entering the ±1*SE boundary, where it stays until straying just outside at 140 patients. Therefore, the estimate does not strictly remain in this boundary until 150 patients, 74% through the trial duration. The reverse estimate however (Figure C.36b), does not provide a valid estimate until 30 patients in, where it lies in the red lower boundary (-4*SE). From patient 70 onwards, 35% through the trial, the estimate remains within the ±1*SE boundaries.



(a) *(b)*

Figure C.36: Stability of the estimate in the Corn plasters trial. A comparison of sequential order (a), reverse order (b), and the median of 1000 random orders (a) and (b)

C.13 AMAZE trial

Figure C.37 shows CP values using the AMAZE trial data under four treatment effect assumptions. A similar pattern to that observed in the Corn plasters trial can be seen here, with the current trend gradually increasing from 0 to 1, with some small fluctuations in between, and the remaining three lines starting and ending at 1, dropping in CP in the middle portion of the trial. However, here peaks from the optimistic limits momentarily reach even the unfavourable zone (only in n_{max} =1.5 for the 90% limit). The current trend assumption falls below the futility bound 126 times (Table C.17), and therefore would have stopped the trial if an interim analysis had taken place at any of these points. After patient 270, CP values lie in the favourable zone for all assumptions of future treatment effect.

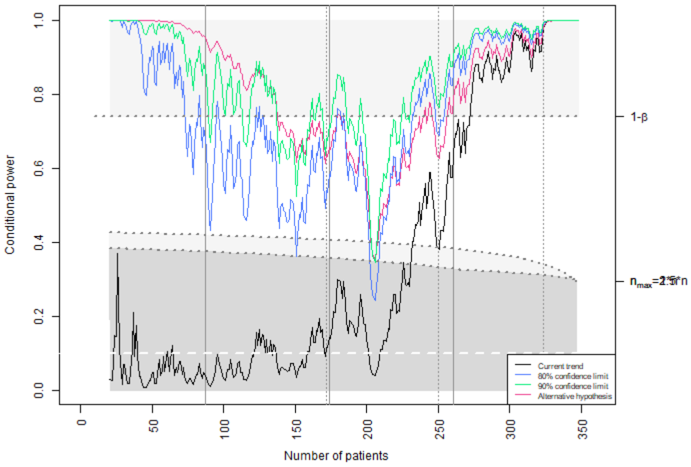


Figure C.37: Conditional power calculated after every patient in the AMAZE trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	76	217	175	251
	Promising	39	108	145	74
	Unfavourable	88	4	9	4
	(Futility)	126	0	0	0
n_{max}=2	Favourable	76	217	175	251
	Promising	42	112	147	77
	Unfavourable	85	0	7	1
	(Futility)	126	0	0	0

Table C.17: Number of times CP values fall in each zone for the promising zone design for the AMAZE trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table C.18 describes the three chosen interim time points using each SSR design, and decisions on new total sample size n^*. The promising zone design n^* increases range from just 3% increase (90% limit, 50% data available), to 24% (80% limit, 50% data available). The combination test only sees decreases in sample size under the current trend (25 and 50% data available) and the hypothesised effect (25% data available), with decreases ranging from 49% to 93% of the original planned sample size. The maximum increase observed is 47%, under the 80% limit at 50% data available. The stepwise design sees the greatest increase in sample size, reaching the maximum sample size allowed at 75% data available under the current trend, 50% under the hypothesised effect, and 25% under the 80% limit. Extending the investigation of n^* for every n_1 patients (Figure C.38) sees the largest region of increases under the optimistic limits, and a small region of moderate increase under the current trend for the combination test design. Despite a small region of sample size increase under the current trend for promising zone and stepwise designs, the increases are much sharper than the combination test, even reaching n_{max}=2. Promising zone sees smaller increases in sample size than both stepwise and combination test designs under the hypothesised treatment effect.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 88	n_{max}=525	0.040	0.419	Unfavourable	348	172	0.466	348
	Recruited: 172	n_{max}=700	0.040	0.374	Unfavourable	348	172	0.44	348
50%	Available: 175	n_{max}=525	0.142	0.406	Unfavourable	348	250	0.46	348
	Recruited: 250	n_{max}=700	0.142	0.357	Unfavourable	348	250	0.43	348
75%	Available: 263	n_{max}=525	0.641	0.382	Promising	419	442	0.447	522
	Recruited: 325	n_{max}=700	0.641	0.328	Promising	419	442	0.41	696
HYPOTHESISED EFFECT									
25%	Available: 88	n_{max}=525	0.954	0.419	Favourable	348	323	0.466	348
	Recruited: 172	n_{max}=700	0.954	0.374	Favourable	348	323	0.440	348
50%	Available: 175	n_{max}=525	0.657	0.406	Promising	389	482	0.460	522
	Recruited: 250	n_{max}=700	0.657	0.357	Promising	389	482	0.430	696
75%	Available: 263	n_{max}=525	0.789	0.382	Favourable	348	418	0.447	406
	Recruited: 325	n_{max}=700	0.789	0.328	Favourable	348	418	0.410	464
80% OPTIMISTIC LIMIT									
25%	Available: 88	n_{max}=525	0.682	0.419	Promising	384	464	0.466	522
	Recruited: 172	n_{max}=700	0.682	0.374	Promising	384	464	0.440	696
50%	Available: 175	n_{max}=525	0.583	0.406	Promising	433	512	0.460	464
	Recruited: 250	n_{max}=700	0.583	0.357	Promising	433	512	0.430	580
75%	Available: 263	n_{max}=525	0.864	0.382	Favourable	348	393	0.447	348
	Recruited: 325	n_{max}=700	0.864	0.328	Favourable	348	393	0.410	348
90% OPTIMISTIC LIMIT									
25%	Available: 88	n_{max}=525	0.865	0.419	Favourable	348	380	0.466	348
	Recruited: 172	n_{max}=700	0.865	0.374	Favourable	348	380	0.440	348
50%	Available: 175	n_{max}=525	0.717	0.406	Promising	359	457	0.460	464
	Recruited: 250	n_{max}=700	0.717	0.357	Promising	359	457	0.430	580
75%	Available: 263	n_{max}=525	0.905	0.382	Favourable	348	377	0.447	348
	Recruited: 325	n_{max}=700	0.905	0.328	Favourable	348	377	0.410	348

Table C.18: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

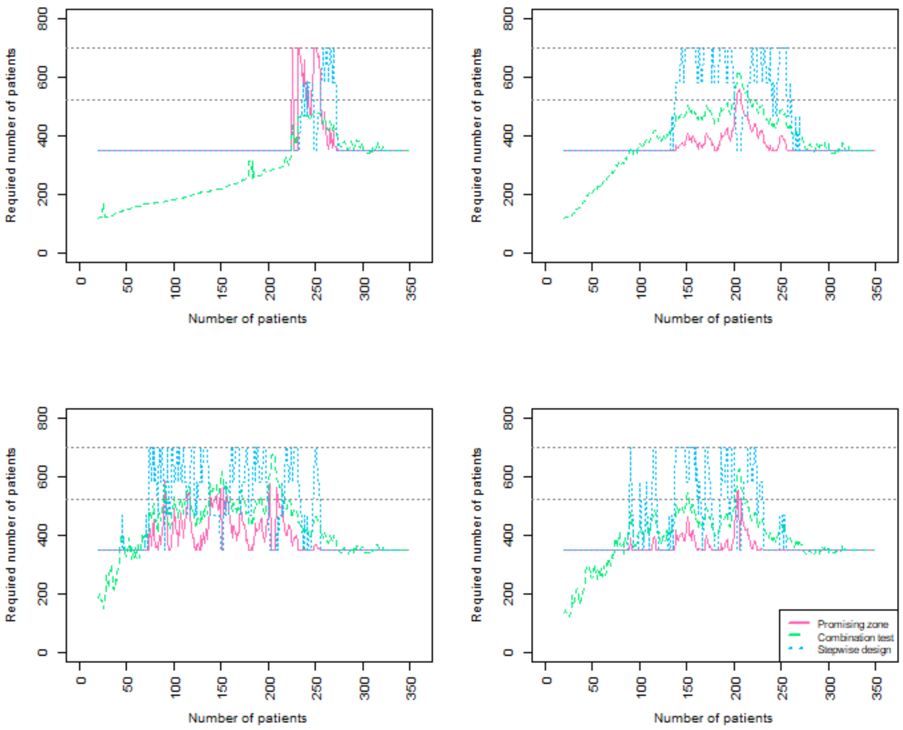
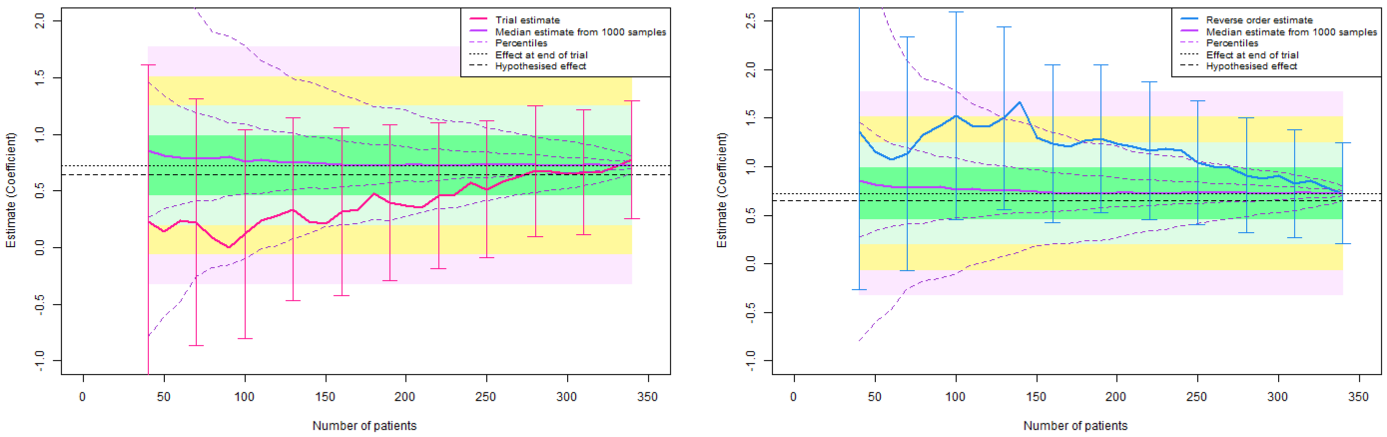


Figure C.38: Comparison of three SSR designs for the AMAZE trial data

Figure C.39 shows the estimate calculated every 10 patients using data from the AMAZE trial, for the original sequential and reverse orders. The original order estimate always lies within ±3*SEs from the assumed true treatment effect, but does not first reach the ±1*SE boundary until 180 patients (52% through the trial). From 230 patients, the estimate remains within the green boundary (66% through). At 140 patients, the reverse estimate goes beyond the yellow (±3*SE) boundary maintained by the original order. It also reaches the green (±1*SE) at a later stage than the original order (270 patients, 78% through the trial). The original order estimate mostly lies below that hypothesised, although hypothesised and that seen in the original analysis are actually very close together.



(a) *(b)*

Figure C.39: Stability of the estimate in the AMAZE trial. A comparison of sequential order (a), reverse order (b), and the median of 1000 random orders (a) and (b)

C.14 3MG trial

Similarly to the Epilepsy trial presented in Section C.7, 3MG reuses the same outcome data more than once, which affects the number recruited at the time of the three % available interim time points chosen. As CP remains the same, one further vertical line can be seen in Figure C.40; the dashed grey line representing the numbers recruited if a 7 day outcome were used (original analysis), or a dotted line for the one year (re-imagined time point) outcome. Additionally, a column has been added to Table C.20, as patients recruited so far affects the combination test design.

CP, calculated from patient 8 onwards, using the current trend largely stays <0.5, with 1020/1076 instances in the unfavourable/(futility) zones using n_{max}=1.5, and 974 for n_{max}=2 (Table C.19). The trial would have stopped for futility at the 25% data available time point, whereas CP values lie just above this boundary at the 75% time point (CP=0.118). The hypothesised effect, however, starts at 1, and does not really start decreasing before this first interim time point. Other than three main peaks of CP, this line gradually decreases, falling below the futility line by patient 917, and again from 926 onwards. The optimistic limit lines generally follow the same pattern as the hypothesised effect, but are more susceptible to fluctuations in treatment estimate; the 80% limit especially. This is particularly prominent in patients 46 to 80, with CP dropping from 0.997, to 0.342, and back up to 0.995 within the space of 35 patients.

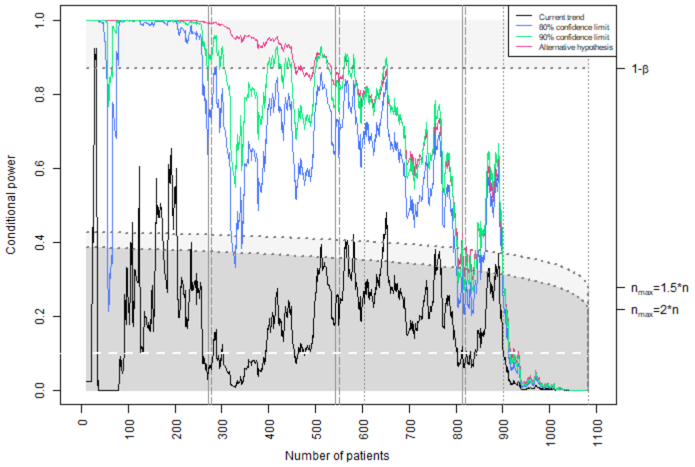


Figure C.40: Conditional power calculated after every patient in the 3MG trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	6	505	231	377
	Promising	50	349	572	473
	Unfavourable	595	58	101	57
	(Futility)	425	164	172	169
n_{max}=2	Favourable	6	505	231	377
	Promising	96	376	599	490
	Unfavourable	549	31	74	40
	(Futility)	425	164	172	169

Table C.19: Number of times CP values fall in each zone for the promising zone design for the 3MG trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table C.20 gives more details on the three chosen interim analyses, comparing three SSR designs. The promising zone design sees no increases under the current trend, and increases ranging from 7% (50% data available, hypothesised treatment effect), to 68% (75% available, 90% limit, n_{max}=2). The stepwise design sees no increase under the current trend and only one increase (75% through, n_{max}=2) under the hypothesised effect (33%).

Using the optimistic limit, n^* ranges from staying at the originally planned sample size, to increaseing to the full n_{max}=2 (50% through, 80% limit).

Figure C.41 shows n^* calculated after every n_1. Combination test design sees only decreases or the original sample size under the current trend, and small regions of a minor increase under the remaining three assumptions.Promising zone sample size rapidly fluctuates between n and n_{max} using the current trend assumption, and large regions of increasing sample size using the optimistic limits. The stepwise design has very few ares of sample size increase using the current trend, compared to a large number of times increasing to the full n_{max} using the 80% optimistic limit.

				Promising zone			Combination test (1 week)	Combination test (1 year)	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	n^*	CP_{min}	n^*
CURRENT TREND										
25%	Available: 271	n_{max}=1626	0.050	0.419	Unfavourable	1084	277	604	0.466	1084
	Recruited: 277/604	n_{max}=2168	0.050	0.374	Unfavourable	1084	277	604	0.440	1084
50%	Available: 542	n_{max}=1626	0.181	0.406	Unfavourable	1084	551	901	0.460	1084
	Recruited: 551/901	n_{max}=2168	0.181	0.357	Unfavourable	1084	551	901	0.430	1084
75%	Available: 813	n_{max}=1626	0.118	0.382	Unfavourable	1084	820	1084	0.447	1084
	Recruited: 820/1084	n_{max}=2168	0.118	0.328	Unfavourable	1084	820	1084	0.410	1084
HYPOTHESISED EFFECT										
25%	Available: 271	n_{max}=1626	0.993	0.419	Favourable	1084	641	641	0.466	1084
	Recruited: 277/604	n_{max}=2168	0.993	0.374	Favourable	1084	641	641	0.440	1084
50%	Available: 542	n_{max}=1626	0.826	0.406	Promising	1164	926	926	0.460	1084
	Recruited: 551/901	n_{max}=2168	0.826	0.357	Promising	1164	926	926	0.430	1084
75%	Available: 813	n_{max}=1626	0.423	0.382	Promising	1626	1149	1149	0.447	1084
	Recruited: 820/1084	n_{max}=2168	0.423	0.328	Promising	1763	1149	1149	0.410	1446
80% OPTIMISTIC LIMIT										
25%	Available: 271	n_{max}=1626	0.718	0.419	Promising	1459	302	604	0.466	1446
	Recruited: 277/604	n_{max}=2168	0.718	0.374	Promising	1459	302	604	0.440	1807
50%	Available: 542	n_{max}=1626	0.644	0.406	Promising	1558	844	901	0.460	1626
	Recruited: 551/901	n_{max}=2168	0.644	0.357	Promising	1558	844	901	0.430	2168
75%	Available: 813	n_{max}=1626	0.328	0.382	Unfavourable	1084	901	1084	0.447	1084
	Recruited: 820/1084	n_{max}=2168	0.328	0.328	Unfavourable	1084	901	1084	0.410	1084
90% OPTIMISTIC LIMIT										
25%	Available: 271	n_{max}=1626	0.886	0.419	Favourable	1084	725	725	0.466	1084
	Recruited: 277/604	n_{max}=2168	0.886	0.374	Favourable	1084	725	725	0.440	1084
50%	Available: 542	n_{max}=1626	0.768	0.406	Promising	1275	923	923	0.460	1265
	Recruited: 551/901	n_{max}=2168	0.768	0.357	Promising	1275	923	923	0.430	1446
75%	Available: 813	n_{max}=1626	0.407	0.382	Promising	1626	1126	1126	0.447	1084
	Recruited: 820/1084	n_{max}=2168	0.407	0.328	Promising	1819	1126	1126	0.410	1084

Table C.20: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

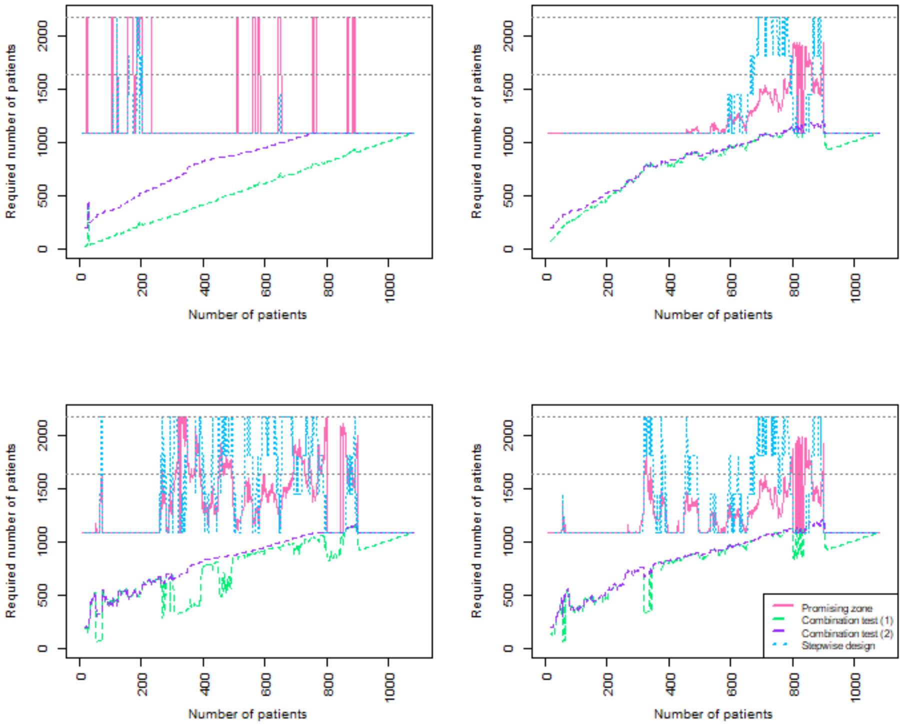
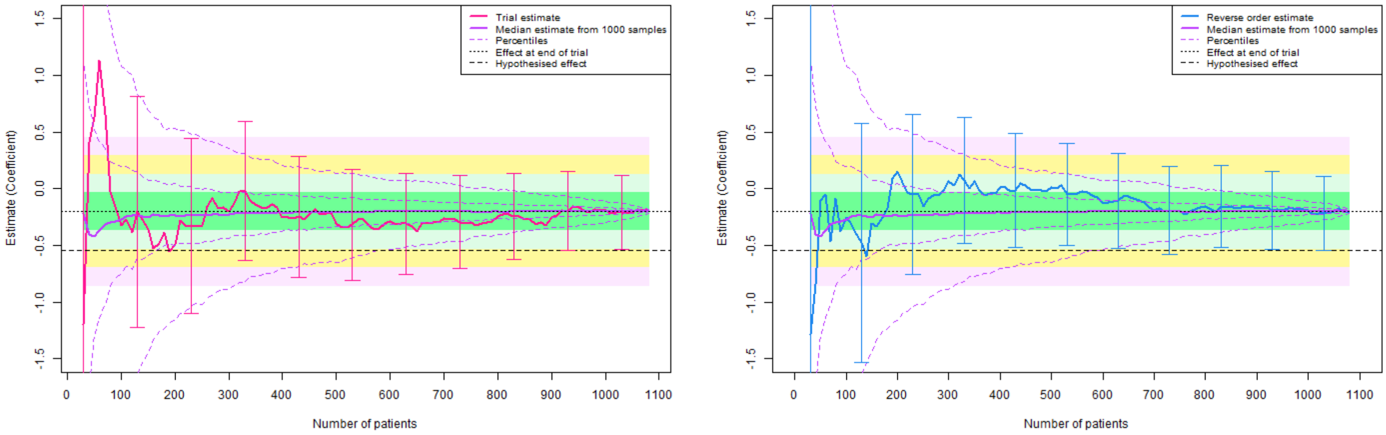


Figure C.41: Comparison of three SSR designs for the 3MG trial data

The original order and reverse estimates calculated after every 10 patients can be seen in Figure C.42. The original estimate is highly variable at the start, and does not keep within even the ±4*SE boundaries until 80 patients onwards. After this point, it quickly reaches the ±2*SE boundaries, and other than straying into the next boundary at 190 patients, stays within these limits. The estimate lies entirely within ±1*SE boundary only from patient 660 onwards (61% through the trial). Whilst also starting outside any investigated boundary, the reverse order estimate is quicker to reach each boundary compared to the original order estimate. The reverse estimate reaches the inner ±1*SE boundaries at 540 patients and remains there from that point forward.



(a) *(b)*

Figure C.42: Stability of the estimate in the 3MG trial. A comparison of sequential order (a), reverse order (b), and the median of 1000 random orders (a) and (b)

C.15 RATPAC trial

Figure C.43 shows CP values calculates after every patient in the RATPAC trial. Both optimistic limits and the hypothesised effect remain consistently at 1 throughout the trial duration. The current trend assumption has one small peak between patients 53 and 188, but remains in the favourable zone. It should be noted that the RATPAC trial terminated early due to slow recruitment and a CP calculation, and so power has been adjusted from the original 80%. Because CP remains consistently high, no sample size increase can be seen for promising zone or stepwise designs at the 3 interim timepoint (Table C.21). Sample size decreases from the combination test range from 44% to 95% of the original 2243 patients. Extending the n^* calculations to every n_1 (Figure C.44), the stepwise design using the current trend has one peak early on in the trial duration, but otherwise remains at 2243 patints. Promising zone design would have always remained at 2243 patients under any assumption, and the combination test would have decreased the sample size anywhere before 1800 patients, or remained at the 2243 at any point after this.

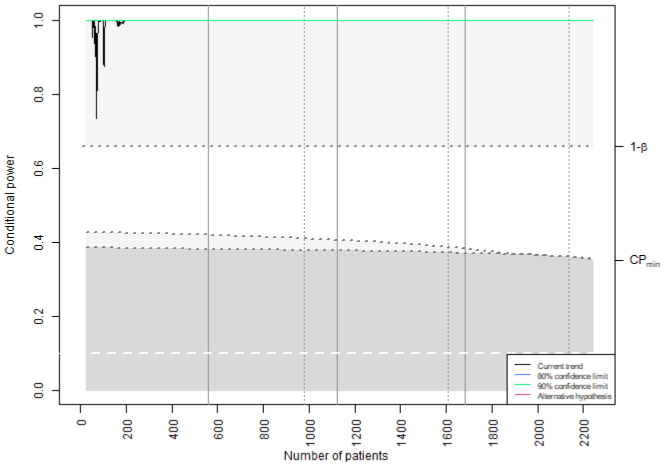


Figure C.43: Conditional power calculated after every patient in the RATPAC trial

			Promising zone			Combination test	Stepwise design	
	Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND								
25% Available: 561	n_{max}=3365	1.000	0.419	Favourable	2243	978	0.466	2243
Recruited: 978	n_{max}=4486	1.000	0.382	Favourable	2243	978	0.440	2243
50% Available: 1122	n_{max}=3365	1.000	0.406	Favourable	2243	1606	0.460	2243
Recruited: 1606	n_{max}=4486	1.000	0.377	Favourable	2243	1606	0.430	2243
75% Available: 1683	n_{max}=3365	1.000	0.382	Favourable	2243	2136	0.447	2243
Recruited: 2136	n_{max}=4486	1.000	0.371	Favourable	2243	2136	0.410	2243
HYPOTHESISED EFFECT								
25% Available: 561	n_{max}=3365	1.000	0.419	Favourable	2243	978	0.466	2243
Recruited: 978	n_{max}=4486	1.000	0.382	Favourable	2243	978	0.440	2243
50% Available: 1122	n_{max}=3365	1.000	0.406	Favourable	2243	1606	0.460	2243
Recruited: 1606	n_{max}=4486	1.000	0.377	Favourable	2243	1606	0.430	2243
75% Available: 1683	n_{max}=3365	1.000	0.382	Favourable	2243	2136	0.447	2243
Recruited: 2136	n_{max}=4486	1.000	0.371	Favourable	2243	2136	0.410	2243
80% OPTIMISTIC LIMIT								
25% Available: 561	n_{max}=3365	1.000	0.419	Favourable	2243	978	0.466	2243
Recruited: 978	n_{max}=4486	1.000	0.382	Favourable	2243	978	0.440	2243
50% Available: 1122	n_{max}=3365	1.000	0.406	Favourable	2243	1606	0.460	2243
Recruited: 1606	n_{max}=4486	1.000	0.377	Favourable	2243	1606	0.430	2243
75% Available: 1683	n_{max}=3365	1.000	0.382	Favourable	2243	2136	0.447	2243
Recruited: 2136	n_{max}=4486	1.000	0.371	Favourable	2243	2136	0.410	2243
90% OPTIMISTIC LIMIT								
25% Available: 561	n_{max}=3365	1.000	0.419	Favourable	2243	978	0.466	2243
Recruited: 978	n_{max}=4486	1.000	0.382	Favourable	2243	978	0.440	2243
50% Available: 1122	n_{max}=3365	1.000	0.406	Favourable	2243	1606	0.460	2243
Recruited: 1606	n_{max}=4486	1.000	0.377	Favourable	2243	1606	0.430	2243
75% Available: 1683	n_{max}=3365	1.000	0.382	Favourable	2243	2136	0.447	2243
Recruited: 2136	n_{max}=4486	1.000	0.371	Favourable	2243	2136	0.410	2243

Table C.21: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

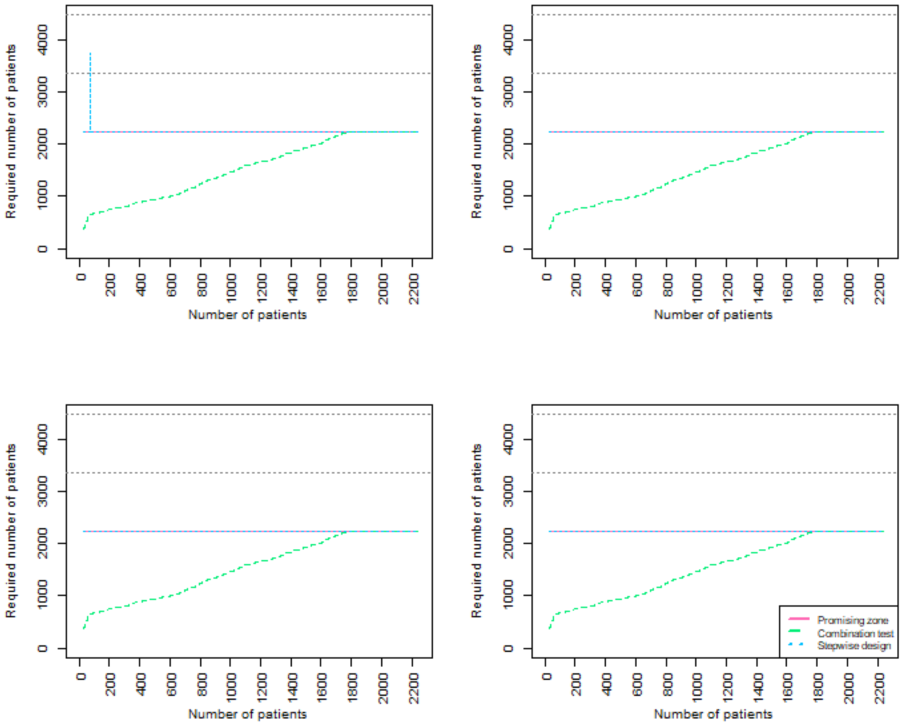
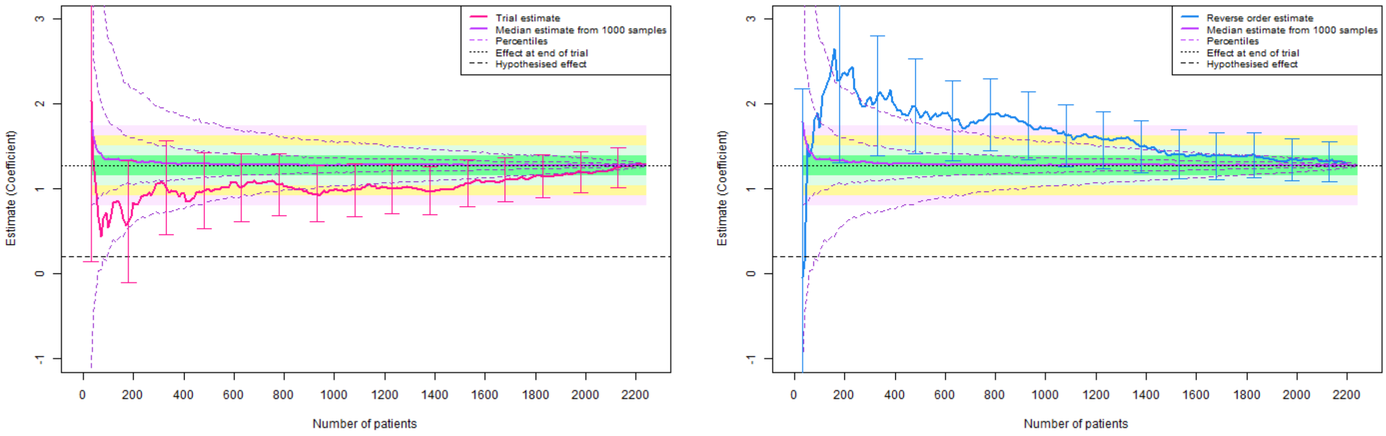


Figure C.44: Comparison of three SSR designs for the RATPAC trial data

Figure C.45 shows original sequential and reverse order estimates from the RATPAC trial. Both original order and reverse order estimates are highly variable towards the beginning of the study. The original order starts high but quickly drops below the -4*SE boundary within 40 patients. From 200 patients onwards, the estimate is contained within the outer boundary investigated (±4*SE). It does not reach the inner (±1*SE) boundary (green) until 1910 patients, 85% through the trial duration (other than briefly at 40 patients). In terms of the ±1*SE boundary, the reverse order is similar to the original sequential order, in that it briefly reaches this limit at 60 patients, but does not remain there until 1840 patients (82% through). However, the reverse order is much higher than the assumed true treatment effect, and does not remain even within the ±4*SE boundaries until patient 930 (41%). This highlights how different patients are at the start than at the end of the trial. Additionally, had an interim analysis stopped the trial earlier than it did, some CIs do not contain the assumed "true" treatment effect, and would therefore have under-estimated the treatment effect with the original order in mind.



(a)

(b)

Figure C.45: Stability of the estimate in the RATPAC trial. A comparison of sequential order (a), reverse order (b), and the median of 1000 random orders (a) and (b)

C.16 Nasal sprays trial

Figure C.46 shows CP values for the four future treatment effect assumptions using the data from the nasal sprays trials. All four lines start and end at 1, despite a late dip in CP from 260 patients, resulting in 7 instances in the promising zone under the current trend, or 5 for the other three assumptions. This peak corresponds with Figure C.47, showing new total sample size n^* for every n_1. A small increase can be seen under any assumption for the combination test and promising zone designs, compared to a large peak of n_{max} using the stepwise design.

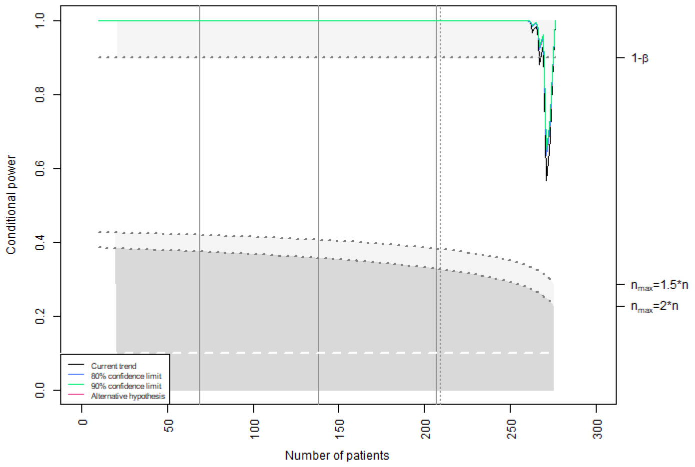


Figure C.46: Conditional power calculated after every patient in the Nasal sprays trial

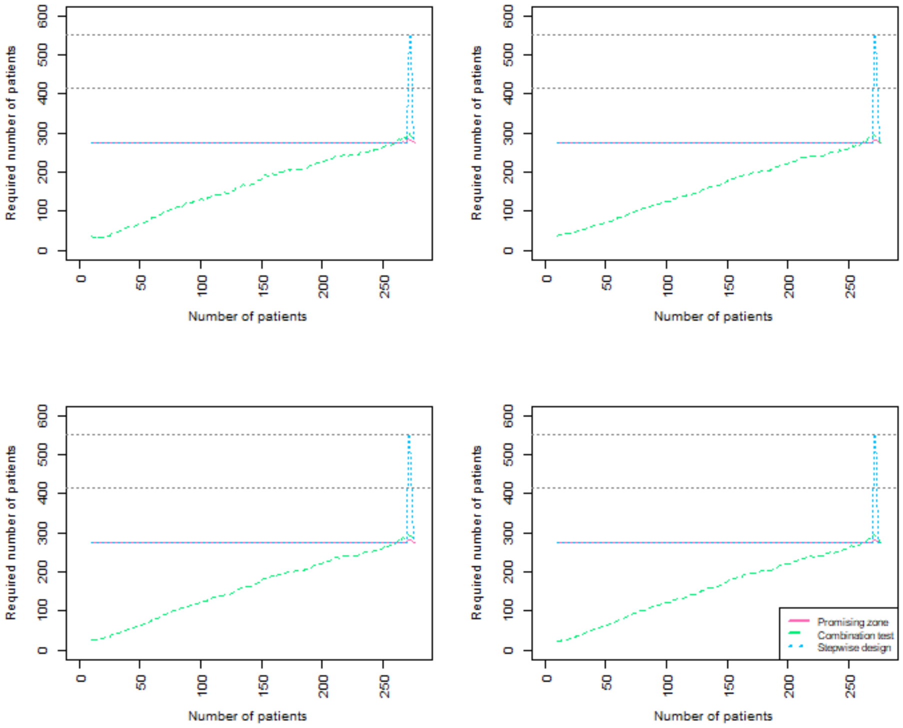


Figure C.47: Comparison of three SSR designs for the Nasal spray trial data

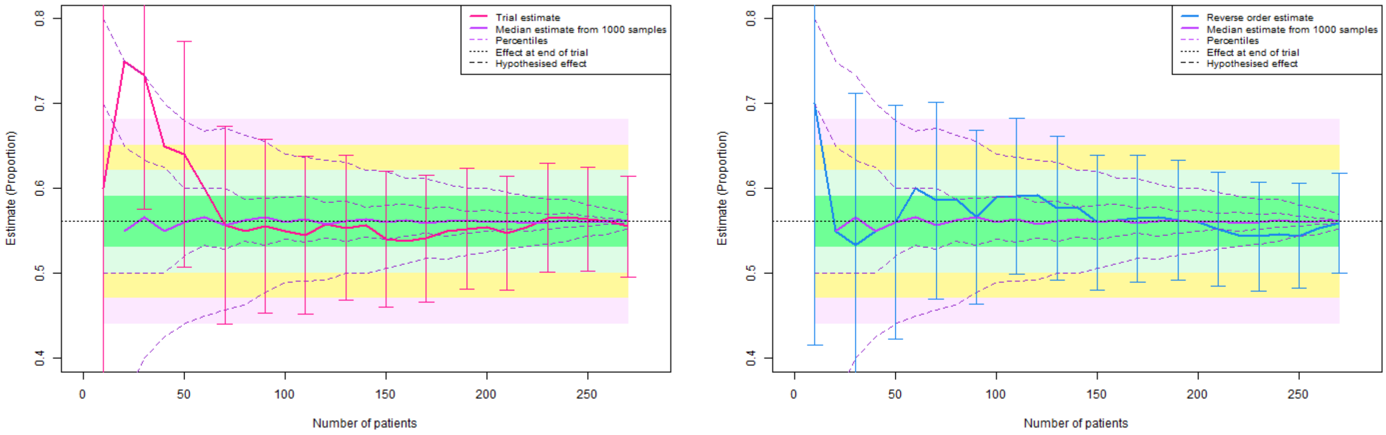
The change in new total sample size further highlighted in Table C.22, looking at the three chosen time points in more depth. No sample size increase is seen at any interim time point for promising zone or stepwise designs. Sample size decreases using the combination test design range from just 32% of the original sample size, to 89%.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 69	n_{max}=414	1.000	0.419	Favourable	276	97	0.466	276
	Recruited: 69	n_{max}=552	1.000	0.374	Favourable	276	97	0.440	276
50%	Available: 138	n_{max}=414	1.000	0.406	Favourable	276	166	0.460	276
	Recruited: 138	n_{max}=552	1.000	0.357	Favourable	276	166	0.430	276
75%	Available: 207	n_{max}=414	1.000	0.382	Favourable	276	233	0.447	276
	Recruited: 209	n_{max}=552	1.000	0.328	Favourable	276	233	0.410	276
HYPOTHESISED EFFECT									
25%	Available: 69	n_{max}=414	1.000	0.419	Favourable	276	94	0.466	276
	Recruited: 69	n_{max}=552	1.000	0.374	Favourable	276	94	0.440	276
50%	Available: 138	n_{max}=414	1.000	0.406	Favourable	276	162	0.460	276
	Recruited: 138	n_{max}=552	1.000	0.357	Favourable	276	162	0.430	276
75%	Available: 207	n_{max}=414	1.000	0.382	Favourable	276	229	0.447	276
	Recruited: 209	n_{max}=552	1.000	0.328	Favourable	276	229	0.410	276
80% OPTIMISTIC LIMIT									
25%	Available: 69	n_{max}=414	1.000	0.419	Favourable	276	90	0.466	276
	Recruited: 69	n_{max}=552	1.000	0.374	Favourable	276	90	0.440	276
50%	Available: 138	n_{max}=414	1.000	0.406	Favourable	276	161	0.460	276
	Recruited: 138	n_{max}=552	1.000	0.357	Favourable	276	161	0.430	276
75%	Available: 207	n_{max}=414	1.000	0.382	Favourable	276	229	0.447	276
	Recruited: 209	n_{max}=552	1.000	0.328	Favourable	276	229	0.410	276
90% OPTIMISTIC LIMIT									
25%	Available: 69	n_{max}=414	1.000	0.419	Favourable	276	88	0.466	276
	Recruited: 69	n_{max}=552	1.000	0.374	Favourable	276	88	0.440	276
50%	Available: 138	n_{max}=414	1.000	0.406	Favourable	276	160	0.460	276
	Recruited: 138	n_{max}=552	1.000	0.357	Favourable	276	160	0.430	276
75%	Available: 207	n_{max}=414	1.000	0.382	Favourable	276	228	0.447	276
	Recruited: 209	n_{max}=552	1.000	0.328	Favourable	276	228	0.410	276

Table C.22: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

The estimate for the nasal spray trial starts in the ±2*SE boundary at 10 patients using the original sequential order (Figure C.48). However, it then ventures beyond the upper 4*SE limit shortly after, over-estimating the treatment effect assumed to be true. By 40 patients, the estimate has moved back within the ±3*SE boundary, and by 70 (25% through the trial), to the ±1*SD boundary, where it remains. The reverse order estimate however, starts just above the ±4*SE zone, but quickly falls below. From patient 130 onwards (47%

through the trial) the reverse order estimate remains within the ±1*SE limit boundaries, where it stays.



(a) *(b)*

Figure C.48: Stability of the estimate in the Nasal sprays trial. A comparison of sequential order (a), reverse order (b), and the median of 1000 random orders (a) and (b)

C.17 Mencevax (Strain A) trial

Figure C.49 shows CP after every patient. The Mencevax outcome assumed rate was very high (98% in both groups), and as no patient had a negative outcome in one group until around patient 70, CP under the current trend starts very low, and suddenly spikes at the first instance of a non-zero cell in the outcome and group 2x2 table. From this point forward, the CP remains high for all four CP lines. Due to the 99% power of the trial, the favourable zone is very small on the graph, but is reached in a large number of instances under all four assumptions (Table C.23).

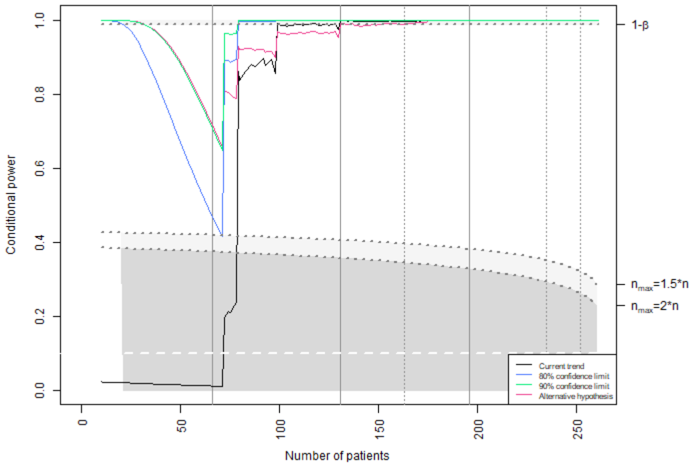


Figure C.49: Conditional power calculated after every patient in the Mencevax (Strain A) trial

		Current trend	Hyp. effect	80% limit	90% limit
n_{max}=1.5	Favourable	139	137	194	205
	Promising	44	115	57	47
	Unfavourable	7	0	1	0
	(Futility)	62	0	0	0
n_{max}=2	Favourable	139	137	194	205
	Promising	44	115	58	47
	Unfavourable	7	0	0	0
	(Futility)	62	0	0	0

Table C.23: Number of times CP values fall in each zone for the promising zone design for the Mencevax (Strain A) trial. For a design where no futility boundary is considered, these values become unfavourable instead.

Table C.24 gives further details of the three interim time points for the three designs. At the first interim time point using the current trend, all designs would have stopped for futility, if the 10% boundary had been used. Otherwise, promising zone designs and stepwise designs would have continued to 261 patients, compared to a decrease of 38% using the combination test design. Using the other three assumptions, all three designs would only see increases at the 25% point, ranging from 17% (stepwise design, 80% limit, n_{max}=1.5), to 100% (promising zone design, n_{max}=2, optimistic limits and hypothesised effect).

Figure C.50 shows n^* after every n_1. The stepwise design has the smallest region of sample size increase in all four assumptions, but still reaches n_max=2 in three of the four.

The promising zone design has the largest region of increase, the largest being under the hypothesised effect. All four instances see n_{max} being reached. The combination test sees n_{max}=1.5 being reached in three cases, almost extending to the n_{max}=2 using the 80% optimistic confidence limit.

				Promising zone			Combination test	Stepwise design	
		Max increase	CP	CP_{min}	Zone	n^*	n^*	CP_{min}	n^*
CURRENT TREND									
25%	Available: 66	n_{max}=392	0.011	0.419	Unfavourable	261	163	0.466	261
	Recruited: 163	n_{max}=522	0.011	0.374	Unfavourable	261	163	0.440	261
50%	Available: 131	n_{max}=392	0.999	0.406	Favourable	261	235	0.459	261
	Recruited: 235	n_{max}=522	0.999	0.357	Favourable	261	235	0.430	261
75%	Available: 196	n_{max}=392	1.000	0.382	Favourable	261	252	0.446	261
	Recruited: 252	n_{max}=522	1.000	0.328	Favourable	261	252	0.410	261
HYPOTHESISED EFFECT									
25%	Available: 66	n_{max}=392	0.718	0.419	Promising	392	372	0.466	348
	Recruited: 163	n_{max}=522	0.718	0.374	Promising	522	372	0.440	435
50%	Available: 131	n_{max}=392	0.990	0.406	Favourable	261	235	0.459	261
	Recruited: 235	n_{max}=522	0.990	0.357	Favourable	261	235	0.430	261
75%	Available: 196	n_{max}=392	1.000	0.382	Favourable	261	252	0.446	261
	Recruited: 252	n_{max}=522	1.000	0.328	Favourable	261	252	0.410	261
80% OPTIMISTIC LIMIT									
25%	Available: 66	n_{max}=392	0.470	0.419	Promising	392	392	0.466	305
	Recruited: 163	n_{max}=522	0.470	0.374	Promising	522	480	0.440	348
50%	Available: 131	n_{max}=392	1.000	0.406	Favourable	261	235	0.459	261
	Recruited: 235	n_{max}=522	1.000	0.357	Favourable	261	235	0.430	261
75%	Available: 196	n_{max}=392	1.000	0.382	Favourable	261	252	0.446	261
	Recruited: 252	n_{max}=522	1.000	0.328	Favourable	261	252	0.410	261
90% OPTIMISTIC LIMIT									
25%	Available: 66	n_{max}=392	0.709	0.419	Promising	392	376	0.466	348
	Recruited: 163	n_{max}=522	0.709	0.374	Promising	522	376	0.440	435
50%	Available: 131	n_{max}=392	1.000	0.406	Favourable	261	235	0.459	261
	Recruited: 235	n_{max}=522	1.000	0.357	Favourable	261	235	0.430	261
75%	Available: 196	n_{max}=392	1.000	0.382	Favourable	261	252	0.446	261
	Recruited: 252	n_{max}=522	1.000	0.328	Favourable	261	252	0.410	261

Table C.24: New total sample size required for each of the three designs investigated at three time points and two values of n_{max}. CP_{min} values are given for the promising zone and stepwise designs, and zone is given for the promising zone design.

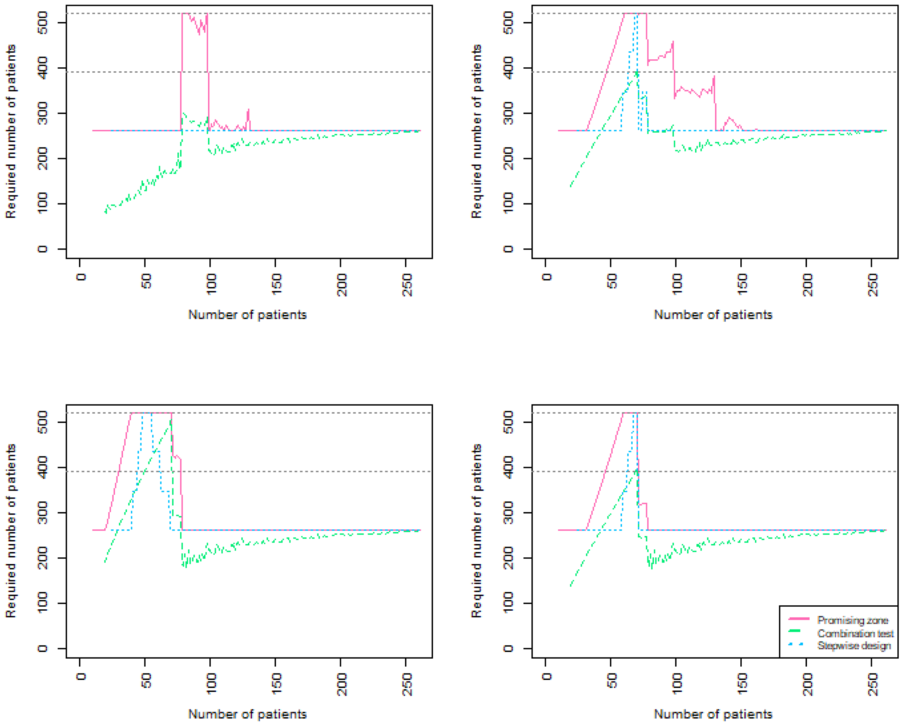
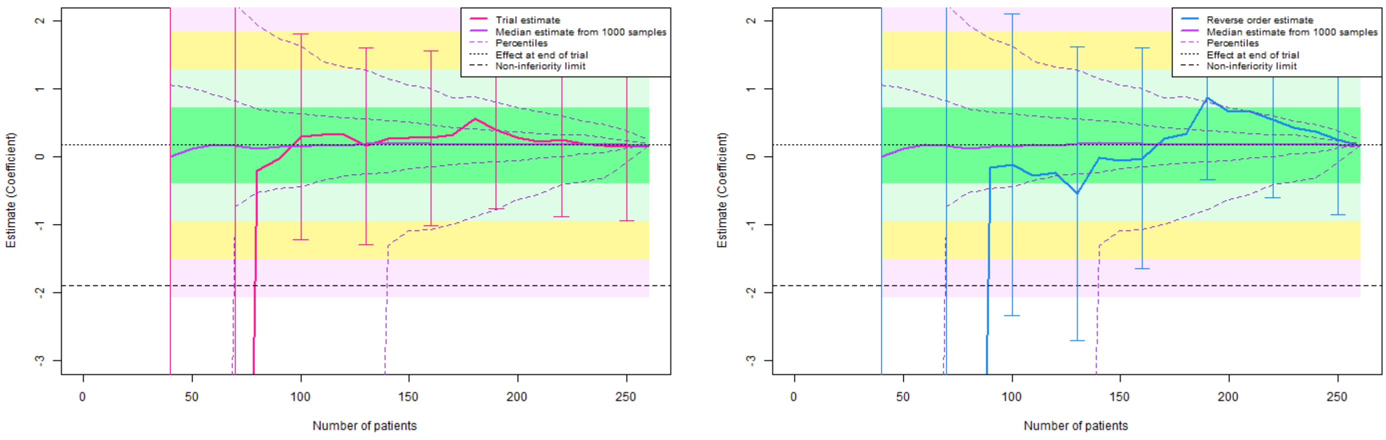


Figure C.50: Comparison of three SSR designs for the Mencevax (Strain A) trial data

Figure C.51 shows the original and reverse ordered estimate calculated every 10 patients from patient 40 onwards. Prior to patient 80, no valid estimate can be obtained due to the rare nature of a negative outcome (2% assumed rate in each group). From this point forward however, the original order estimate always lies within the smallest investigated boundaries, ±1*SE. The reverse order starts in the same way and has no valid estimate prior to patient 90. After this point, it reaches the ±1*SE boundary. However, it also reaches the lower, and later the upper, limit of the ±2*SE boundaries. From patient 200 (77% through) however, the reverse order estimate remains inside the ±1*SE boundary.



(a) *(b)*

Figure C.51: Stability of the estimate in the Mencevax A trial. A comparison of sequential order (a), reverse order (b), and the median of 1000 random orders (a) and (b)

C.18 Mencevax (Strain C) trial

CP (shown in Figure C.52) is 1 throughout the trial under either optimistic limit of hypothesised effect assumption. Under the current trend, CP starts at almost zero, rapidly peaking to ≈1 at just 20 patients, where it remains for the remainder of the trial. Because of the extended time frame for primary outcome data collection (1 year), All patients have been recruited by the first interim time point (25% data available). Therefore no decreases in sample size can be seen using the combination test. Additionally, because of the very high CP values seen almost everywhere, no increase in sample size can be seen using the promising zone or stepwise designs either at any value of n_1 (Figure C.53). Note that if a futility bound had been used, and a very early interim analysis prior to 20 patients, using the current trend would have resulted in the trial stopping for futility.

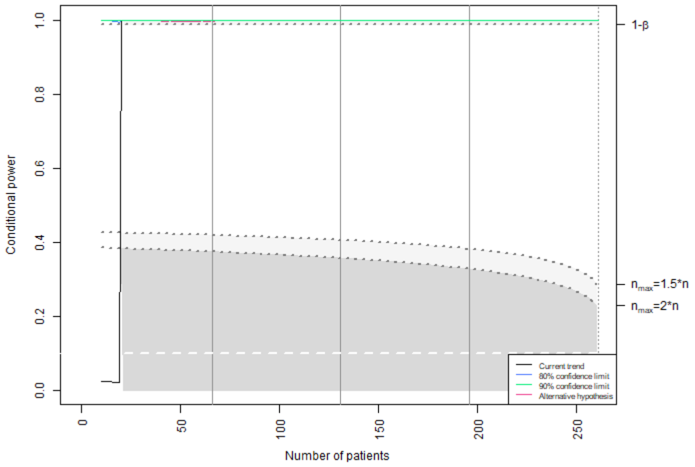


Figure C.52: Conditional power calculated after every patient in the Mencevax (Strain C) trial

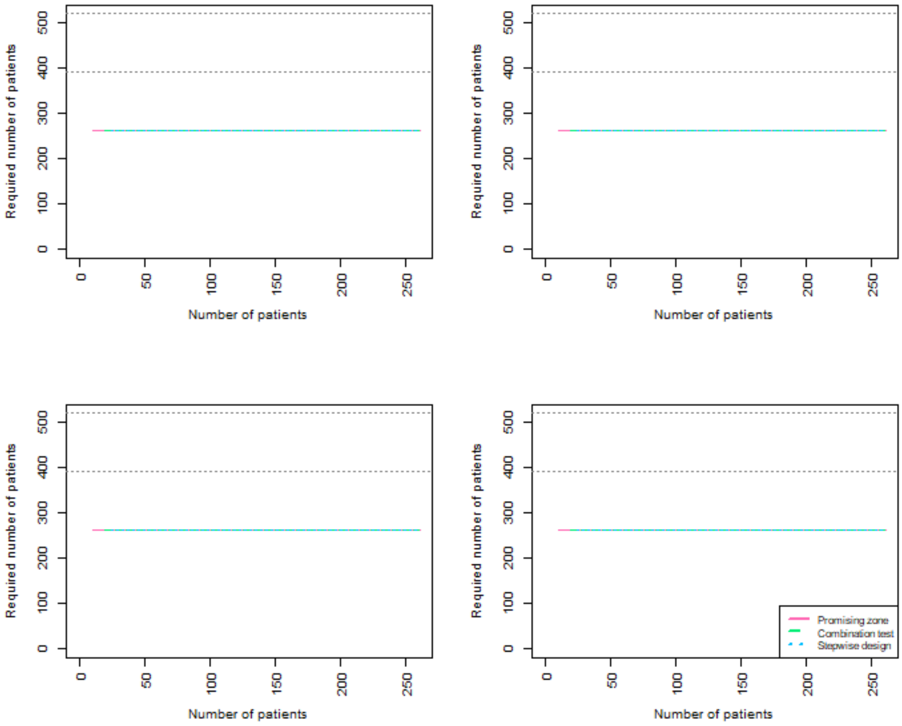
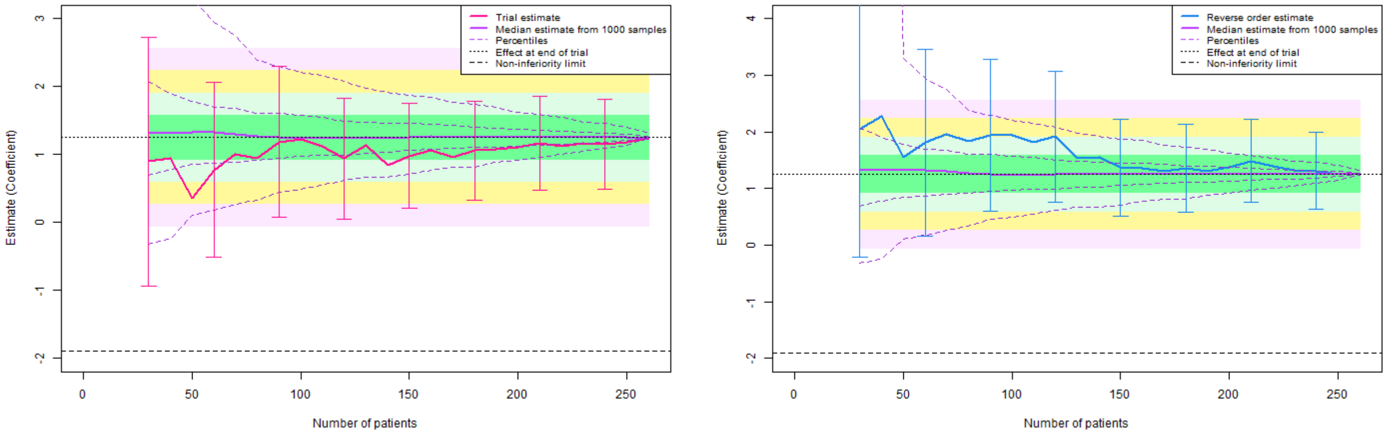


Figure C.53: Comparison of three SSR designs for the Mencevax (Strain C) trial data

Figure C.54 shows the original and reverse order estimates from the Mencevax trial, looking at the strain C data only. In both cases, the estimate lies well above the non-inferiority limit. The original order lies mainly underneath the actual treatment effect from the original analysis, under-estimating this value. The estimate always lies within ±3*SE boundary, first reaching the ±1*SE limit at patient 40, but only remains there from patient 150 onwards. The reverse estimate however slightly over-estimates the treatment effect early in the trial, and only remains within the ±4*SE boundaries throughout. It first reaches the ±1*SE limit at 50 patients (19% through the trial), but remains there from this point forward only from patient 130 (50%).



(a) *(b)*

Figure C.54: Stability of the estimate in the Mencevax C trial. A comparison of sequential order (a), reverse order (b), and the median of 1000 random orders (a) and (b)

C.19 FLU A/H1N1 (Seroconversion) trial

CP values using the A/H1N1 strain data from the flu vaccine trial is shown in Figure C.55. Optimistic confidence limit lines are remain ≈1 throughout the trial. The hypothesised effect assumption line (calculated from patient 10 onwards) begins just outside the favourable zone, and remains intermittently between this boundary for the first 400 patients. After this point, CP remains in the favourable zone, reaching ≈1 by 1600 patients. The current trend line on the other hand, is highly variable, rapidly switching between zones, before steadily increasing to 1 by 1600 patients.

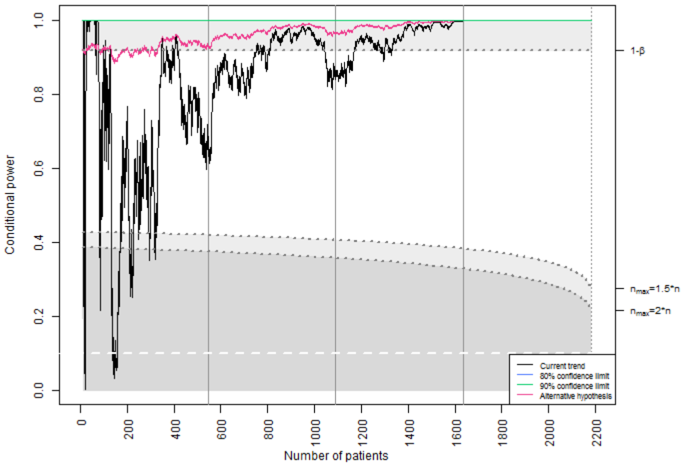


Figure C.55: Conditional power calculated after every patient in the Flu A/H1N1 (Seroconversion) trial

Under the current trend, a sample size increase is seen to the full allowed maximum at 25% data available with promising zone and stepwise designs, and an increase of 26% at either maximum at the 50% data available time point using the promising zone design. Combination test designs see no increase at any of the three interim time points investigated. Figure C.56 shows new n^* after every n_1 patients. A few small increases can be seen in the early stages of the trial (<400 patients) using the promising zone design and hypothesised effect assumption. However, under the current trend assumption, both stepwise and promising zone designs see large increases in the first third of the trial, and more modest increases in the second third for the promising zone design, and almost no increases in the same region for the stepwise design. Again, no decreases in sample size can be seen due to the rapid recruitment rate of the trial, resulting in all patients being recruited before primary outcome data becomes available.

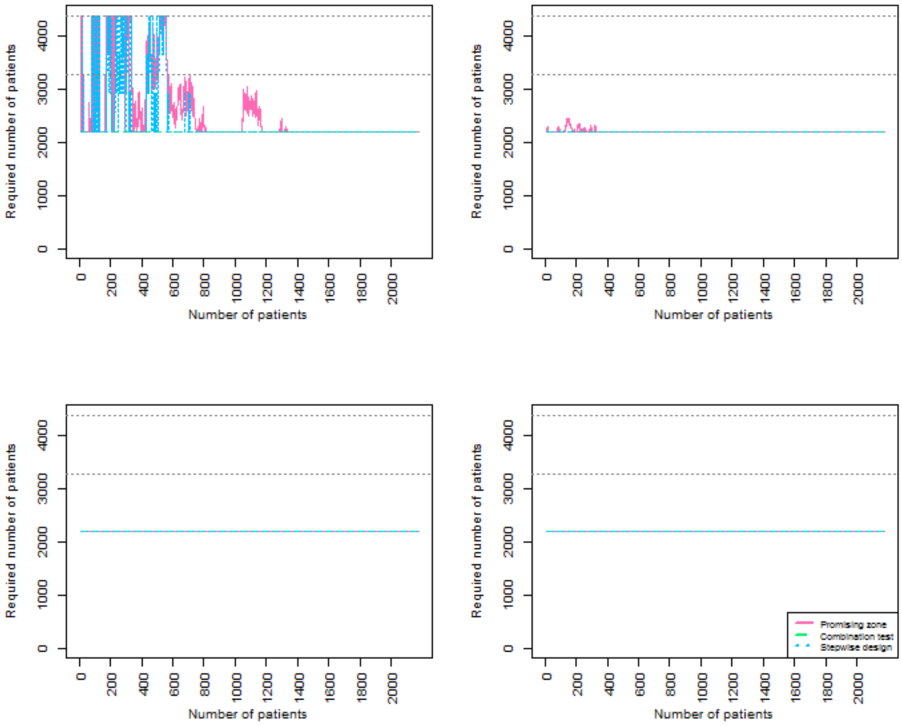
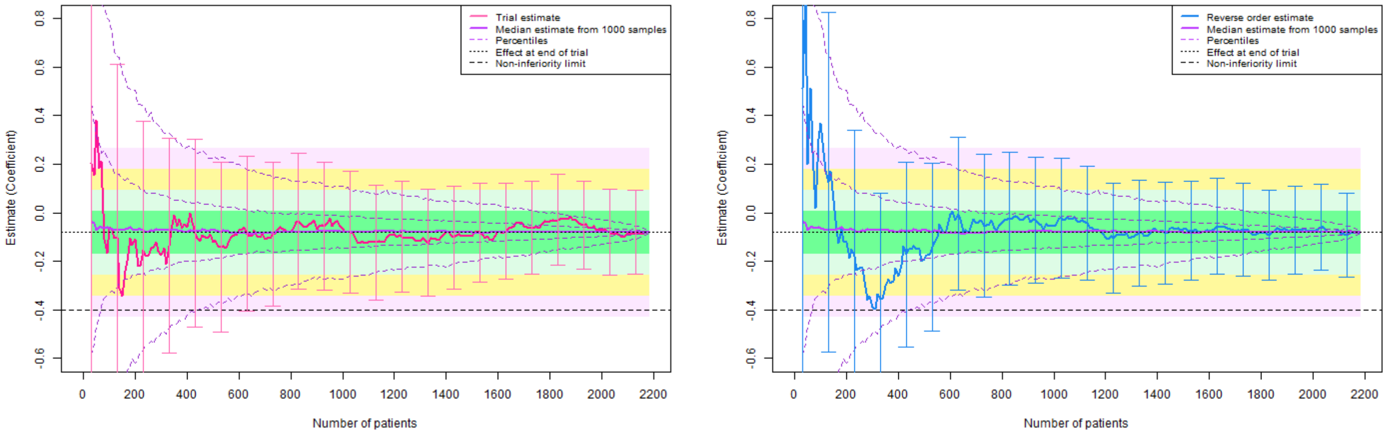


Figure C.56: Comparison of three SSR designs for the Flu A/H1N1 (Seroconversion) trial data

The estimate for the flu A/H1N1 vaccine using the seroconversion (binary) endpoint remains above the non-inferiority limit at all times for both the original sequential order and the reverse order, although it does get close in one instance for the reverse order estimate. The original order estimate remains in the ±1*SE boundary from an earlier time point than the reverse order (330 patients (15%), compared to 620 (28%)). The biggest differences (furthest away) from the end estimate value ("true" treatment effect) are seen in the reverse order, which starts well above the upper 4*SE limit, and does not stay within these bounds until patient 110 (5% onwards).



(a) *(b)*

Figure C.57: Stability of the estimate in the Flu A/H1N1 vaccine trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.20 FLU A/H3N2 (Seroconversion) trial

Figure C.58 shows CP values calculated after every patient. In the first 600 patients, the optimistic limits vary considerably, ranging from zero to one. After this point however, both lines setlle and remain ≈1 until the end of the trial. The current trend on the other hand, whilst still inclined to variable peaks, have less rapid changes than seen in the optimistic limits assumptions. The hypothesised treatment effect remains ≥0.7 at all times, but does somewhat fluctuate, before steadily increasing to ≈1 with the other three lines (∼ patient 1800).

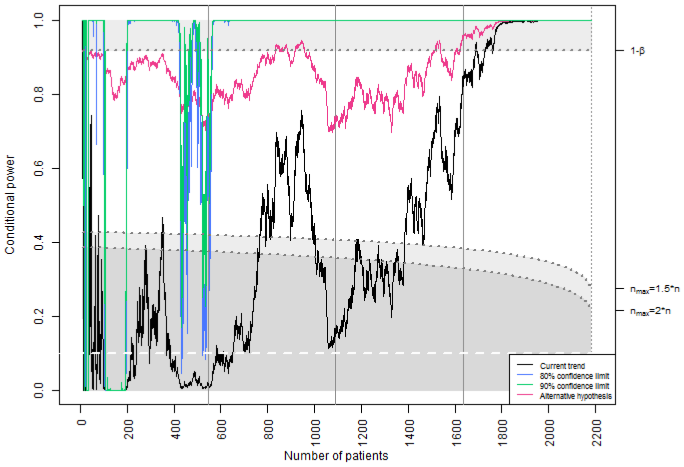


Figure C.58: Conditional power calculated after every patient in the Flu A/H3N2 (Seroconversion) trial

Again, no increase is seen for the combination test design. Promising zone design increases range from 26% (25% data available, 90% limit), to 82% (25% time point, 80% limit, n_{max}=2). Figure C.59 shows n^* calculated after every n_1 patients. Current trend and hypothesised assumptions have a large region of increased sample sizes for the promising zone and stepwise designs, with more moderate increases in the hypothesised effect, compared to maximum increases being reached under the current trend. Both optimistic limits have smaller regions of increase for both designs, but again, maximum values are reached, leading to up to doubling of sample size.

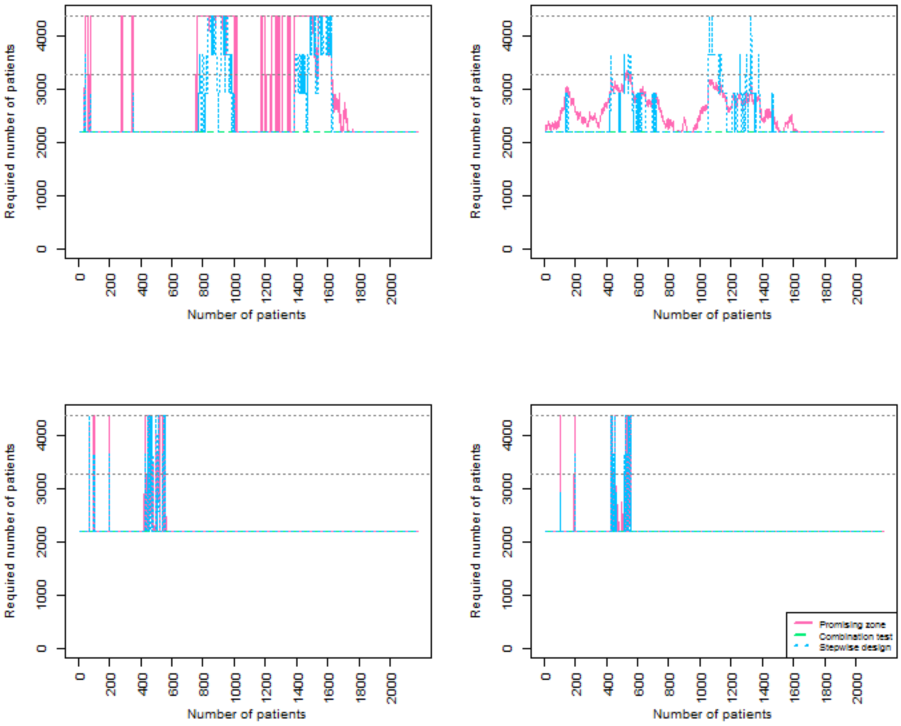
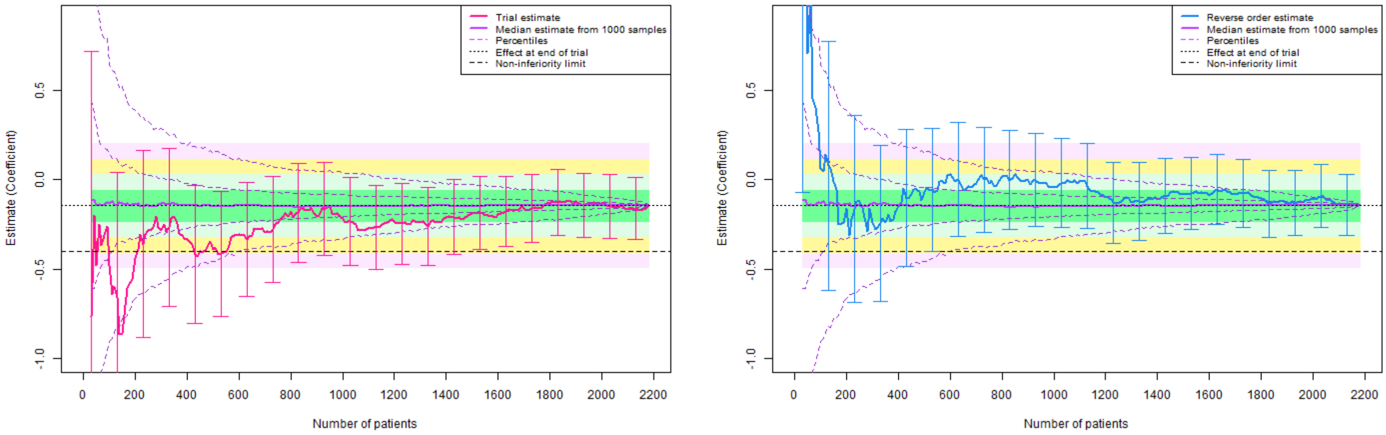


Figure C.59: Comparison of three SSR designs for the Flu A/H3N2 (Seroconversion) trial data

Figure C.60 shows the sample estimate through the trial for the A/H3N2 strain and the flu vaccine trial in the original (a) and reverse (b) order. The original order estimate starts by alternating above and below the non-inferiority limit, first reaching the ±1*SE boundary at 40 patients, but not remaining there until after 1390 patients (64% through the trial). The estimate largely stays within the ±4*SE limits, except for a peak outside this boundary early in the trial, resolving by 200 patients. The reverse order however, always remains above the non-inferiority limit, even above the upper +4*SE boundary for the first 200 patients (i.e. the last 200 patients recruited in the original order). The reverse order estimate is slower to reach and remain in the ±1*SE boundary; from 1670 patients, 77% through the trial.



(a) *(b)*

Figure C.60: Stability of the estimate in the Flu A/H3N2 vaccine trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.21 FLU B1 (Seroconversion) trial

Figure C.61 shows CP values for the B1 strain data in the flu vaccine trial. All four lines have reached ≈1 by 400 patients, and remain there throughout the trial duration. Optimistic limits and current trend lines rapidly fluctuate between minimum and maximum values of CP very early on in the trial, but settle by 100 patients, other than some small drops in CP under the current trend assumption, which still remain in the favourable zone. Because of the relatively quick convergence to 1 in CP values, only the current trend assumption yields any increase in sample size in both promising zone and stepwise design (Figure C.62). However, the trial would remain at the original sample size after this early peak. Specifically at the three chosen interim time points, all designs would have remained at the original 1634 patients.

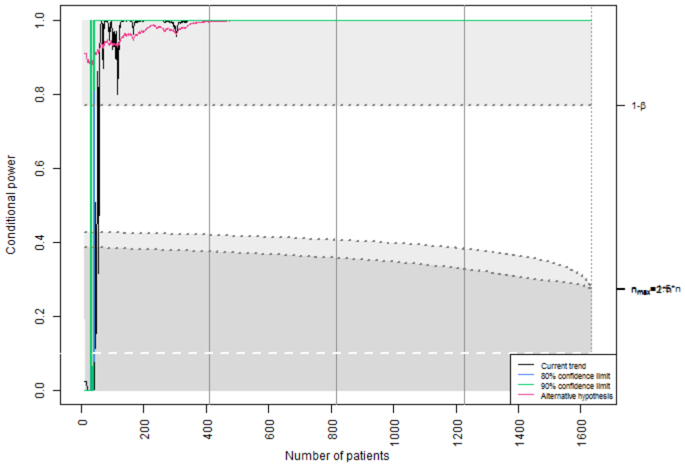


Figure C.61: Conditional power calculated after every patient in the Flu B1 trial using the Seroconversion endpoint

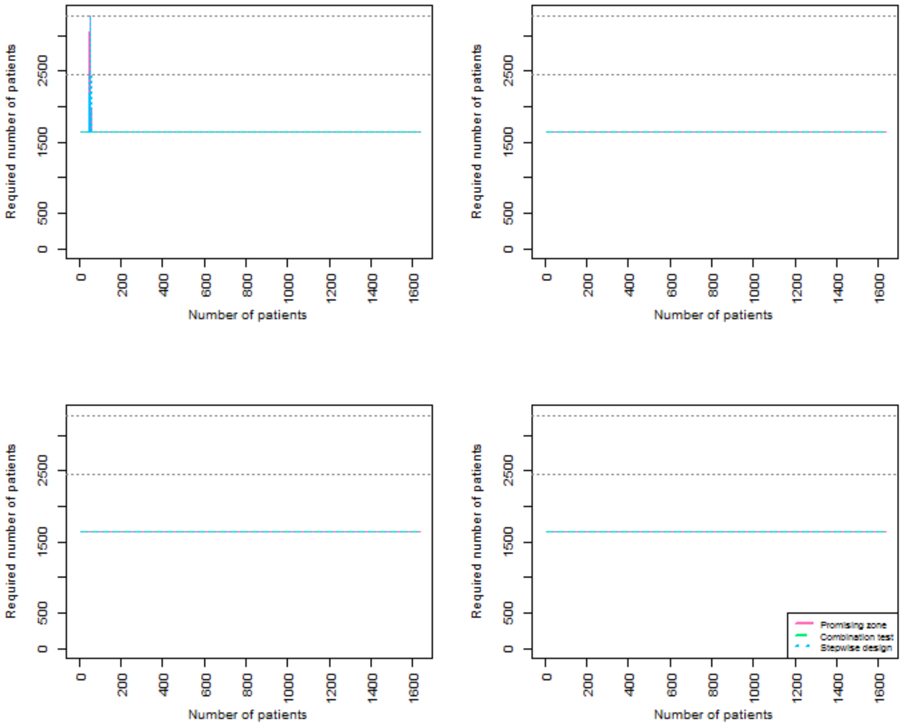
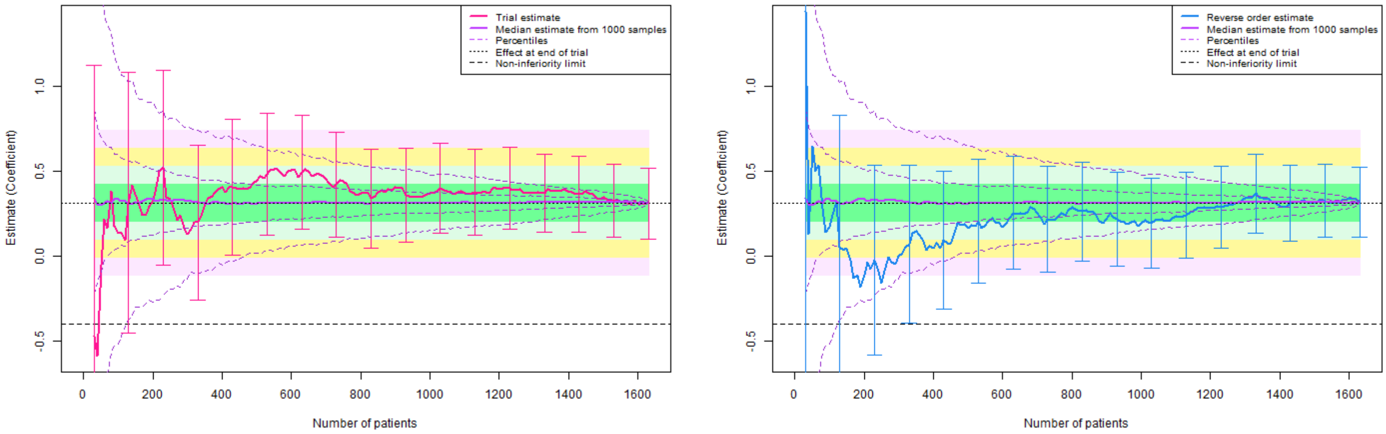


Figure C.62: Comparison of three SSR designs for the Flu BN1 trial data

The investigation of the estimate through the Flu vaccine trial for the B1 strain can be seen in Figure C.63. Similarly to the other two strains, the original sequential order starts by under-estimating the treatment effect, whereas the reverse order starts with an over-estimate. However, all estimates (both original order and reverse) are largely contained within the ±4*SE limits from 50 patients onwards, with 2 small dips outside this boundary in the reverse order estimate. The original order first reaches the inner ±1*SE boundaries marginally earlier than the reverse order (60 compared to 80 patients), and maintains its position earlier (from patient 770, 47% through the trial, compared to 1090 patients, 67% through the trial).



(a)

(b)

Figure C.63: Stability of the estimate in the Flu B1 vaccine trial. A comparison of sequential order, reverse order, and the median of 1000 random orders

C.22 Additional summary results

n*			Not significant (N=6)	Significant (N=15)	All (N=21)
			Median (LQ, UQ)	Median (LQ, UQ)	Median (LQ, UQ)
nmax=2*n					
Trend	25%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 118.4)	100.0 (100.0, 100.0)
		CT	44.9 (26.4, 100.0)	83.3 (49.4, 100.0)	62.5 (46.8, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	59.1 (52.2, 69.8)	100.0 (71.8, 100.0)	90.0 (63.4, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	94.0 (78.1, 131.9)	100.0 (96.6, 100.0)	100.0 (95.2, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Hypothesised	25%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	90.3 (72.9, 142.3)	100.0 (67.2, 100.0)	100.0 (67.3, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	PZ	100.0 (100.0, 107.4)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	133.6 (100.7, 200.0)	100.0 (76.5, 100.0)	100.0 (90.0, 105.4)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	PZ	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	109.1 (90.2, 115.3)	100.0 (96.6, 100.0)	100.0 (96.6, 100.0)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
80% limit	25%	PZ	100.0 (100.0, 134.6)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	65.9 (27.9, 200.0)	100.0 (62.4, 102.3)	100.0 (43.6, 102.3)
		SW	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	PZ	100.0 (100.0, 142.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	119.1 (63.4, 200.0)	100.0 (90.0, 100.0)	100.0 (77.9, 100.0)
		SW	100.0 (100.0, 166.7)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	PZ	100.0 (100.0, 142.4)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	94.0 (83.1, 163.5)	100.0 (96.5, 100.0)	99.0 (95.2, 100.0)
		SW	100.0 (100.0, 133.3)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
90% limit	25%	PZ	129.7 (100.0, 164.9)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	120.2 (66.9, 200.0)	100.0 (52.0, 100.0)	100.0 (66.9, 109.2)
		SW	100.0 (100.0, 133.7)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	PZ	107.8 (100.0, 117.6)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	113.1 (63.4, 199.6)	100.0 (85.6, 100.0)	100.0 (85.1, 100.0)
		SW	100.0 (100.0, 133.4)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	PZ	112.7 (100.0, 151.5)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
		CT	100.8 (90.2, 151.0)	100.0 (96.5, 100.0)	100.0 (96.5, 100.0)
		SW	100.0 (100.0, 166.7)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)

Table C.25: Percentage of n required for the new sample size at three interim time points, implementing three SSR designs using four future treatment effect assumptions. 100% indicates no change in sample size, >100% indicates an increase, up to 200% of the original sample size

Original order	±1*SE				±2*SE				±3*SE				±4*SE			
	first	(%)	last	(%)	first	(%)	last	(%)	first	(%)	last	(%)	first	(%)	last	(%)
FAST INdiCATE	50	(17%)	50	(17%)	20	(7%)	40	(14%)	20	(7%)	20	(7%)	20	(7%)	20	(7%)
Acupuncture	50	(21%)	110	(46%)	40	(17%)	40	(17%)	40	(17%)	40	(17%)	40	(17%)	40	(17%)
SELF	20	(23%)	20	(23%)	20	(23%)	20	(23%)	20	(23%)	20	(23%)	20	(23%)	20	(23%)
CASPER MINUS	50	(7%)	360	(51%)	30	(4%)	90	(13%)	30	(4%)	80	(11%)	30	(4%)	30	(4%)
CASPER	100	(14%)	120	(17%)	30	(4%)	80	(11%)	30	(4%)	80	(11%)	30	(4%)	80	(11%)
CASPER PLUS	180	(37%)	300	(62%)	130	(27%)	170	(35%)	120	(25%)	120	(25%)	110	(23%)	110	(23%)
IMPROVE	160	(26%)	460	(75%)	50	(8%)	130	(21%)	50	(8%)	130	(21%)	50	(8%)	110	(18%)
Corn plasters	40	(20%)	150	(74%)	40	(20%)	40	(20%)	20	(10%)	20	(10%)	20	(10%)	20	(10%)
AMAZE	180	(52%)	230	(66%)	40	(11%)	110	(32%)	40	(11%)	40	(11%)	40	(11%)	40	(11%)
3MG	90	(8%)	660	(61%)	80	(7%)	200	(18%)	80	(7%)	80	(7%)	40	(4%)	80	(7%)
RATPAC	40	(2%)	1910	(85%)	40	(2%)	1520	(68%)	40	(2%)	950	(42%)	40	(2%)	200	(9%)
Nasal sprays	70	(25%)	70	(25%)	10	(4%)	60	(22%)	10	(4%)	40	(14%)	10	(4%)	40	(14%)
Mencevax (A)	80	(31%)	80	(31%)	80	(31%)	80	(31%)	80	(31%)	80	(31%)	80	(31%)	80	(31%)
Mencevax (C)	40	(15%)	150	(57%)	30	(11%)	60	(23%)	30	(11%)	30	(11%)	30	(11%)	30	(11%)

Table C.26: Table to show the first time the original sequential order estimate enters each boundary and the last time (i.e. the estimate remains within this boundary for the remainder of the trial) split by outcomes type

Reverse order	±1*SE				±2*SE				±3*SE				±4*SE			
	first	(%)	last	(%)	first	(%)	last	(%)	first	(%)	last	(%)	first	(%)	last	(%)
FAST INdiCATE	20	(7%)	120	(42%)	20	(7%)	50	(17%)	20	(7%)	20	(7%)	20	(7%)	20	(7%)
Acupuncture	50	(21%)	120	(50%)	40	(17%)	40	(17%)	20	(8%)	20	(8%)	20	(8%)	20	(8%)
SELF	20	(23%)	20	(23%)	20	(23%)	20	(23%)	20	(23%)	20	(23%)	20	(23%)	20	(23%)
CASPER MINUS	70	(10%)	280	(40%)	60	(9%)	60	(9%)	60	(9%)	60	(9%)	50	(7%)	50	(7%)
CASPER	30	(4%)	170	(24%)	30	(4%)	150	(21%)	30	(4%)	70	(10%)	30	(4%)	30	(4%)
CASPER PLUS	240	(49%)	410	(85%)	110	(23%)	110	(23%)	100	(21%)	100	(21%)	40	(8%)	90	(19%)
IMPROVE	40	(7%)	350	(57%)	40	(7%)	280	(46%)	40	(7%)	70	(11%)	40	(7%)	40	(7%)
Corn plasters	70	(35%)	70	(35%)	70	(35%)	70	(35%)	40	(20%)	60	(30%)	30	(15%)	30	(15%)
AMAZE	270	(78%)	270	(78%)	50	(14%)	200	(57%)	40	(11%)	150	(43%)	40	(11%)	40	(11%)
3MG	50	(5%)	580	(54%)	50	(5%)	330	(30%)	50	(5%)	50	(5%)	40	(4%)	40	(4%)
RATPAC	60	(3%)	1840	(82%)	60	(3%)	1380	(62%)	50	(2%)	1170	(52%)	50	(2%)	930	(41%)
Nasal sprays	20	(7%)	130	(47%)	20	(7%)	20	(7%)	20	(7%)	20	(7%)	20	(7%)	20	(7%)
Mencevax (A)	90	(34%)	200	(77%)	90	(34%)	90	(34%)	90	(34%)	90	(34%)	90	(34%)	90	(34%)
Mencevax (C)	50	(19%)	130	(50%)	50	(19%)	130	(50%)	30	(11%)	50	(19%)	30	(11%)	30	(11%)

Table C.27: Table to show the first time the reverse order estimate enters each boundary and the last time (i.e. the estimate remains within this boundary for the remainder of the trial) split by outcome type

D δ investigation results

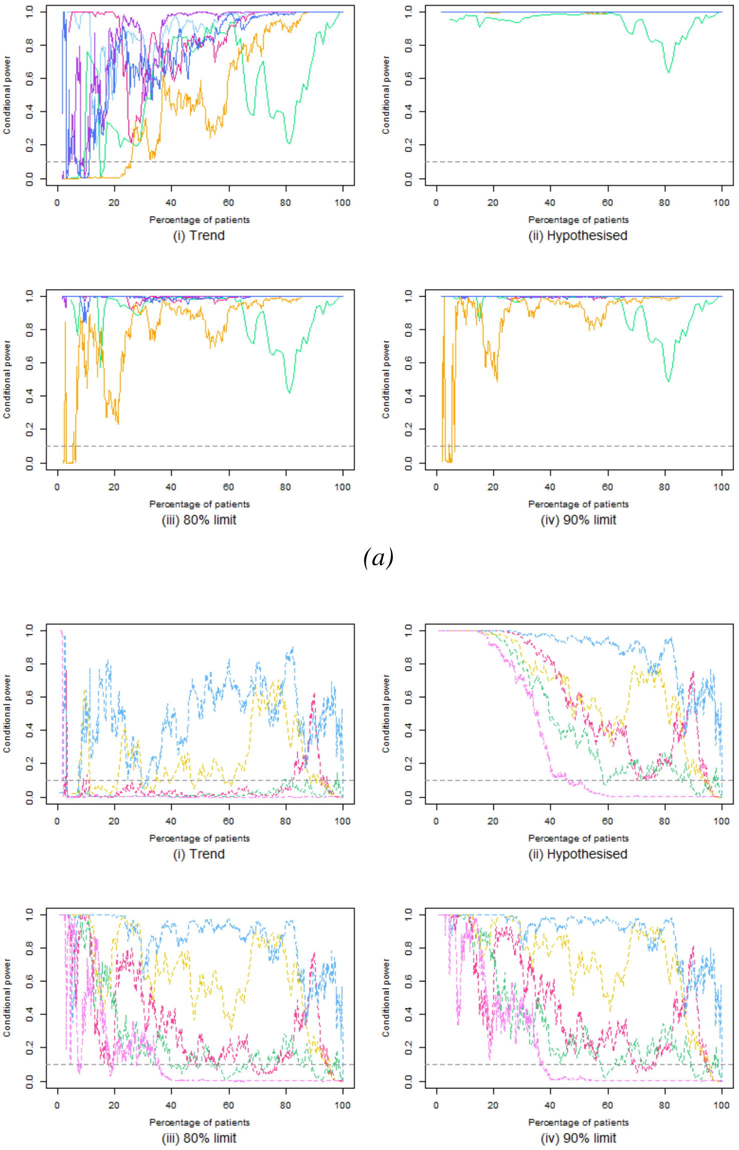
D.1 Observed effect = half planned

n*		Continuous (N=6)	Binary (N=5)	All (N=11)
Trend				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	102.3 (100.0, 137.7)	100.0 (100.0, 100.0)	100.0 (100.0, 137.7)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 124.3)	100.0 (100.0, 100.0)
Combination test	25%	92.8 (63.8, 118.4)	43.6 (26.4, 49.4)	63.8 (43.6, 118.4)
	50%	99.9 (99.7, 128.0)	71.6 (54.0, 71.8)	99.7 (61.1, 128.0)
	75%	99.5 (83.3, 100.0)	93.1 (82.9, 95.2)	95.2 (82.9, 100.0)
Hypothesised				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 131.7)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 102.0)	100.0 (100.0, 100.0)
Combination test	50%	75.3 (61.1, 100.0)	70.3 (55.8, 71.1)	70.3 (55.8, 100.0)
	75%	88.2 (72.2, 100.0)	124.6 (76.9, 132.7)	100.0 (72.2, 124.6)
	25%	99.5 (83.3, 100.0)	126.2 (95.2, 149.9)	100.0 (94.0, 143.0)
80% limit				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	91.8 (59.0, 103.1)	73.3 (63.2, 100.7)	83.3 (59.0, 103.1)
	50%	95.8 (78.1, 100.0)	71.8 (71.6, 77.5)	78.1 (71.6, 100.0)
	75%	99.5 (83.3, 100.0)	95.2 (94.0, 105.9)	99.0 (83.3, 105.9)
90% limit				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 126.3)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	85.8 (53.1, 100.0)	61.5 (59.9, 91.7)	83.3 (53.1, 100.0)
	50%	91.3 (74.3, 100.0)	110.4 (76.4, 129.7)	96.5 (74.3, 110.4)
	75%	99.5 (83.3, 100.0)	95.2 (93.7, 99.0)	99.0 (83.3, 100.0)

Table D.1: New total sample size (n^*) as a percentage of original planned sample size (n) comparison at three interim time points and four treatment effect assumptions when the observed treatment effect is half of the planned effect

Zones		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	Futility	4.5 (0, 9)	87.9 (37, 95)	9.1 (4, 88)
	Unfav.	7.3 (5, 19)	9.4 (2, 26)	8.5 (2, 25)
	Prom.	26.0 (15, 36)	2.6 (0, 20)	19.7 (3, 36)
	Fav.	59.6 (33, 76)	0.0 (0, 1)	30.9 (0, 60)
Combination test	Decrease	46.2 (14, 64)	88.2 (80, 91)	64.3 (31, 88)
	Remain	20.6 (5, 45)	3.6 (0, 7)	6.5 (3, 25)
	Increase	20.5 (0, 42)	4.1 (2, 6)	6.3 (0, 42)
Hypothesised				
Promising zone	Futility	0.0 (0, 0)	6.2 (6, 13)	0.0 (0, 6)
	Unfav.	0.0 (0, 0)	11.9 (10, 27)	0.0 (0, 12)
	Prom.	0.0 (0, 0)	25.6 (13, 38)	3.9 (0, 26)
	Fav.	100.0 (100, 100)	29.5 (29, 30)	99.7 (30, 100)
Combination test	Decrease	53.4 (14, 82)	37.9 (21, 80)	37.9 (20, 82)
	Remain	20.5 (10, 65)	2.4 (0, 4)	10.4 (2, 24)
	Increase	0.0 (0, 0)	56.8 (16, 74)	0.0 (0, 74)
80% limit				
Promising zone	Futility	0.0 (0, 0)	16.9 (5, 33)	0.1 (0, 17)
	Unfav.	0.0 (0, 0)	20.1 (13, 49)	1.1 (0, 20)
	Prom.	0.0 (0, 21)	26.3 (9, 38)	8.6 (0, 26)
	Fav.	99.7 (84, 100)	7.0 (7, 29)	69.1 (7, 100)
Combination test	Decrease	40.2 (27, 82)	71.2 (40, 80)	47.1 (27, 82)
	Remain	20.6 (5, 65)	2.1 (0, 4)	5.2 (2, 25)
	Increase	1.8 (0, 26)	23.5 (16, 55)	16.3 (0, 55)
90% limit				
Promising zone	Futility	0.0 (0, 0)	11.9 (5, 16)	0.1 (0, 12)
	Unfav.	0.0 (0, 0)	13.6 (8, 39)	0.4 (0, 14)
	Prom.	0.0 (0, 8)	23.0 (10, 37)	9.0 (0, 23)
	Fav.	99.7 (91, 100)	14.5 (14, 49)	88.9 (15, 100)
Combination test	Decrease	54.4 (33, 82)	49.7 (29, 80)	50.6 (29, 82)
	Remain	20.8 (5, 65)	1.5 (1, 4)	5.2 (2, 25)
	Increase	0.0 (0, 15)	45.0 (16, 66)	14.7 (0, 51)

Table D.2: Percentage of trial duration spent in each zone at three interim time points and four treatment effect assumptions when the observed treatment effect is half the effect planned



(a)

(b)

Figure D.1: Conditional power when $\hat{\delta}_{obs} = \frac{1}{2}\delta_{plan}$ for (a) Continuous trials (b) Binary trials

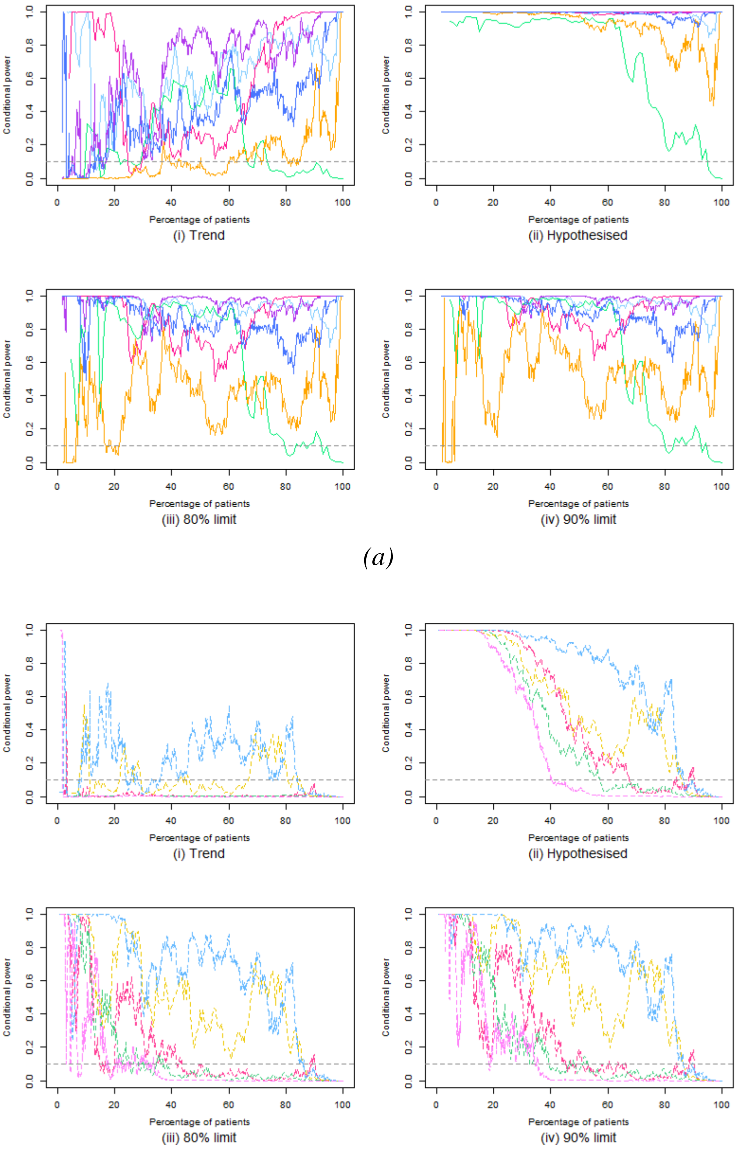
D.2 Observed effect = one third planned

n*		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	25%	100.0 (100.0, 146.1)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 150.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	110.9 (100.0, 129.4)	100.0 (100.0, 100.0)	100.0 (100.0, 121.8)
Combination test	25%	70.1 (56.7, 100.0)	43.6 (26.4, 49.4)	56.7 (43.6, 100.0)
	50%	100.0 (76.5, 150.0)	55.4 (52.2, 71.6)	75.9 (55.4, 100.0)
	75%	99.5 (97.7, 117.0)	93.1 (78.1, 95.2)	97.7 (93.1, 117.0)
Hypothesised				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	104.0 (100.0, 150.0)	100.0 (100.0, 104.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 140.6)	100.0 (100.0, 133.7)
Combination test	50%	78.8 (67.4, 100.0)	72.8 (58.6, 74.9)	74.4 (58.6, 100.0)
	75%	94.8 (86.1, 100.0)	139.2 (83.9, 145.5)	100.0 (83.9, 139.2)
	25%	100.0 (92.4, 104.9)	95.2 (93.1, 105.4)	100.0 (92.4, 105.4)
80% limit				
Promising zone	25%	100.0 (100.0, 150.1)	100.0 (100.0, 126.4)	100.0 (100.0, 150.0)
	50%	100.0 (100.0, 145.2)	100.0 (100.0, 133.2)	100.0 (100.0, 145.2)
	75%	100.0 (100.0, 103.5)	100.0 (100.0, 117.8)	100.0 (100.0, 117.8)
Combination test	50%	104.3 (73.6, 150.0)	49.4 (43.6, 55.4)	73.6 (49.4, 125.4)
	75%	121.1 (100.0, 146.5)	71.8 (71.6, 80.2)	100.0 (71.6, 146.5)
	25%	105.1 (102.8, 136.9)	95.2 (93.1, 95.6)	102.8 (95.2, 130.7)
90% limit				
Promising zone	25%	100.0 (100.0, 108.8)	100.0 (100.0, 100.0)	100.0 (100.0, 108.8)
	50%	100.0 (100.0, 117.5)	100.0 (100.0, 110.6)	100.0 (100.0, 117.5)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 107.4)	100.0 (100.0, 107.4)
Combination test	50%	100.8 (64.2, 120.5)	66.3 (66.2, 103.6)	83.3 (64.2, 120.5)
	75%	111.7 (100.0, 127.9)	71.8 (71.6, 84.0)	100.0 (71.6, 127.9)
	25%	102.3 (99.7, 130.7)	95.2 (93.1, 105.4)	100.8 (95.2, 119.3)

Table D.3: New total sample size (n^*) as a percentage of original planned sample size (n) comparison at three interim time points and four treatment effect assumptions when the observed treatment effect equals one third of the planned effect

Zones		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	Futility	10.1 (5, 51)	97.1 (71, 99)	50.6 (9, 97)
	Unfav.	29.3 (15, 38)	0.0 (0, 28)	24.7 (0, 38)
	Prom.	32.7 (20, 40)	0.2 (0, 1)	10.6 (0, 33)
	Fav.	17.8 (1, 37)	0.0 (0, 1)	0.6 (0, 28)
Combination test	Decrease	27.0 (11, 74)	88.2 (80, 95)	73.5 (21, 88)
	Remain	14.8 (11, 44)	8.8 (4, 14)	13.0 (4, 19)
	Increase	29.9 (11, 73)	0.0 (0, 0)	10.5 (0, 47)
Hypothesised				
Promising zone	Futility	0.0 (0, 0)	30.8 (16, 44)	6.5 (0, 31)
	Unfav.	0.0 (0, 0)	18.3 (6, 21)	3.7 (0, 18)
	Prom.	0.0 (0, 14)	19.4 (10, 32)	9.7 (0, 19)
	Fav.	99.7 (86, 100)	26.9 (26, 29)	68.8 (27, 100)
Combination test	Decrease	42.0 (14, 82)	60.8 (42, 80)	50.6 (21, 82)
	Remain	9.1 (1, 56)	3.6 (1, 9)	3.6 (1, 19)
	Increase	9.9 (0, 46)	33.9 (12, 46)	12.1 (0, 46)
80% limit				
Promising zone	Futility	0.0 (0, 8)	61.3 (15, 72)	14.3 (0, 61)
	Unfav.	0.0 (0, 21)	15.3 (13, 21)	13.3 (0, 21)
	Prom.	23.5 (2, 45)	12.9 (4, 38)	15.6 (3, 45)
	Fav.	61.7 (55, 98)	5.6 (5, 17)	22.8 (5, 68)
Combination test	Decrease	27.0 (14, 47)	80.1 (77, 91)	47.3 (18, 81)
	Remain	9.9 (3, 33)	3.8 (1, 9)	3.8 (1, 19)
	Increase	49.6 (32, 76)	5.4 (3, 11)	32.3 (3, 76)
90% limit				
Promising zone	Futility	0.0 (0, 4)	49.5 (15, 60)	11.7 (0, 50)
	Unfav.	0.0 (0, 21)	19.0 (15, 21)	15.0 (0, 21)
	Prom.	10.8 (1, 36)	16.1 (6, 43)	12.5 (3, 43)
	Fav.	76.1 (64, 99)	11.5 (10, 21)	39.2 (12, 87)
Combination test	Decrease	23.8 (13, 60)	80.1 (61, 83)	59.9 (14, 81)
	Remain	10.2 (2, 37)	3.6 (1, 9)	3.9 (2, 19)
	Increase	41.5 (26, 84)	11.9 (12, 28)	27.7 (12, 74)

Table D.4: Percentage of trial duration spent in each zone at three interim time points and four treatment effect assumptions when the observed treatment effect is one third of the planned effect



(a)

(b)

Figure D.2: Conditional power when $\hat{\delta}_{obs} = \frac{1}{3}\delta_{plan}$ for (a) Continuous trials (b) Binary trials

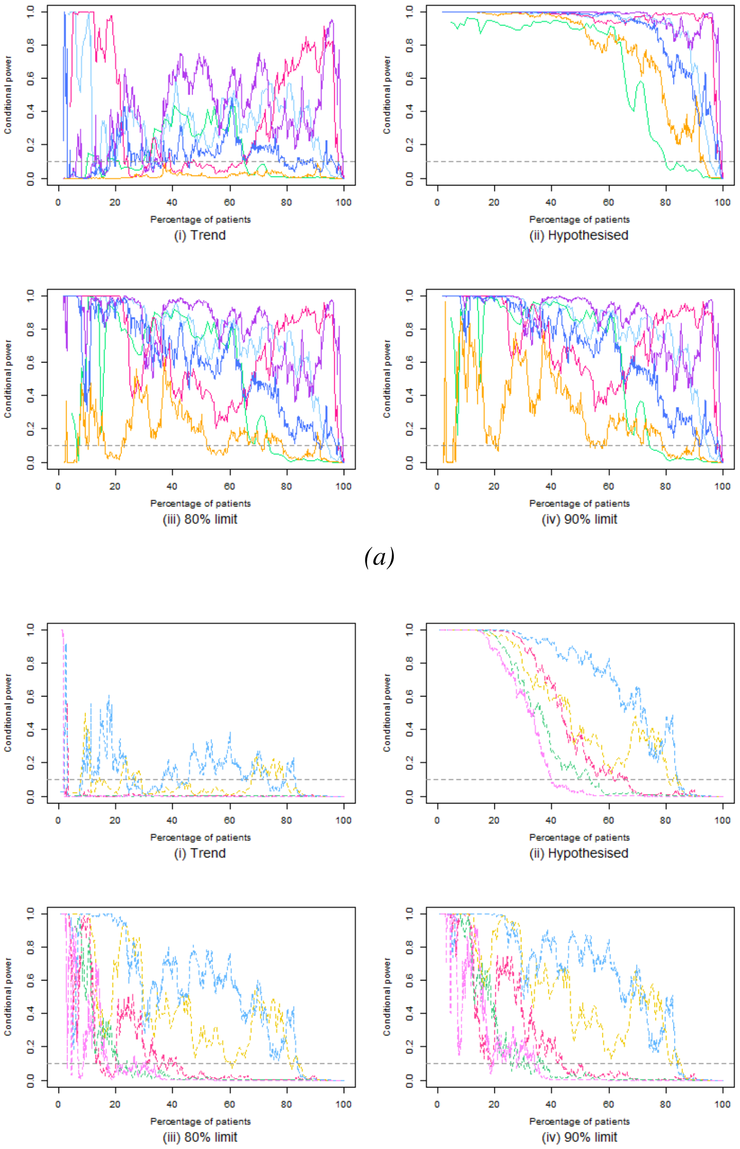
D.3 Observed effect = one quarter planned

n*		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	25%	100.0 (100.0, 150.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 150.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	25%	66.6 (46.8, 100.0)	36.6 (26.4, 43.6)	46.8 (29.5, 83.3)
	50%	95.1 (78.3, 100.0)	55.4 (52.2, 71.6)	76.5 (55.4, 100.0)
	75%	99.5 (97.7, 150.0)	93.1 (78.1, 95.2)	97.7 (83.3, 150.0)
Hypothesised				
Promising zone	25%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
	50%	100.0 (100.0, 100.0)	100.0 (100.0, 112.5)	100.0 (100.0, 100.0)
	75%	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)	100.0 (100.0, 100.0)
Combination test	50%	80.6 (70.5, 100.0)	74.3 (60.1, 76.8)	76.8 (60.1, 100.0)
	75%	98.9 (94.4, 100.4)	146.8 (87.5, 150.0)	100.0 (87.5, 146.8)
	25%	104.5 (103.8, 120.8)	95.2 (93.1, 100.1)	103.8 (95.2, 120.8)
80% limit				
Promising zone	25%	100.0 (100.0, 111.6)	100.0 (100.0, 100.0)	100.0 (100.0, 111.6)
	50%	105.8 (100.0, 111.8)	100.0 (100.0, 100.0)	100.0 (100.0, 111.8)
	75%	103.5 (100.0, 134.3)	100.0 (100.0, 100.0)	100.0 (100.0, 134.3)
Combination test	50%	83.5 (66.7, 150.0)	43.6 (26.4, 49.4)	66.7 (43.6, 89.1)
	75%	139.9 (106.1, 150.0)	71.6 (66.6, 71.8)	106.1 (71.6, 150.0)
	25%	138.1 (125.4, 150.0)	93.1 (78.1, 95.2)	125.4 (93.1, 147.5)
90% limit				
Promising zone	25%	100.0 (100.0, 141.8)	100.0 (100.0, 104.0)	100.0 (100.0, 141.8)
	50%	105.1 (100.0, 150.1)	100.0 (100.0, 129.3)	100.0 (100.0, 150.0)
	75%	110.0 (100.0, 120.4)	100.0 (100.0, 100.0)	100.0 (100.0, 120.4)
Combination test	50%	108.1 (71.9, 140.7)	67.3 (49.4, 70.3)	71.9 (60.1, 133.0)
	75%	134.5 (104.3, 150.0)	71.8 (71.6, 85.7)	104.3 (71.8, 147.4)
	25%	130.5 (124.3, 150.0)	93.1 (78.1, 95.2)	124.3 (93.1, 133.9)

Table D.5: New total sample size (n^*) as a percentage of original planned sample size (n) comparison at three interim time points and four treatment effect assumptions when the observed treatment effect equals one quarter of the planned effect

Zones		**Continuous (N=6)**	**Binary (N=5)**	**All (N=11)**
Trend				
Promising zone	Futility	38.3 (18, 58)	97.1 (76, 99)	58.4 (38, 99)
	Unfav.	37.3 (19, 55)	0.0 (0, 23)	23.3 (0, 49)
	Prom.	17.8 (2, 30)	0.2 (0, 1)	2.3 (0, 29)
	Fav.	1.8 (0, 6)	0.0 (0, 1)	0.6 (0, 3)
Combination test	Decrease	32.3 (30, 74)	88.2 (80, 95)	73.5 (31, 88)
	Remain	23.3 (8, 48)	8.8 (4, 17)	16.6 (4, 24)
	Increase	25.9 (1, 53)	0.0 (0, 0)	0.9 (0, 43)
Hypothesised				
Promising zone	Futility	3.3 (1, 7)	35.4 (19, 48)	16.0 (1, 35)
	Unfav.	2.8 (1, 8)	15.6 (8, 20)	7.8 (2, 16)
	Prom.	7.1 (2, 16)	16.4 (10, 24)	10.0 (5, 19)
	Fav.	83.1 (68, 96)	26.5 (26, 28)	62.8 (27, 88)
Combination test	Decrease	39.3 (14, 57)	73.2 (53, 80)	53.1 (21, 76)
	Remain	4.6 (2, 37)	3.6 (1, 9)	4.0 (1, 19)
	Increase	41.2 (28, 52)	21.5 (13, 35)	35.1 (19, 52)
80% limit				
Promising zone	Futility	5.1 (3, 31)	68.5 (20, 81)	19.7 (4, 69)
	Unfav.	15.9 (4, 22)	15.3 (10, 19)	15.6 (8, 22)
	Prom.	36.0 (31, 46)	7.9 (3, 20)	31.2 (6, 46)
	Fav.	28.9 (23, 52)	4.6 (4, 17)	20.9 (4, 30)
Combination test	Decrease	27.0 (14, 52)	81.1 (80, 95)	51.5 (15, 81)
	Remain	8.1 (2, 13)	3.6 (2, 9)	5.7 (2, 13)
	Increase	62.8 (31, 80)	2.0 (0, 7)	30.5 (2, 78)
90% limit				
Promising zone	Futility	4.5 (3, 26)	57.9 (17, 68)	17.1 (3, 58)
	Unfav.	7.6 (3, 19)	17.9 (12, 20)	11.5 (5, 20)
	Prom.	24.6 (23, 36)	14.4 (8, 38)	24.0 (9, 38)
	Fav.	50.1 (33, 67)	10.3 (8, 20)	26.3 (8, 58)
Combination test	Decrease	25.5 (14, 33)	80.1 (76, 85)	33.3 (20, 80)
	Remain	7.2 (2, 14)	3.6 (1, 9)	6.5 (2, 14)
	Increase	63.3 (51, 79)	9.4 (7, 12)	51.3 (9, 74)

Table D.6: Percentage of trial duration spent in each zone at three interim time points and four treatment effect assumptions when the observed treatment effect is one quarter of the planned effect



(a)

(b)

Figure D.3: Conditional power when $\hat{\delta}_{obs} = \frac{1}{4}\delta_{plan}$ *for (a) Continuous trials (b) Binary trials*

E | Simulation results

E.1 Planned effect = 0.2

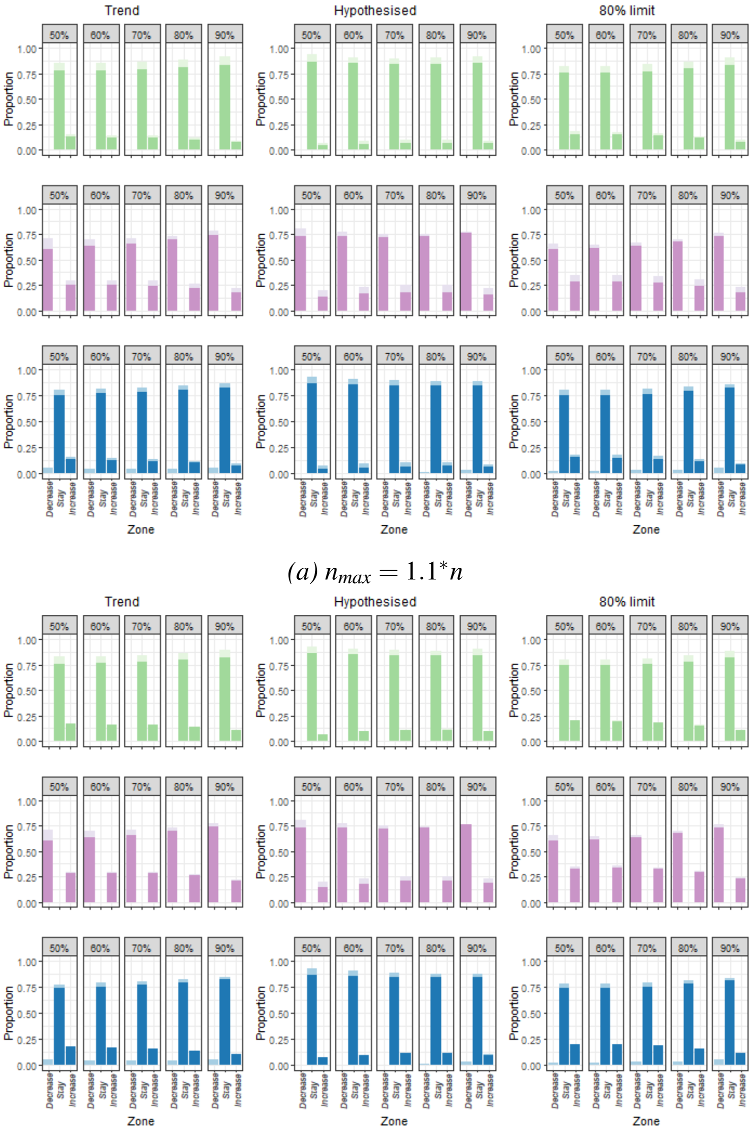
$\hat{\delta}_{obs} = \delta_{plan} = 0.2$	**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**	**100%**
Mean $\hat{d}$	3.9978	3.9921	3.9897	3.9892	3.9879	3.9879	3.9896	3.9911	3.9900	3.9891	3.9891
Mean SD	19.9924	19.9938	19.9939	19.995	19.9955	19.9958	19.996	19.9967	19.9978	19.9978	19.9976
Mean δ	0.2002	0.1998	0.1997	0.1997	0.1996	0.1996	0.1996	0.1997	0.1996	0.1996	0.1996
Mean $(d-\hat{d})$	-0.0022	-0.0079	-0.0103	-0.0108	-0.0121	-0.0121	-0.0104	-0.0089	-0.0100	-0.0109	-0.0109
$\hat{\delta}_{obs} = \frac{2}{3}\delta_{plan}$											
Mean $\hat{d}$	2.6645	2.6588	2.6563	2.6559	2.6546	2.6545	2.6563	2.6578	2.6566	2.6558	2.6557
Mean SD	19.9924	19.9938	19.9939	19.995	19.9955	19.9958	19.996	19.9967	19.9978	19.9978	19.9976
Mean δ	0.1334	0.1331	0.1330	0.1329	0.1328	0.1328	0.1329	0.1330	0.1329	0.1329	0.1329
Mean $(d-\hat{d})$	-1.3355	-1.3412	-1.3437	-1.3441	-1.3454	-1.3455	-1.3437	-1.3422	-1.3434	-1.3442	-1.3443
$\hat{\delta}_{obs} = \frac{1}{3}\delta_{plan}$											
Mean $\hat{d}$	1.3311	1.3255	1.323	1.3226	1.3212	1.3212	1.3229	1.3244	1.3233	1.3225	1.3224
Mean SD	19.9924	19.9938	19.9939	19.995	19.9955	19.9958	19.996	19.9967	19.9978	19.9978	19.9976
Mean δ	0.0666	0.0663	0.0662	0.0662	0.0661	0.0661	0.0662	0.0663	0.0662	0.0662	0.0662
Mean $(d-\hat{d})$	-2.6689	-2.6745	-2.6770	-2.6774	-2.6788	-2.6788	-2.6771	-2.6756	-2.6767	-2.6775	-2.6776
$\hat{\delta}_{obs} = 0$											
Mean $\hat{d}$	-0.0022	-0.0079	-0.0103	-0.0108	-0.0121	-0.0121	-0.0104	-0.0089	-0.0100	-0.0109	-0.0109
Mean SD	19.9924	19.9938	19.9939	19.995	19.9955	19.9958	19.996	19.9967	19.9978	19.9978	19.9976
Mean δ	-0.0001	-0.0004	-0.0005	-0.0005	-0.0006	-0.0006	-0.0005	-0.0005	-0.0005	-0.0005	-0.0006
Mean $(d-\hat{d})$	-4.0022	-4.0079	-4.0103	-4.0108	-4.0121	-4.0121	-4.0104	-4.0089	-4.0100	-4.0109	-4.0109

Table E.1: Mean difference, SD, treatment effect and difference from the true population value for values between 50% and 100% of the originally planned n=1052 when $\delta = 0.2$

		50%	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**
δ=δ_{plan}	Trend	0.853	0.860	0.866	0.871	0.876	0.88	0.884	0.887	0.89
	Hypothesised	0.974	0.970	0.965	0.960	0.953	0.946	0.938	0.929	0.919
	80% limit	0.889	0.885	0.883	0.882	0.882	0.883	0.884	0.887	0.889
δ=$\frac{2}{3}\delta_{plan}$	Trend	0.636	0.634	0.631	0.628	0.623	0.619	0.613	0.607	0.598
	Hypothesised	0.921	0.900	0.878	0.851	0.821	0.787	0.751	0.711	0.668
	80% limit	0.713	0.694	0.677	0.662	0.646	0.633	0.621	0.61	0.598
δ=$\frac{1}{3}\delta_{plan}$	Trend	0.362	0.347	0.331	0.317	0.301	0.285	0.268	0.25	0.231
	Hypothesised	0.808	0.754	0.696	0.633	0.563	0.495	0.426	0.357	0.295
	80% limit	0.462	0.426	0.393	0.363	0.333	0.306	0.281	0.256	0.233
δ=0	Trend	0.145	0.129	0.115	0.101	0.088	0.076	0.065	0.053	0.043
	Hypothesised	0.630	0.537	0.446	0.359	0.275	0.205	0.148	0.100	0.066
	80% limit	0.226	0.189	0.159	0.132	0.108	0.088	0.072	0.057	0.044

Table E.2: Mean conditional power values from 50000 repetitions for three treatment effect assumptions when δ=0.2 and n=1052

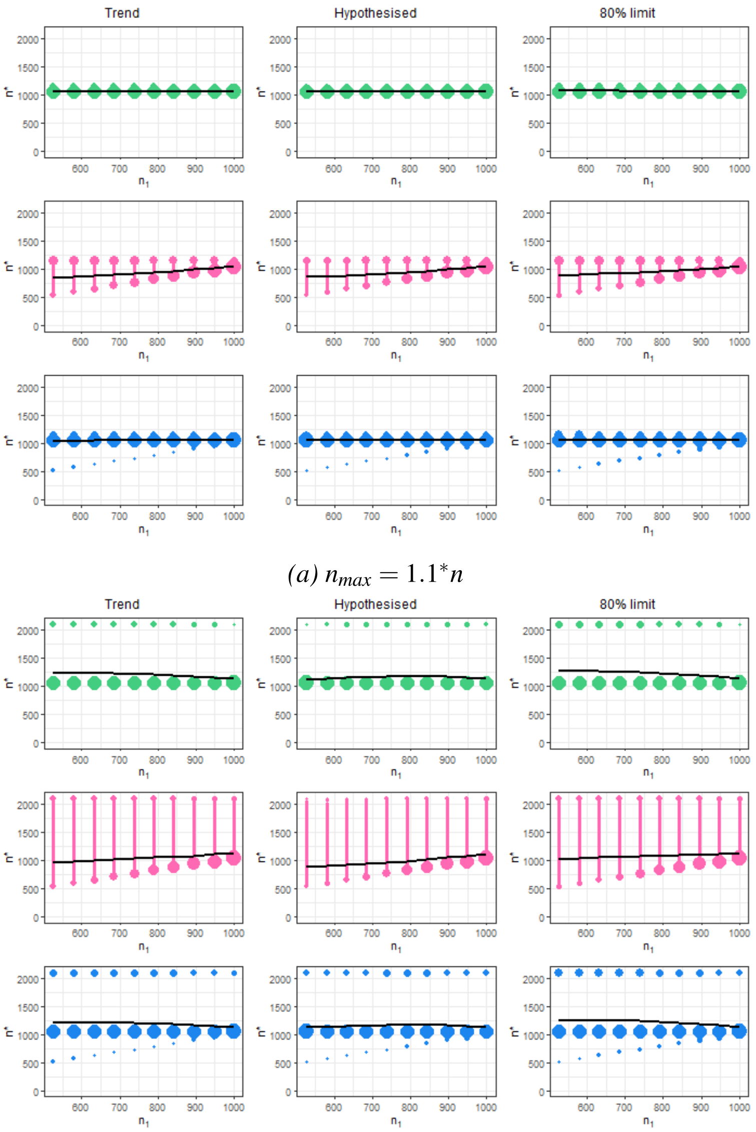
E.1.1 Observed effect=planned effect



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

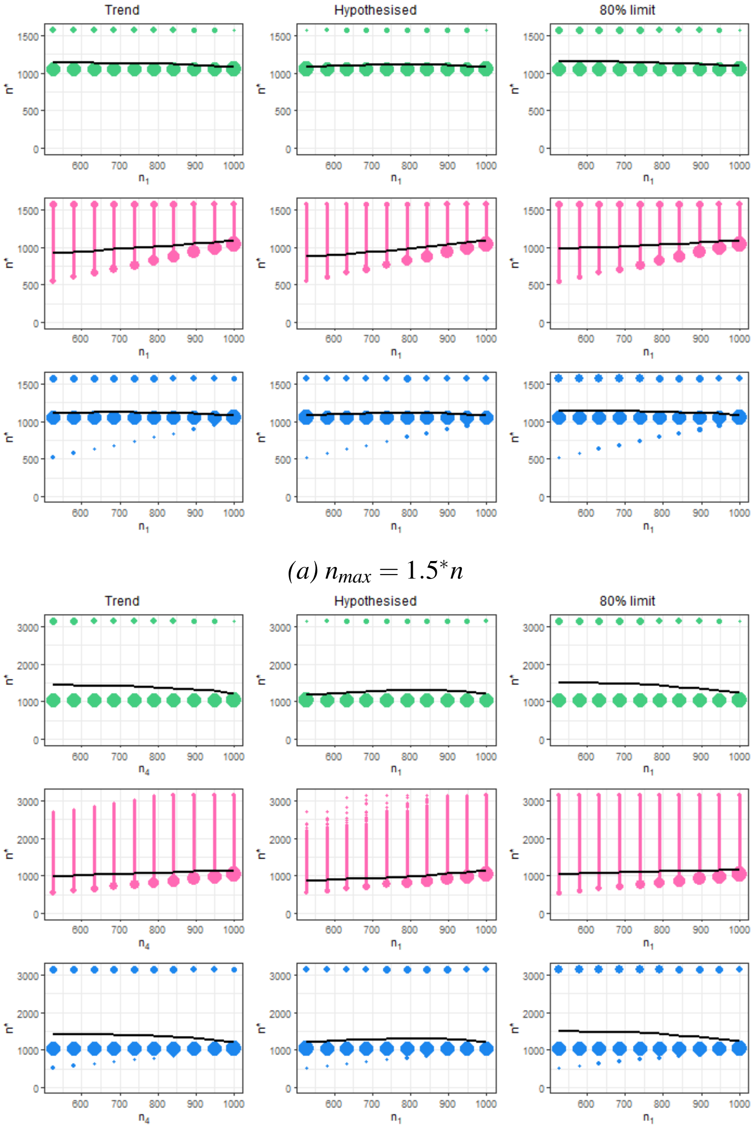
Figure E.1: Sample size zones from 50000 simulations when $\delta = \delta_{plan}$ and n=1052, for two values of n_{max}: comparing three designs and four observed treatment effects



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure E.2: Sample size from 50000 simulations when $\delta = \delta_{plan}$ and n=1052, for two values of n_{max}: comparing three designs and three treatment effect assumptions



(a) $n_{max} = 1.5^{*}n$

(b) $n_{max} = 3^{*}n$

Figure E.3: Sample size from 50000 simulations when $\delta = \delta_{plan}$ and n=1052, for two values of n_{max}: comparing three designs and three treatment effect assumptions

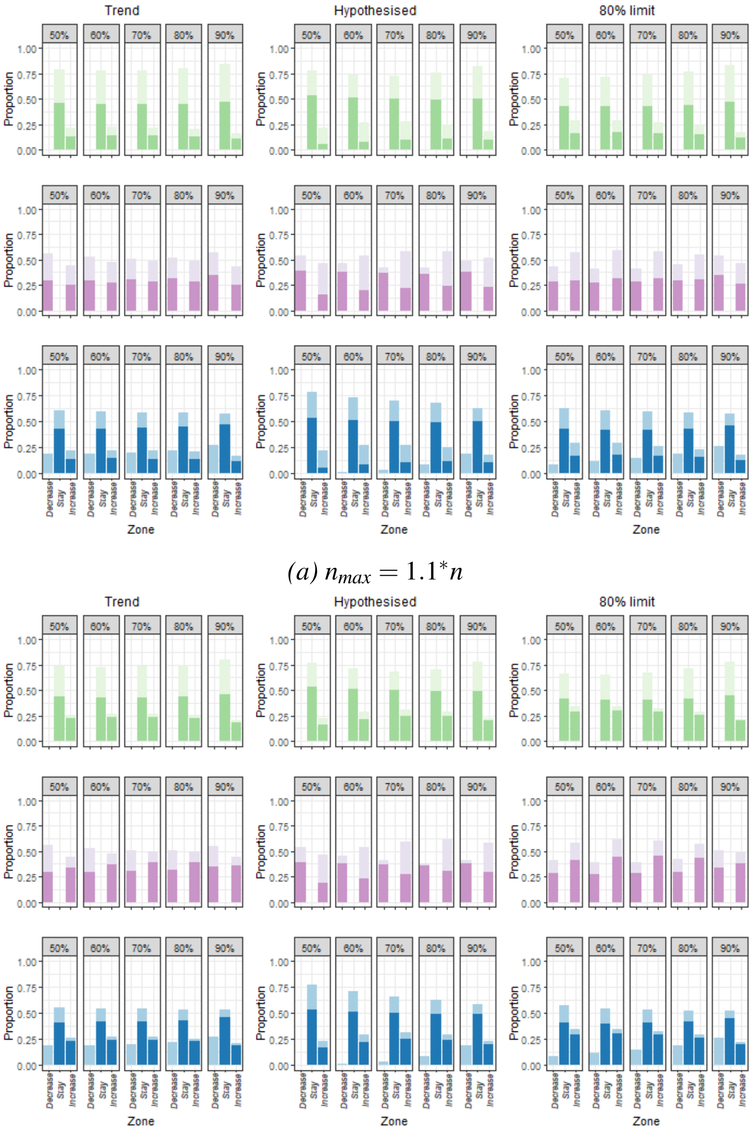
δ_{plan}=0.2		**SHORT**					**MEDIUM**				
δ=δ_{plan}											
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	1071	1092	1105	1119	1140	1110	1150	1192	1220	1185
	Power	91.37	94.24	95.59	96.63	96.68	92.63	95.12	96.27	96.96	96.73
γ_2	ASN	969	1006	1035	1068	1108	1019	1076	1135	1179	1155
	Power	88.13	91.73	93.89	95.47	95.83	90.08	93.30	94.98	96.06	95.95
γ_3	ASN	766	833	899	967	1043	838	927	1020	1093	1094
	Power	79.58	84.66	88.65	91.52	93.29	83.47	87.94	91.14	92.90	93.57
γ_4	ASN	653	737	822	910	1006	753	859	969	1054	1061
	Power	69.39	76.24	82.39	86.93	90.24	69.39	76.24	82.39	86.93	90.24
HYPOTHESISED											
γ_1	ASN	956	980	1009	1049	1106	972	1021	1088	1148	1150
	Power	91.69	93.46	94.90	95.97	96.36	91.88	93.79	95.3	96.27	96.42
γ_2	ASN	882	919	962	1016	1087	907	973	1054	1125	1133
	Power	88.52	91.10	92.99	94.62	95.62	88.99	91.79	93.79	95.14	95.74
γ_3	ASN	783	838	900	973	1058	828	917	1015	1096	1107
	Power	82.16	85.83	89.29	92.09	94.05	83.77	87.88	91.26	93.41	94.29
γ_4	ASN	706	775	851	933	1025	777	880	987	1071	1078
	Power	75.13	80.09	85.12	89.06	92.00	79.20	84.61	88.88	91.47	92.46
80% Limit											
γ_1	ASN	1127	1143	1145	1145	1154	1150	1186	1220	1240	1198
	Power	95.53	96.83	97.36	97.61	97.22	95.89	97.16	97.69	97.83	97.26
γ_2	ASN	1023	1050	1069	1089	1120	1055	1105	1157	1194	1167
	Power	92.92	94.91	95.95	96.61	96.56	93.68	95.63	96.62	97.01	96.66
γ_3	ASN	820	873	926	983	1051	874	954	1038	1106	1102
	Power	84.60	88.07	90.76	92.72	93.92	86.72	90.18	92.51	93.88	94.16
γ_4	ASN	678	752	830	914	1007	762	864	972	1055	1062
	Power	72.81	78.41	83.38	87.39	90.48	77.97	83.30	87.61	90.10	91.01

*Table E.3: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when δ=δ_{plan}, n=1052, n_{max}=2*n*

$\delta = \delta_{plan}$		**Information fraction**									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	1236	1232	1228	1224	1219	1209	1196	1181	1163	1134
	Power	93.01	93.10	93.17	93.25	93.28	93.35	93.20	93.06	92.79	92.18
Hypothesised	ASN	1125	1140	1153	1163	1168	1173	1172	1167	1155	1131
	Power	93.28	94.00	94.65	95.10	95.28	95.44	95.35	94.94	94.14	93.01
80% limit	ASN	1265	1266	1263	1257	1249	1236	1217	1197	1172	1137
	Power	94.40	94.55	94.45	94.32	94.16	94.08	93.70	93.39	92.95	92.21
Promising zone with Futility											
Trend	ASN	1210	1211	1211	1210	1207	1199	1188	1175	1158	1130
	Power	91.11	91.59	92.03	92.35	92.55	92.80	92.76	92.75	92.54	92.02
Hypothesised	ASN	1125	1140	1153	1162	1168	1172	1170	1164	1152	1128
	Power	93.28	94.00	94.66	95.10	95.28	95.44	95.34	94.92	94.11	92.97
80% limit	ASN	1257	1258	1255	1249	1242	1228	1210	1191	1167	1134
	Power	94.04	94.21	94.15	94.00	93.85	93.80	93.41	93.16	92.74	92.07

Table E.4: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions for the promising zone design with and without a futility boundary when δ=δ_{plan}, *n=1052 and* $n_{max} = 2^{*}n$

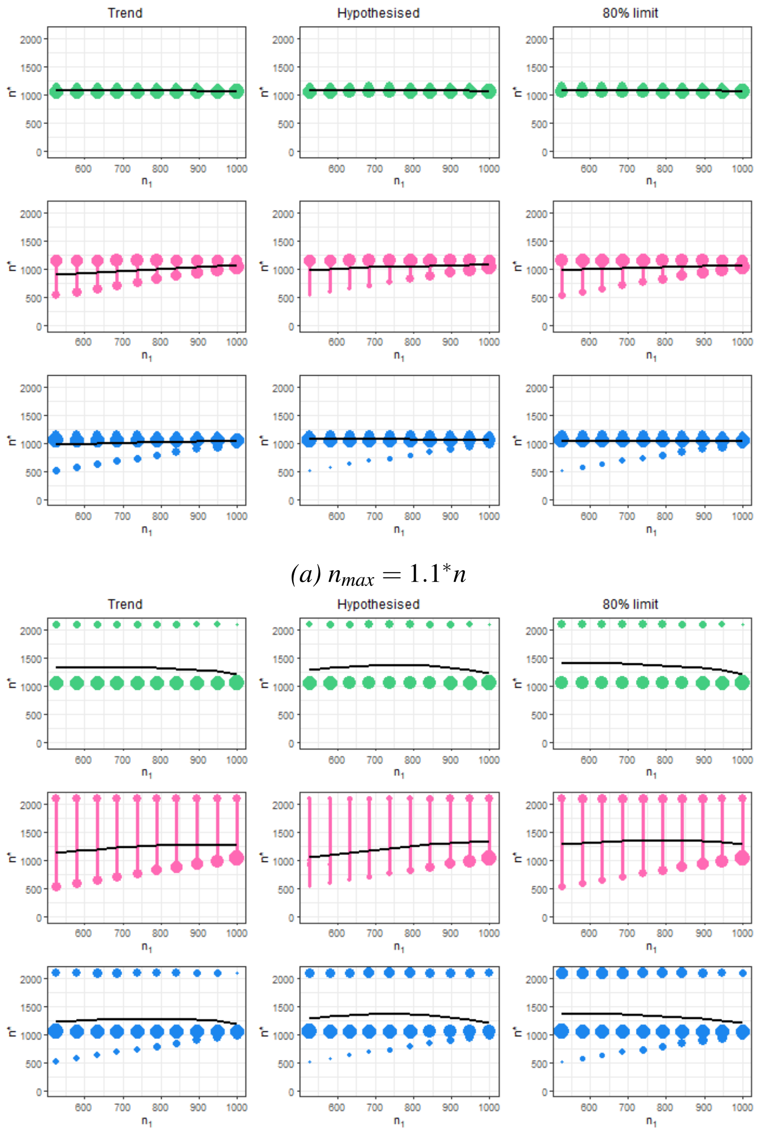
E.1.2 Observed effect = two thirds planned



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

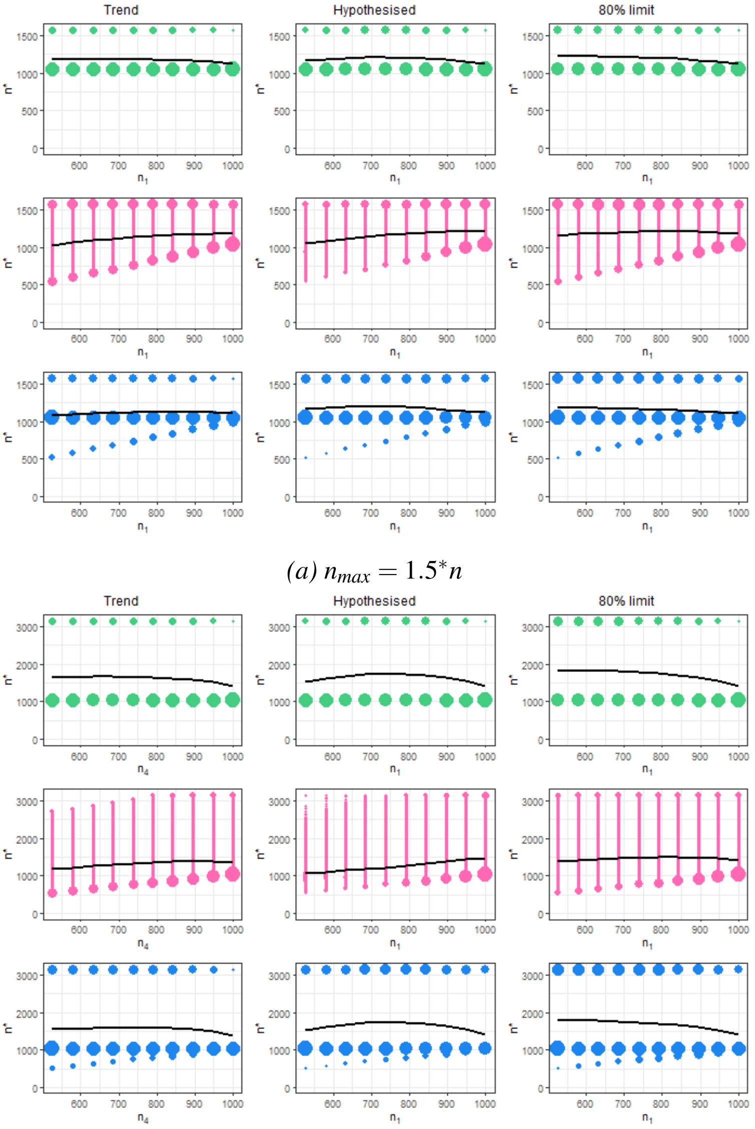
Figure E.4: Sample size zones from 50000 simulations when $\delta = \frac{2}{3}\delta_{plan}$ and n=1052, for two values of n_{max}: comparing three designs and four observed treatment effects



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure E.5: Sample size from 50000 simulations when $\delta = \frac{2}{3}\delta_{plan}$ and n=1052, for two values of n_{max}: comparing three designs and three treatment effect assumptions



(a) $n_{max} = 1.5^{*}n$

(b) $n_{max} = 3^{*}n$

Figure E.6: Sample size from 50000 simulations when $\delta = \frac{2}{3}\delta_{plan}$ and n=1052, for two values of n_{max}: comparing three designs and three treatment effect assumptions

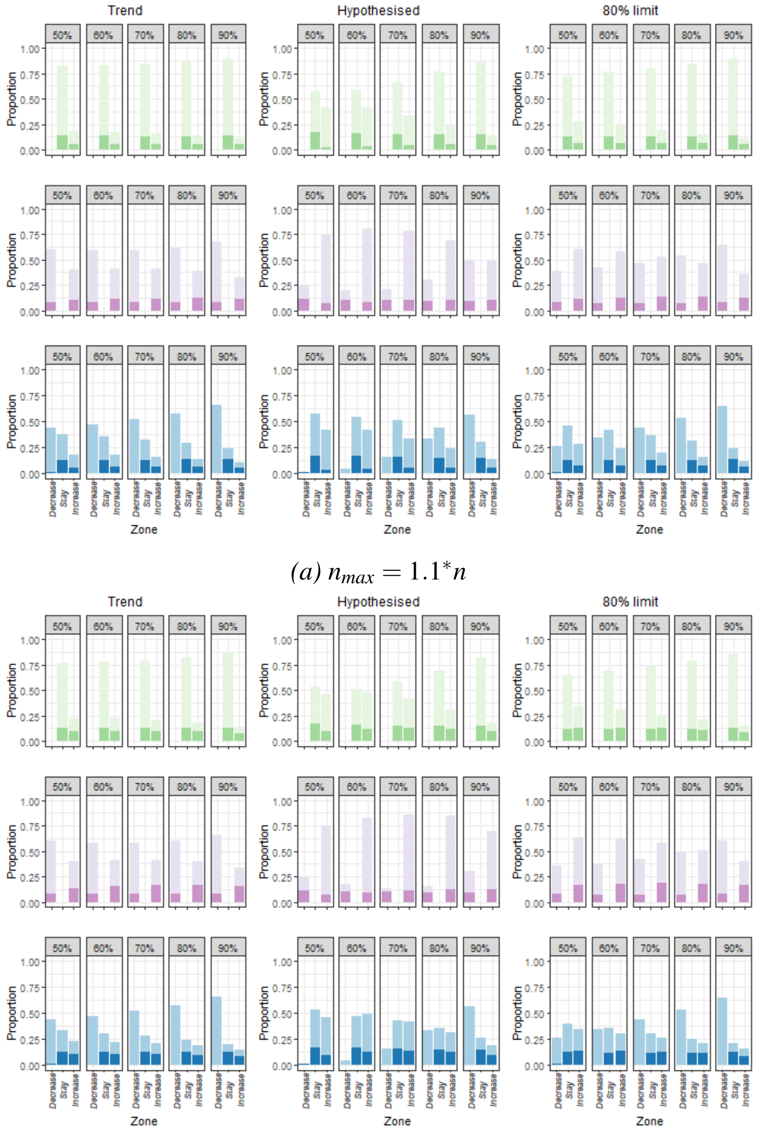
δ_{plan}=**0.2**		**SHORT**					**MEDIUM**				
δ=$\frac{2}{3}\delta_{plan}$											
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	1277	1330	1364	1375	1353	1316	1375	1420	1439	1383
	Power	68.40	71.49	73.57	74.36	72.88	69.08	72.19	74.14	74.74	72.93
γ_2	ASN	1129	1196	1245	1274	1281	1179	1253	1315	1350	1314
	Power	63.05	66.8	69.63	71.31	71.07	64.24	68.06	70.63	72.02	71.24
γ_3	ASN	804	893	972	1046	1113	877	973	1067	1143	1153
	Power	48.79	53.73	57.82	61.52	64.29	51.27	56.30	60.18	63.36	64.68
γ_4	ASN	654	749	841	932	1024	756	862	971	1056	1070
	Power	38.39	43.94	49.27	54.13	58.60	38.39	43.94	49.27	54.13	58.60
HYPOTHESISED											
γ_1	ASN	1165	1230	1292	1348	1385	1168	1241	1318	1388	1407
	Power	62.71	65.42	67.65	69.25	69.28	62.77	65.61	67.94	69.55	69.33
γ_2	ASN	1067	1140	1213	1282	1331	1073	1158	1248	1331	1356
	Power	57.94	61.30	63.87	66.19	67.16	58.15	61.74	64.48	66.77	67.31
γ_3	ASN	929	1011	1094	1171	1231	945	1042	1147	1238	1263
	Power	50.99	54.72	58.76	61.64	63.99	51.74	56.12	60.20	63.02	64.28
γ_4	ASN	799	876	953	1027	1096	839	940	1044	1127	1139
	Power	44.26	48.89	53.44	57.49	60.83	46.41	51.87	56.57	60.07	61.44
80% Limit											
γ_1	ASN	1435	1469	1474	1456	1400	1451	1493	1513	1508	1428
	Power	74.29	76.17	76.63	76.14	73.64	74.47	76.42	76.92	76.35	73.67
γ_2	ASN	1284	1326	1348	1347	1320	1309	1363	1401	1413	1352
	Power	69.14	71.92	73.20	73.55	72.02	69.54	72.50	73.78	74.04	72.15
γ_3	ASN	915	980	1040	1093	1140	964	1042	1121	1182	1179
	Power	54.40	58.09	61.31	63.86	65.61	55.59	59.66	62.91	65.25	65.94
γ_4	ASN	692	775	857	941	1028	771	871	975	1059	1073
	Power	41.20	45.93	50.68	54.84	58.82	44.41	49.77	54.65	58.11	59.56

Table E.5: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when $\delta=\frac{2}{3}\delta_{plan}$, n=1052, $n_{max} = 2^{*}n$

$\delta = \frac{2}{3}\delta_{plan}$		Information fraction									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	1326	1334	1338	1337	1333	1327	1319	1297	1269	1217
	Power	66.43	66.74	66.88	66.84	66.84	66.62	66.29	65.80	64.72	62.86
Hypothesised	ASN	1294	1330	1357	1375	1383	1376	1361	1330	1291	1228
	Power	69.29	71.24	72.81	73.96	74.42	74.08	73.15	71.27	68.97	65.24
80% limit	ASN	1414	1415	1415	1406	1393	1375	1355	1322	1284	1223
	Power	70.75	70.63	70.51	69.87	69.32	68.48	67.57	66.58	65.09	62.92
Promising zone with Futility											
Trend	ASN	1230	1247	1260	1268	1272	1273	1273	1260	1241	1201
	Power	63.95	64.62	65.10	65.38	65.66	65.66	65.53	65.21	64.27	62.57
Hypothesised	ASN	1294	1329	1355	1371	1374	1363	1344	1309	1271	1214
	Power	69.32	71.26	72.83	73.97	74.43	74.07	73.14	71.23	68.91	65.17
80% limit	ASN	1372	1371	1369	1360	1348	1332	1316	1288	1257	1207
	Power	70.23	70.05	69.95	69.33	68.79	68.00	67.13	66.18	64.72	62.65

Table E.6: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions for the promising zone design with and without a futility boundary when $\delta = \frac{2}{3}\delta_{plan}$, n=1052 and $n_{max} = 2^{*}n$

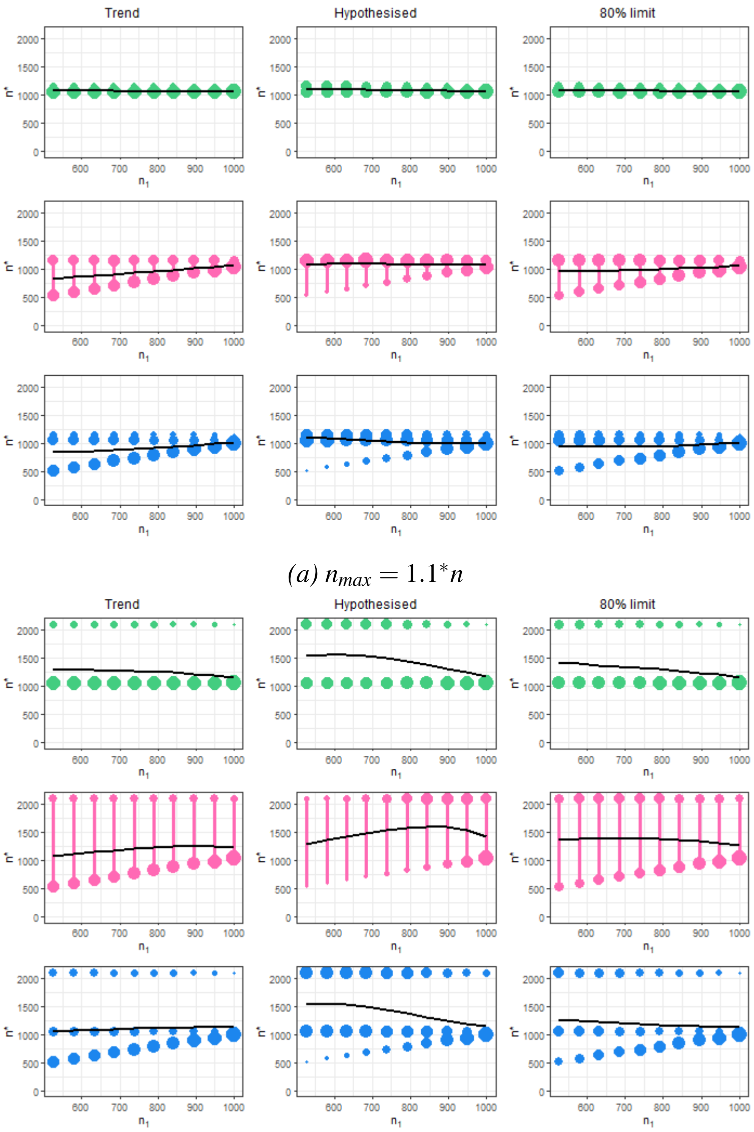
E.1.3 Observed effect = one third planned



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

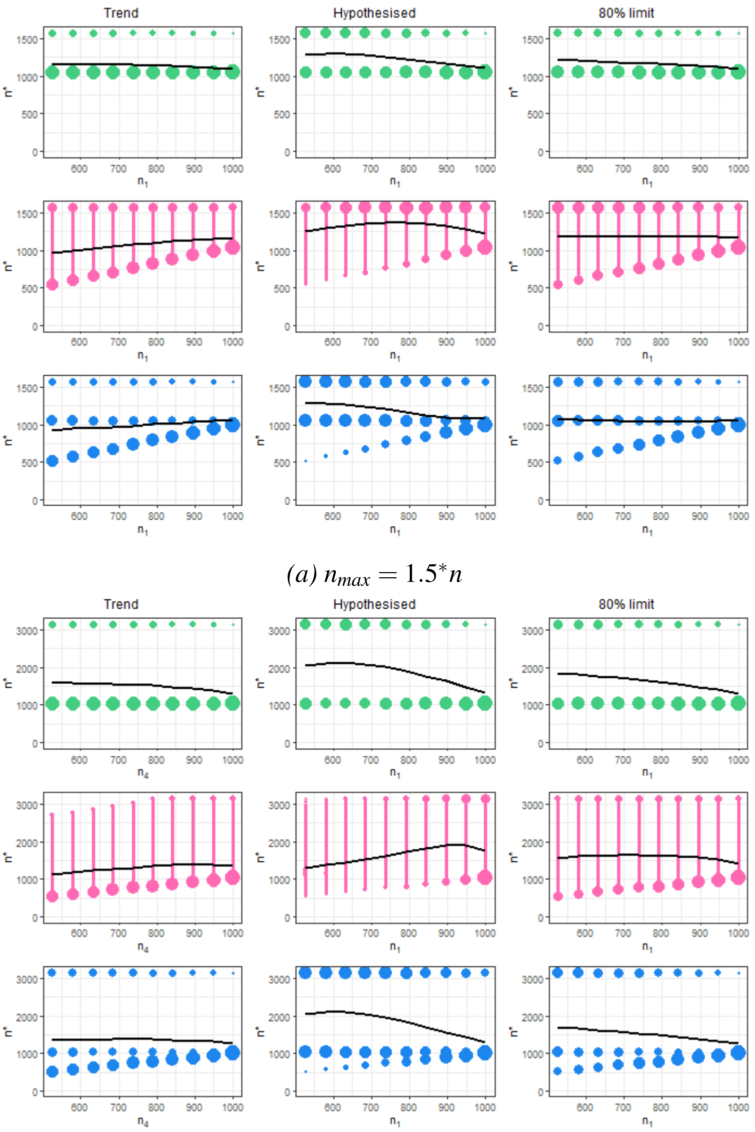
Figure E.7: Sample size zones from 50000 simulations when $\delta = \frac{1}{3}\delta_{plan}$ and n=1052, for two values of n_{max}: comparing three designs and four observed treatment effects



(a) $n_{max} = 1.1^{*}n$

(b) $n_{max} = 2^{*}n$

Figure E.8: Sample size from 50000 simulations when $\delta = \frac{1}{3}\delta_{plan}$ and n=1052, for two values of n_{max}: comparing three designs and three treatment effect assumptions



(a) $n_{max} = 1.5^{*}n$

(b) $n_{max} = 3^{*}n$

Figure E.9: Sample size from 50000 simulations when $\delta = \frac{1}{3}\delta_{plan}$ *and n=1052, for two values of* n_{max}*: comparing three designs and three treatment effect assumptions*

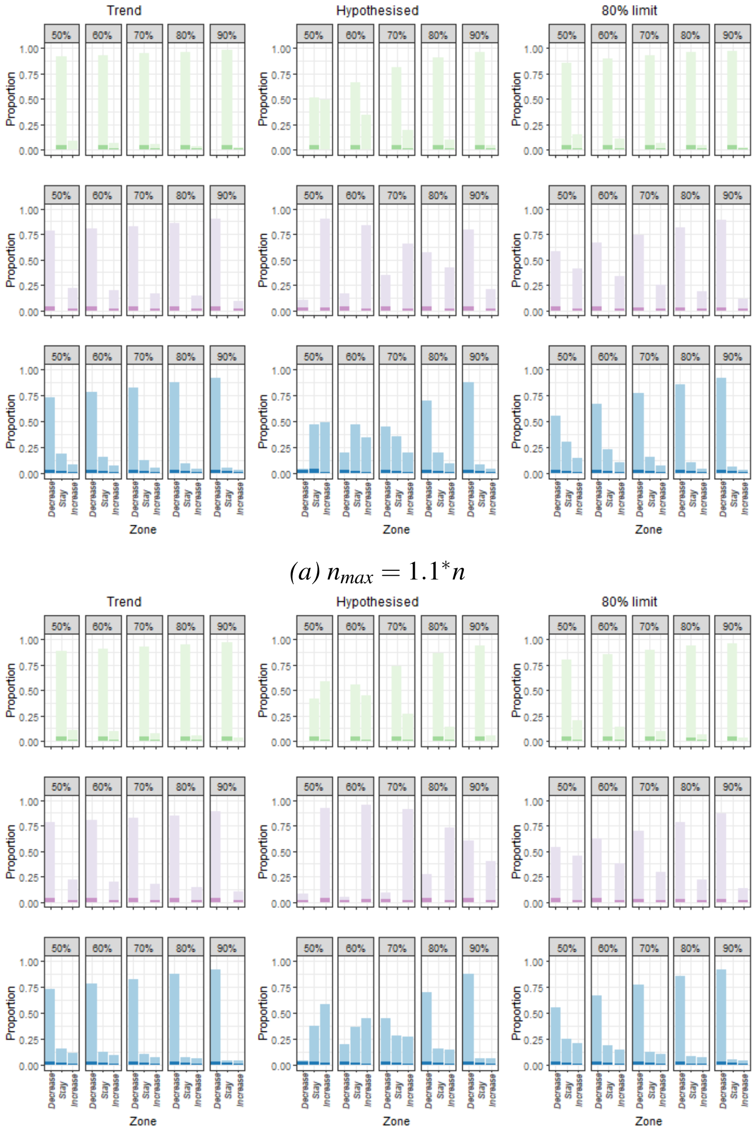
δ_{plan}=**0.2**		**SHORT**					**MEDIUM**				
δ=$\frac{1}{3}\delta_{plan}$											
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	1225	1285	1327	1348	1330	1304	1369	1420	1440	1368
	Power	24.61	25.63	26.02	25.90	24.87	24.46	25.69	25.99	25.90	24.88
γ_2	ASN	1072	1148	1204	1245	1257	1163	1244	1311	1347	1299
	Power	22.54	23.63	24.24	24.68	24.16	22.50	23.83	24.38	24.75	24.18
γ_3	ASN	743	836	927	1013	1091	855	954	1054	1132	1138
	Power	16.99	18.13	19.63	20.81	21.67	17.26	18.70	20.13	21.25	21.75
γ_4	ASN	616	716	817	915	1015	752	860	970	1055	1066
	Power	13.78	15.09	16.68	18.09	19.54	13.78	15.09	16.68	18.09	19.54
HYPOTHESISED											
γ_1	ASN	1414	1535	1640	1697	1623	1415	1537	1646	1713	1641
	Power	20.93	21.73	22.54	23.18	22.87	20.95	21.76	22.55	23.13	22.89
γ_2	ASN	1296	1428	1545	1605	1540	1297	1431	1556	1631	1562
	Power	19.24	20.38	21.38	22.13	22.21	19.30	20.46	21.37	22.13	22.25
γ_3	ASN	1113	1238	1343	1395	1360	1119	1255	1381	1455	1395
	Power	17.05	18.30	19.61	20.58	21.12	17.14	18.50	19.81	20.93	21.22
γ_4	ASN	867	943	1012	1072	1117	922	1021	1116	1183	1166
	Power	15.49	16.78	18.12	19.13	20.11	15.64	17.33	18.75	19.91	20.28
80% Limit											
γ_1	ASN	1563	1568	1542	1493	1403	1604	1619	1609	1569	1439
	Power	26.36	27.06	26.87	26.36	25.10	26.09	26.91	26.82	26.31	25.11
γ_2	ASN	1375	1396	1390	1369	1316	1431	1465	1475	1458	1356
	Power	24.38	25.30	25.45	25.35	24.42	24.16	25.25	25.42	25.34	24.43
γ_3	ASN	887	951	1011	1069	1122	972	1048	1123	1180	1168
	Power	18.78	19.67	20.81	21.62	21.98	18.78	19.91	21.09	21.93	22.06
γ_4	ASN	652	740	832	924	1018	767	868	974	1058	1069
	Power	14.63	15.72	17.10	18.32	19.71	15.05	16.65	18.08	19.27	19.87

Table E.7: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when δ=$\frac{1}{3}\delta_{plan}$, n=1052, $n_{max} = 2^{*}n$

$\delta = \frac{1}{3}\delta_{plan}$		**Information fraction**									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	1293	1291	1285	1280	1270	1261	1244	1223	1201	1162
	Power	23.08	23.07	22.95	22.88	22.75	22.63	22.27	21.75	21.3	20.35
Hypothesised	ASN	1539	1564	1563	1537	1493	1440	1379	1315	1247	1180
	Power	26.28	27.32	28.19	28.66	28.54	28.01	27.11	25.77	23.94	21.76
80% limit	ASN	1416	1400	1375	1354	1328	1303	1274	1242	1211	1165
	Power	25.31	25.24	24.81	24.43	23.99	23.50	22.80	22.10	21.40	20.34
Promising zone with Futility											
Trend	ASN	1060	1076	1086	1098	1108	1119	1123	1128	1133	1125
	Power	22.12	22.24	22.21	22.29	22.23	22.20	21.89	21.43	21.04	20.13
Hypothesised	ASN	1536	1555	1545	1505	1445	1378	1309	1246	1189	1145
	Power	26.56	27.5	28.33	28.77	28.61	28.05	27.14	25.77	23.94	21.71
80% limit	ASN	1277	1257	1230	1211	1190	1176	1161	1150	1143	1129
	Power	25.29	25.14	24.69	24.30	23.80	23.28	22.59	21.89	21.18	20.14

Table E.8: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions for the promising zone design with and without a futility boundary when $\delta = \frac{1}{3}\delta_{plan}$, *n=1052 and* $n_{max} = 2^{*}n$

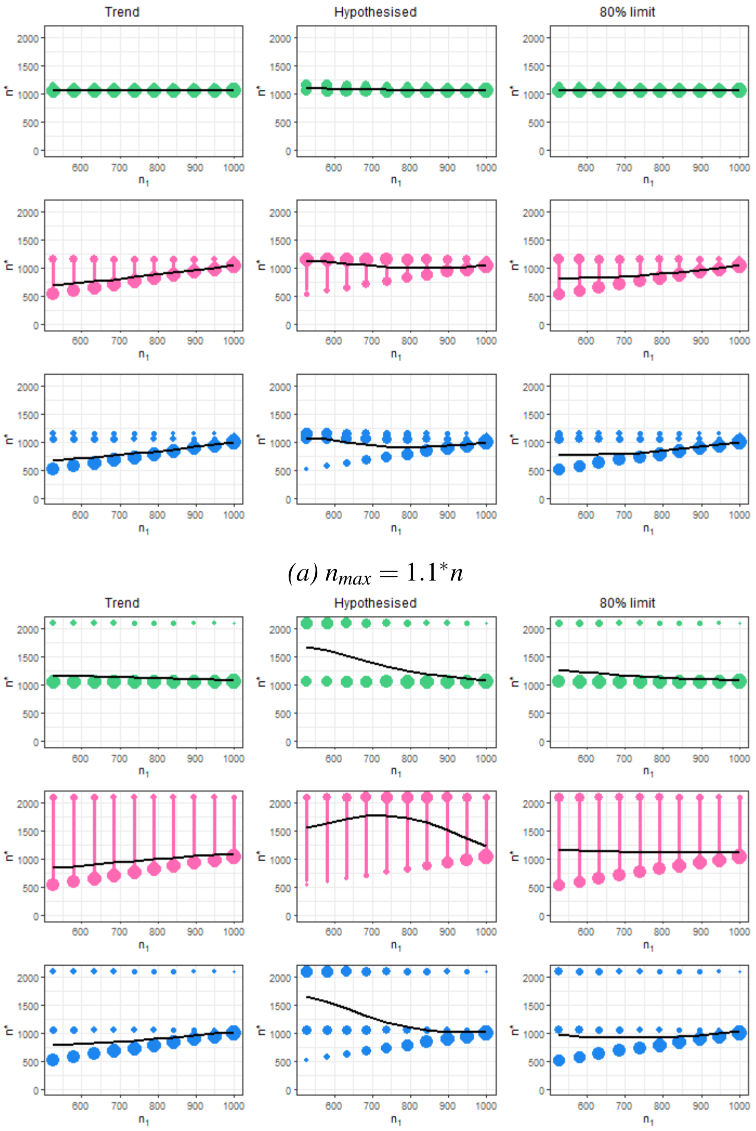
E.1.4 Observed effect = zero



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

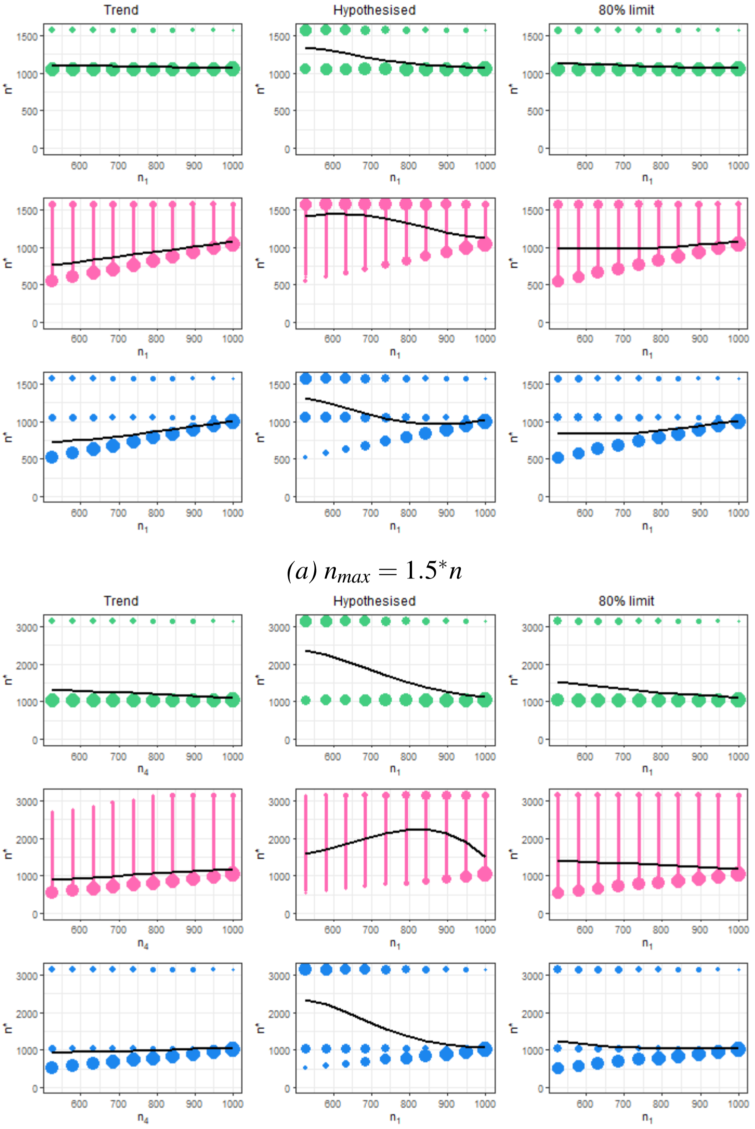
Figure E.10: Sample size zones from 50000 simulations when $\delta = 0$ and n=1052, for two values of n_{max}: comparing three designs and four observed treatment effects



(a) $n_{max} = 1.1{*}n$

(b) $n_{max} = 2{*}n$

Figure E.11: Sample size from 50000 simulations when $\delta = 0$ *and n=1052, for two values of* n_{max}*: comparing three designs and three treatment effect assumptions*



(a) $n_{max} = 1.5{*}n$

(b) $n_{max} = 3{*}n$

Figure E.12: Sample size from 50000 simulations when $\delta = 0$ and n=1052, for two values of n_{max}: comparing three designs and three treatment effect assumptions

δ_{plan}=0.2		SHORT					MEDIUM				
δ=0											
TREND		**50%**	**60%**	**70%**	**80%**	**90%**	**50%**	**60%**	**70%**	**80%**	**90%**
γ_1	ASN	942	989	1038	1081	1115	1076	1133	1192	1225	1171
	Power	5.53	5.47	5.25	5.22	5.16	5.10	5.24	5.09	5.04	5.07
γ_2	ASN	841	906	969	1030	1085	984	1058	1131	1179	1142
	Power	5.57	5.44	5.27	5.26	5.17	5.09	5.20	5.09	5.09	5.06
γ_3	ASN	638	734	830	926	1022	796	898	1003	1083	1080
	Power	5.60	5.47	5.28	5.32	5.17	5.08	5.20	5.06	5.12	5.10
γ_4	ASN	575	679	785	890	996	745	856	968	1053	1056
	Power	5.64	5.44	5.31	5.32	5.18	5.64	5.44	5.31	5.32	5.18
HYPOTHESISED											
γ_1	ASN	1676	1819	1879	1769	1451	1676	1820	1893	1815	1489
	Power	5.01	5.07	5.09	5.20	5.22	5.01	5.08	5.05	5.08	5.15
γ_2	ASN	1553	1716	1773	1650	1370	1553	1721	1800	1715	1413
	Power	5.04	5.10	5.06	5.18	5.16	5.04	5.15	5.05	5.04	5.10
γ_3	ASN	1312	1430	1446	1358	1216	1329	1483	1545	1480	1270
	Power	5.09	5.13	5.15	5.23	5.14	5.01	5.12	5.05	5.06	5.09
γ_4	ASN	851	899	944	993	1047	970	1044	1109	1149	1106
	Power	5.36	5.32	5.24	5.26	5.13	5.09	5.18	5.07	5.10	5.09
80% Limit											
γ_1	ASN	1363	1296	1235	1190	1154	1454	1410	1370	1324	1209
	Power	5.45	5.38	5.22	5.19	5.16	5.19	5.18	5.09	5.02	5.07
γ_2	ASN	1176	1141	1119	1113	1115	1288	1271	1268	1255	1171
	Power	5.43	5.42	5.25	5.28	5.16	5.08	5.2	5.13	5.11	5.07
γ_3	ASN	748	808	879	955	1034	886	959	1043	1107	1092
	Power	5.54	5.44	5.28	5.28	5.18	5.10	5.19	5.08	5.09	5.11
γ_4	ASN	595	691	791	893	998	753	860	970	1054	1057
	Power	5.61	5.49	5.30	5.31	5.18	5.09	5.24	5.07	5.10	5.10

Table E.9: Average sample number (ASN) and power for the combination test, comparing short and medium times to primary outcome data becoming available and 4 values of γ when δ=0, n=1052, $n_{max} = 2^{*}n$

$\delta = 0$		**Information fraction**									
Promising zone		**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**
Trend	ASN	1170	1161	1152	1142	1131	1121	1112	1101	1090	1078
	Power	4.98	4.89	4.93	4.88	4.85	4.80	4.75	4.66	4.65	4.62
Hypothesised	ASN	1669	1609	1521	1426	1335	1260	1199	1151	1113	1086
	Power	6.09	6.04	5.81	5.65	5.55	5.47	5.29	5.11	4.96	4.80
80% limit	ASN	1265	1233	1205	1180	1158	1138	1123	1107	1092	1079
	Power	5.12	5.00	4.98	4.88	4.83	4.76	4.72	4.62	4.61	4.60
Promising zone with Futility											
Trend	ASN	785	804	825	846	872	898	929	961	994	1029
	Power	4.78	4.77	4.76	4.75	4.72	4.75	4.63	4.65	4.63	4.63
Hypothesised	ASN	1647	1561	1439	1311	1193	1107	1052	1025	1022	1037
	Power	7.49	6.78	5.97	5.71	5.54	5.53	5.26	5.17	4.99	4.84
80% limit	ASN	977	944	924	914	915	924	944	968	997	1029
	Power	5.06	5.01	4.91	4.85	4.78	4.76	4.63	4.62	4.60	4.61

Table E.10: Average sample number (ASN) and power from 50000 repetitions for three treatment effect assumptions for the promising zone design with and without a futility boundary when δ=0, n=1052 and $n_{max} = 2^{*}n$