Statistical issues when incorporating emerging therapies into ongoing randomised clinical trials

Dena Rose Howard

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Leeds Institute of Clinical Trials Research

Faculty of Medicine and Health

September, 2018

Intellectual property and publication statements

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The supervision team for this research are Professor Julia M Brown, Professor Susan Todd and Professor Walter M Gregory. The supervision team provided advice, direction and review for each chapter and related research output. DRH is solely responsible for conducting the research and writing the thesis.

DRH is the Methodological Lead and Supervising Statistician for the FLAIR trial described in Chapter 6, with responsibility for the trial design, statistical validity and statistical oversight. Mrs Anna Hockaday, Head of Trial Management, reviewed Chapter 6 and input into the Trial Management sections. Professor Peter Hillmen, Chief Investigator, has overall responsibility for the trial and input into the protocol from which the chapter is based, along with other members of the Trial Management Group.

Journal articles

1. Cohen DR[i], Todd S, Gregory WM and Brown JM. Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. *Trials*. 2015; 16: 179.

   Chapter 2 is based on the work within this publication. The literature review was conducted by DRH with input into the search terms from the supervision team. The relevant statistical considerations were identified by DRH, and were agreed in discussion with the supervision team. DRH prepared the manuscript, which was reviewed by the supervision team.

[i] Cohen was Dena Howard's maiden name

2. Howard DR, Brown JM, Todd S and Gregory WM. Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. *Statistical Methods in Medical Research*. 2018; 27: 1513-30.

Chapters 3 and 4 are based on the work within this publication. DRH was responsible for the content and preparation of the manuscript in its entirety. The supervision team input into the research at monthly meetings, and reviewed the research output and manuscript.

Other Research Output

Incorporating emerging therapies into ongoing randomised clinical trials.
Cohen DR, Gregory WM, Todd S and Brown JM.
Oral presentation (invited speaker)
Royal Statistical Society local group, Leeds UK, October 2012.
Includes work presented in Chapter 2.

Incorporating emerging therapies into ongoing randomised clinical trials: literature review findings.
Cohen DR, Gregory WM, Todd S and Brown JM.
Oral presentation (contributed session)
Society for Clinical Trials international conference, Boston USA, May 2013.
Includes work presented in Chapter 2.

Statistical Considerations when Adding a New Treatment Arm into Ongoing Clinical Trials: Review of Literature and Comparison to Current Practice.
Cohen DR, Todd S, Gregory WM and Brown JM.
Oral presentation (invited speaker)
MRC Hub Network's ‘Adaptive designs for clinical trials’ workshop, Cambridge UK, January 2014.
Includes work presented in Chapter 2.

Efficient Trial Amendments in Practice: the FLAIR and Myeloma XI amendments.
Cohen DR and Gregory WM.
Oral presentations (invited speaker)
NIHR/NCRI 9th Annual Meeting of Cancer Clinical Trials Units, UK, July 2014, and
UKCRC Registered CTUs Network: Bi-annual Statistician's Operational Group Meeting, Leeds UK, October 2014.
Includes work presented in Chapters 2 and 6.

Incorporating Emerging Therapies into Ongoing Randomised Clinical Trials: Improving Efficiency and Relevance.
Cohen DR, Gregory WM, Todd S and Brown JM.
Seminars
Institute of Cancer Research, London UK, November 2014, and
University of Sydney National Health and Medical Research Council Clinical Trials Centre, Sydney Australia, September 2015, and
University of New South Wales, Sydney Australia, September 2015, and
Clinical and Transitional Radiotherapy workshop, Leeds UK, October 2015.
Includes work presented in Chapters 2 and 6.

Recommendations on multiple testing adjustment in multi-arm trials with correlated hypotheses.
Howard DR, Brown JM, Todd S and Gregory WM.
Oral presentations (contributed sessions)
Society for Clinical Trials international conference, Montreal Canada, May 2016, and
International Society for Clinical Biostatistics (ISCB), Birmingham UK, August 2016 [ST delivered the ISCB presentation on behalf of DRH].
Includes work presented in Chapters 3 and 4.

Incorporating emerging experimental therapies into ongoing randomised clinical trials with recommendations on multiple-testing adjustment.
Howard DR, Todd S, Gregory WM and Brown JM.
Oral presentation (invited speaker)
Royal Statistical Society international conference, Manchester UK, September 2016.
Includes work presented in Chapters 2, 3 and 4.

Acknowledgements

I would like to thank my supervisors, Professor Julia Brown, Professor Susan Todd and Professor Walter Gregory, for their continued guidance, encouragement and support. I have been extremely lucky to have not one, but three, supervisors who cared so much about my work, dedicated their time, shared their expertise and responded to my queries so promptly.

I am extremely grateful to the National Institute for Health Research for awarding a Doctoral Research Fellowship to provide financial support which allowed me to undertake this research, and gave me the opportunity to attend conferences and meet so many interesting people.

Thanks are due to staff and friends at the Leeds Clinical Trials Research Unit and the Trial Management Group for our CLL portfolio for their support and understanding. I am especially grateful to Professor Peter Hillmen, Mrs Anna Hockaday and all members of the FLAIR Trial Management Group for their roles in enabling the FLAIR trial to successfully incorporate a new experimental treatment arm.

My family and friends have provided a great deal of encouragement and support throughout this fellowship; in particular my husband Steven Howard, and parents Andrea and Philip Cohen, without whom I would not have come this far. Finally, I am thankful to my son, Alistair, who makes me laugh every day.

Funding acknowledgement

This research was funded by the NIHR Doctoral Research Fellowship scheme (reference number NIHR-DRF-2012-05-364).

Department of Health and Social Care disclaimer

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Abstract

This research investigates the incorporation of an emerging therapy as a new randomised arm in a confirmatory clinical trial that is open to recruitment. It may take many years to run confirmatory trials from concept to reporting within a rapidly changing drug development environment, hence in order to optimally inform policy and practice it is advantageous for trials to be able to adapt to emerging developments. It is becoming increasingly desirable to researchers, regulators and patients to allow such an adaptation to be made within a clinical trial to ensure that new treatments are evaluated as quickly as possible, and resources are optimised.

A comprehensive literature review confirmed that there is currently no clear methodological guidance on this topic, although treatment arms have been added into confirmatory trials in practice. Unfortunately the statistical considerations were not always appropriately addressed, often leading to uninterpretable or invalid outcomes. In this research, the necessary considerations to ensure statistical validity are identified and considered. The probability of false positive conclusions must be controlled, whilst ensuring that trial outcomes are not compromised or biased by the amendment. The need for multiple testing adjustment due to assessing multiple hypotheses within the same protocol and with shared control data is investigated, and recommendations are provided that apply to multi-arm trials in general. Adaptive analysis methods using p-value combination across the trial stages are compared to non-adaptive methods, with appropriate multiplicity adjustment considered. The findings are implemented to successfully incorporate a new experimental therapy within a large, confirmatory leukaemia trial.

Guidance is presented detailing the requirements deemed necessary to ensure statistical validity, including recommendations so that the adaptation to add a new experimental arm to an ongoing trial is appropriate and acceptable. It is hoped that this will encourage consideration of adding arms more widely in future.

Table of Contents

List of Tables

List of Figures

Abbreviations

ANOVA	Analysis of Variance
BCR	B-cell receptor
Btk	Bruton's Tyrosine Kinase
CCRD	Carfilzomib, Cyclophosphamide, Lenalidomide and Dexamethasone
CDF	Cumulative Distribution Function
CI	Confidence Interval
CLL	Chronic Lymphocytic Leukaemia
CONSORT	Consolidated Standards of Reporting Trials
CRD	Cyclophosphamide, Lenalidomide and Dexamethasone
CRF	Case Record Form
CRUK	Cancer Research UK
CTAAC	Clinical Trials Advisory and Award Committee
CTD	Cyclophosphamide, Thalidomide and Dexamethasone
DMEC	Data Monitoring and Ethics Committee
EFC	Expected number of False Claims
EMA	European Medicines Agency
FCR	Fludarabine, Cyclophosphamide and Rituximab
FDA	Food and Drug Administration
FISHER	Fisher's p-value combination method
FLAIR	Front Line therapy in CLL: Assessment of Ibrutinib-containing Regimes
FMER	Family Multiple Error Rate
FMER2	At least two false-positive errors within the family
FMER3	At least three false-positive errors within the family
FWER	Family-Wise Error Rate
H_0	Null Hypothesis
H_1	Alternative Hypothesis
HR	Hazard Ratio
HTA	Health Technology Assessment
I	Ibrutinib monotherapy
I+V	Ibrutinib and Venetoclax
ICH	International Conference on Harmonisation
INVN	Weighted inverse normal p-value combination method
IR	Ibrutinib and Rituximab
MAMS	Multi-Arm Multi-Stage
med	Median

MEER	Maximum Experiment-wise Error Rate
MeSH	Medical Subject Heading
MHRA	Medicines and Healthcare Products Regulatory Authority
MRD	Minimal Residual Disease
MSFP	Multiple Superior False Positives
MSFP2	At least two superior false positive outcomes
MSFP3	At least three superior false positive outcomes
MULTI	Pooled analysis adjusting for stage using a multivariable model
NICE	National Institute for Health and Care Excellence
NIHR HTA	National Institute for Health Research Health Technology Assessment
NRES	National Research Ethics Service
OxFP	Oxaliplatin and Fluoropyrimidine
PASI	Psoriasis Area and Severity Index
PCER	Per-Comparison Error Rate
PDF	Probability Density Function
PFS	Progression-Free Survival
POOLED	Pooled analysis without adjustment
PPI	Patient and Public Involvement
PSI	Statisticians in the Pharmaceutical Industry
PWER	Pairwise Error Rate
RCT	Randomised Controlled Trial
sPGA	Static Physician Global Assessment
TAP	Trials Acceleration Programme
t_d	Dunnett's t
TMG	Trial Management Group
TSC	Trial Steering Committee
UK	United Kingdom
UKCRC	UK Clinical Research Collaboration
V	Venetoclax
V+R	Venetoclax and Rituximab

Chapter 1
Introduction

1.1 Overarching aim

The aim of this research is to identify, review and investigate statistical considerations when incorporating an emerging therapy as a new randomised arm to be included in a confirmatory clinical trial that is already open to recruitment. When adapting a trial in this way, it is vital that the methodology is appropriate in order to maintain the validity of the conclusions. As a result of this work, recommendations are made to help researchers feel confident in considering and applying this type of adaptation to their trial designs efficiently and without statistical bias.

1.2 Background and rationale

Confirmatory clinical trials can take many years to run, requiring considerable resources. In addition, new therapies or therapy combinations are often in different stages of development, and new evidence of promising therapies for a particular population may emerge from early stages of development at different times. It would therefore be advantageous to be able to incorporate emerging therapies into ongoing trials as a new randomised arm. This would help to ensure that the outcomes of trials are relevant at the time of reporting; whilst benefitting patients, funders, trialists and regulatory bodies by shortening the overall process of comparing and selecting experimental treatments. This allows optimal therapies to be determined faster than would otherwise be the case, and can reduce costs and patient numbers. In addition, increasing the number of experimental arms in a trial increases the probability of identifying a successful treatment[1].

Ongoing treatment advances are continually improving survival rates in many therapeutic areas, including, for example, in most types of cancer[2]. The Cancer Research UK (CRUK) website states that “Half (50%) of people diagnosed with cancer in England and Wales survive their disease for ten years or more”[3]. Improving survival times are fantastic for patients, but presents challenges to researchers in continuing to

progress and further improve these survival rates within feasible trials settings. Whilst shorter term surrogate outcome measures can help to improve the efficiency of trials, where good substitutes exist, trial set up times and recruitment periods can still be long and further efficiencies would be beneficial. New promising treatments are continually being developed and tested in early phase trials, and it is difficult for researchers to address them in a confirmatory setting in a timely manner. It can be considered unethical to delay a phase III trial to assess currently available treatments whilst waiting for the results of ongoing promising early phase trials. The ability to add new arms to ongoing trials could help to advance the pace of research by allowing emerging therapies to be investigated in populations where trials already exist without introducing competition, and by reducing the set-up time for designing a new trial.

This type of amendment falls under the umbrella topic of 'adaptive designs'. An adaptive design refers to a "clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial"[4]. The aim is to improve efficiency by reducing the overall resources needed to be able to answer the relevant questions for a particular group of patients. Adaptive design methodology for confirmatory trials has been discussed in statistical literature for over 25 years[5] with continually growing popularity, and is a key topic in the statistical community with significant contributions in statistical literature including major regulatory guidance documents. Adding a new arm, however, is not strictly an adaptive design because the evidence informing the amendment is likely to emerge from data external to the trial being adapted rather than accumulating data internal to the study, which has a different impact on the statistical implications. It is not clear which of the considerations related to adaptive designs are relevant for this type of amendment, or how to address them. Such an adaptation could be just as advantageous as other, standard, adaptive designs for improving efficiency, and comes without some of the complexity or controversy due to using internal trial data. Therefore guidance is needed in this area to encourage and inform this type of amendment.

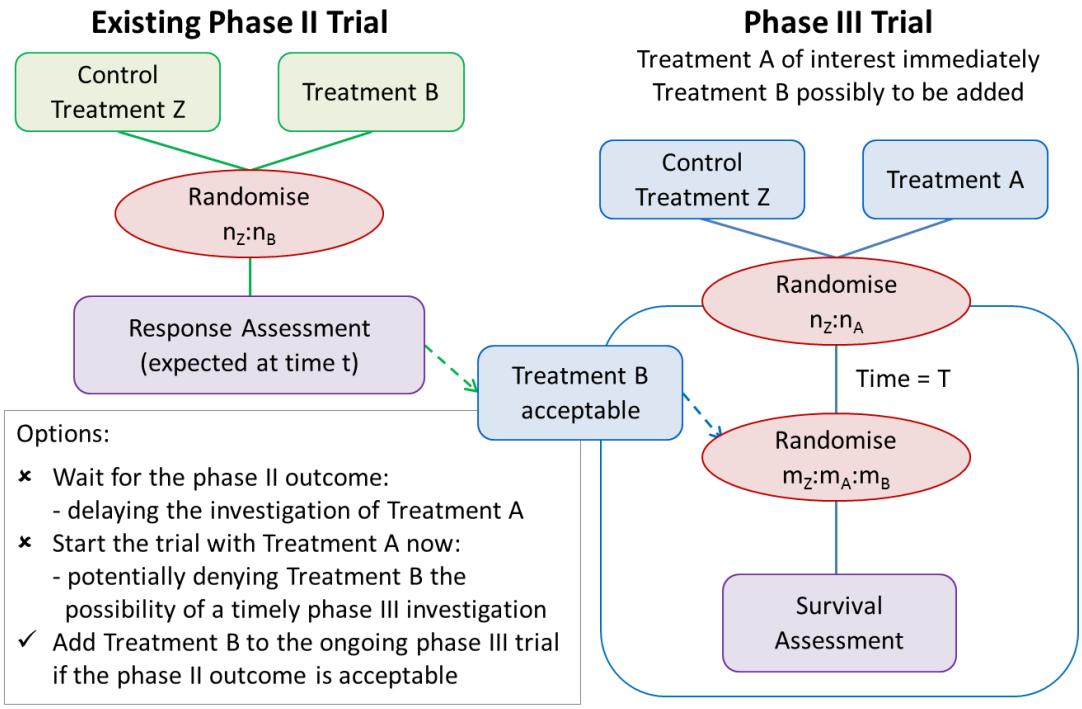
1.3 Motivating example

Treatment for Chronic Lymphocytic Leukaemia (CLL), a cancer of the white blood cells, is currently a rapidly moving field in which multiple new therapies are being developed which show considerable early promise. Early phase trials are continually researching new targeted therapies, both alone and in combination, and a platform is needed for

promising emerging therapies to be rapidly included within a confirmatory trial. The generally accepted primary endpoint in a CLL trial is progression-free survival (PFS), which is defined as time from randomisation to clinical progression or death from any cause. Currently the median PFS in newly diagnosed patients is approximately 5 years, meaning that a confirmatory trial assessing a survival improvement is likely to take at least 10 years to be reported. If it is possible to add a new promising treatment to an ongoing trial should one emerge during the recruitment phase, this would clearly be beneficial for the reasons outlined above.

The motivating example for this research is the FLAIR trial in newly diagnosed CLL (detailed in Chapter 6). An example of the scenario at the time of designing the original trial is used as the basis for this research, illustrated in Figure 1-1. Treatment A was immediately available for assessment in a large, confirmatory phase III trial in newly diagnosed CLL patients in the UK. However, a promising Treatment B was undergoing assessment in a phase II trial in the same population. The phase II trial was shortly due to complete recruitment, but required 12 months of follow-up for the outcomes. The tchoice was either to delay the assessment of Treatment A, therefore denying patients that promising new therapy in a trial setting and delaying the research; or opening the phase III trial and denying Treatment B the possibility of a timely phase III investigation in that population. Ideally the phase III trial assessing Treatment A would be opened immediately, with Treatment B incorporated at a later time if the phase II evidence was promising.

Figure 1-1 Scenario in which it would be beneficial to add a treatment arm to a phase III trial



1.4 Scope of the research

This thesis investigates the addition of a new treatment arm to an ongoing trial under the following situations:

- The trial has a confirmatory primary objective.
- The trial has already begun recruitment and the randomisation is still open when the new treatment is to be added.
- The new therapy is to be assessed within the current trial population, against the same standard-of-care. If any changes are necessary to eligibility criteria, these are minimal and do not materially change the trial population.
- The entire treatment arm will be new; an amendment to an existing arm to include a new treatment is not relevant.
- The treatment arm will be added to the existing randomisation; rather than including a new separate randomisation for a subgroup of patients within a master protocol.
- The trial is designed using frequentist methodology.
- The evidence for the new treatment has arisen externally to the trial being adapted, rather than due to the findings within an internal trial analysis. Adaptive dose finding trials, for example, are not relevant.

Since the motivating example is in a cancer trial, this research focuses on trials with survival outcomes, using trials in haematological oncology as exemplars, with generalisability into other disease areas and endpoints. It is assumed that the control treatment is the same for all experimental arms, and primary comparisons will be pairwise between each experimental treatment and control.

1.5 Framework

In Chapter 2 we report a comprehensive literature review on methodologies for, and practical examples of, amending an ongoing trial by adding a new treatment arm. Relevant methodological literature describing statistical considerations required when making this specific type of amendment is identified, and the key statistical concepts when planning the addition of a new treatment arm are extracted, assessed and summarised. This includes an assessment of statistical recommendations within general adaptive design guidance documents, for completeness. An assessment is made as to how the relevant statistical considerations have been addressed in practice, and the related implications to the statistical validity of the trial outcomes. Uncertainties or inconsistencies within the literature around the statistical concepts that were identified were used to inform the focus of the remainder of this research. The findings from this review were published in the Trials journal[6] to help researchers wishing to implement this type of amendment in practice.

When an arm is added to an ongoing trial, by definition it becomes a multi-arm trial with multiple hypotheses assessed within the same protocol. A key statistical consideration, with mixed views in the literature, was identified to be the necessity for familywise error rate control in order to maintain the chance of at least one false positive outcome to be less than the required significance level. In Chapter 3 the concept of adjustment and conflicting viewpoints within the literature are reviewed, and common adjustment methods are described. Whilst multiplicity adjustment is a very common statistical issue, there is surprisingly little literature that considers the need to adjust for multiple hypotheses with shared control data. In a recent conference expert panel discussion on adaptive clinical trial designs[7], Michael Proschan commented “I think there needs to be more research on whether you really need to adjust and when you need to adjust. I try to be consistent and coherent with multiple comparisons and I’m not. There are too many papers on how to do multiple comparisons and not enough on when you really need to do such adjustments.” The reasons why false positive findings may be inflated

in multi-arm trials over independent trials are broken down into: multiple chances of making a claim of effectiveness due to the efficiency of testing more than one hypothesis in the same protocol; and the shared use of control data. Whilst the first point is widely debated, the second is less well addressed and generally not well understood. In Chapter 4 the effect of the correlation between the test statistics for the hypotheses due to a shared control group on the probability of one or more false positive errors is comprehensively investigated. In addition, various multiple adjustment methods are reviewed as to how well they control these errors, leading to some unexpected results. A decision diagram is included to aid researchers in determining the requirement for multiple testing adjustment in their multi-arm trial, with examples. Since adjustment can have a significant effect on the power for hypotheses in a multi-arm trial, it is important that it is only implemented if necessary, in order to ensure that the trial design is as efficient as possible. The research within these chapters has been published in Statistical Methods in Medical Research[8].

Under the standard definition of adaptive designs in which a trial is amended based on accumulating data from within the trial, the design of the stage after the amendment is dependent on data from the stage before. In this case, it is required that the analysis accounts for this using adaptive design analysis methodology. One common technique is to use a combination test, where the p-values are calculated separately for each trial stage and then combined to produce the overall p-value. In the case of adding an arm, since the new treatment to be added is likely to have emerged from evidence external to the trial being adapted, such as from an exploratory phase of development, adaptive design analysis methodology may not be necessary. It is still possible, however, that the amendment could have consequences for the hypotheses of interest within each stage, causing a stage effect. In addition, although the evidence to add an arm is assumed to have arisen externally, data internal to the trial may have been analysed, and could influence the design characteristics. The requirements for analysis of a trial adding an arm are therefore assessed in Chapter 5. Depending on the nature of the experimental therapies, multiplicity adjustment may or may not be necessary, and this is also considered alongside the recommended analysis methods. In the case of adding am arm, it may be that the stage prior to the amendment does not include multiple hypotheses, but the stage after does. Where p-value combination methods are used, it is possible to adjust within stage rather than adjusting the final p-values across the stages, which is often the method recommended in adaptive design literature. However, in Chapter 5 it is shown that this may not be appropriate. A summary is provided on the recommended methods of analysis when a new arm is added to an ongoing trial, with and without multiple testing adjustment.

In Chapter 6 a real trial in Chronic Lymphocytic Leukaemia is described, in which a new experimental arm has been added based on the findings within this research. In addition, a second control arm was also added to protect the trial in case of a change in practice, and the original experimental treatment was discontinued on completion. The trial was increased from a planned 754 patients being randomised between two arms, to 1515 being randomised to up to four arms. By the time the amendment was implemented, there were only 61 control patients remaining for the original randomisation. Although the savings in patient numbers due to sharing control patients was relatively small, large advantages were gained in the time to assess the promising new treatment within a confirmatory trial, and the use of the existing trial structure. The amendment obtained all necessary approvals, was positively received and the trial remained ahead of its recruitment target. The methodological and logistical considerations are presented in detail.

The discussion and summary of the research in Chapter 7 describes all of the key results from the work. It also includes guidance and recommendations for researchers wishing to add a new experimental arm into an ongoing trial.

Chapter 2
Framing the research: review and summary of statistical considerations and practical implementation when adding a treatment arm into an ongoing clinical trial

2.1 Introduction

2.1.1 Aims

A literature review was conducted to identify and assess existing literature regarding statistical methods and design considerations, or their implementation, when adapting an ongoing trial by adding a new treatment arm. Literature was deemed to be relevant if it included either methodological considerations or practical discussions of trials which had implemented this type of adaptation, and it was within the scope of the research defined in Section 1.4. In addition, adaptive designs guidance documents were included in the review to identify the key statistical considerations when adapting a trial generally and to consider their relevance in this situation. The aim was to identify the potential statistical issues when adding an arm to an ongoing clinical trial, so that these can be appraised and summarised to form guidance when planning or reviewing this type of amendment. Areas of uncertainty and contradictions within the literature are identified to generate objectives for investigation within this research.

Firstly, some clinical trials and survival analysis terminology is introduced to inform the work in this chapter and the wider research. In Section 2.2 the literature review methods are described and the literature identified is summarised. In Section 2.3 the relevant considerations and methodology identified are reviewed, and it is assessed how these have been addressed in practice. Where recommendations are not clear, whether it is an area that has not arisen previously or there is contradiction in the literature, future chapters addressing these issues are signposted.

Part of the work presented in this chapter has been published as a review article in the Trials journal (Cohen et al. 2015)[6]. The publication is entitled “Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice”. The work contained within the publication is directly attributable to myself as first author, with input from the co-authors who are all part of my PhD supervisory team.

2.1.2 Clinical trials terminology

Randomised controlled trials (RCTs) are the gold standard in clinical trials, in which, in the simplest case, an experimental therapy of interest is prospectively compared to the current standard of care in a population to establish whether it is efficacious and safe. A confirmatory, or phase III, clinical trial is typically designed to conclusively demonstrate efficacy within acceptably small and controlled margins of error. Trial participants must meet the eligibility criteria to enter the trial, and are randomised to either the experimental or control arm. Once complete, the trial is analysed in order to assess whether there is evidence to reject the null hypothesis (H_0) of no significant difference between the therapies under comparison. The treatment effect, θ, can be calculated to be the difference in outcomes between participants randomised to the experimental arm and those randomised to the control arm. The null hypothesis can therefore be written as $H_0: \theta = 0$, with the two-sided alternative hypothesis $H_1: \theta \neq 0$. A minimal clinically important difference is determined, which is the smallest treatment effect that would influence practice. The trial is designed so that the minimal clinically important effect size is able to be reliably detected[9], and reported as a significant difference between treatment arms with appropriate and controlled levels of error, whilst preventing smaller true effect sizes from influencing practice. That is, the type I error (or false positive) rate, which is the probability of incorrectly declaring that there is a significant difference of clinical importance between the therapies under comparison when actually H_0 is true, is less than the significance level, α, as follows:

$$P(Reject\ H_0) \leq \alpha.$$

Conventionally, a two-sided α is generally accepted to equal 5%, such that the chance of falsely declaring the experimental therapy to be better than control is 2.5%. Trial results typically report a p-value (p), which is the probability of observing data as or more extreme than the observed difference between treatment outcomes given that the null hypothesis is true, thus the convention is that a clinical trial outcome is declared significant when $p \leq 0.05$.

Also of interest is the probability of correctly rejecting the null hypothesis and declaring a significant difference of clinical importance between the therapies under comparison, known as the power. A type II error (β) is the false negative rate, or probability of failure to reject a false null hypothesis, and therefore the power $= 1 - \beta$. Conventionally this is not required to be as stringent as the probability of a type I error, and typically the power is set to be 80% to 90%.

It is possible for a clinical trial to have 'multiple arms' such that it includes more than one experimental arm to be compared to the control, and therefore to have more than one primary hypothesis. In this case, there is an increased chance of a type I error over the set of hypotheses across the trial as a whole, known as the family-wise error rate (FWER). 'Multiplicity' is the term used to describe this increased chance of at least one false positive conclusion over the protocol as a whole due to multiple testing. Multiple testing adjustment methods can be applied to adjust for multiplicity to control the overall type I error across the set of hypothesis to be less than 5%, and will be discussed in Chapter 3. In contrast, the pairwise error rate (PWER) is the individual probability of a type I error for a single comparison within the protocol.

2.1.3 Survival analysis

2.1.3.1 Survival and hazard functions

In confirmatory cancer clinical trials, the primary endpoint is often a survival outcome such as progression-free or overall survival. Since the motivation for this research arose within a cancer trial, it will primarily focus on trials with survival primary outcomes, also known as time-to-event outcomes.

For each patient, they either have an event at time *t*, or they don't have an event and they are censored at time *c*, which is when they were last known to be event free. The survival function *S(t)* is the probability of surviving longer than time *t*,

$$S(t) = P(T > t) = 1 - F(t),$$

where *T* is a random variable associated with the survival time. *F(t)* is the lifetime distribution function or cumulative distribution function (CDF), which can be expressed as the integral of the probability density function (PDF) $f(t)$, so

$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(t)dt,$$

where $f(t)$ is the rate of events per unit time.

The hazard function of *T*, *h(t)*, gives the instantaneous probability of an event at time *t*,

$$h(t) = \frac{f(t)}{S(t)} = \frac{-d/dt S(t)}{S(t)} = -\frac{d}{dt} logS(t).$$

The cumulative hazard function, $H(t)$, is the probability of an event over time, so

$$H(t) = \int_0^t h(u)du = -logS(t).$$

2.1.3.2 The logrank test

The logrank test is a non-parametric test that compares the survival patterns of two or more trial arms, typically defined by treatment, by comparing estimates of their hazard functions. The logrank statistic is calculated based on the number of events observed (o) minus the number of events expected (e) if there were no difference between the arms, over each event time $r = 1,2,\dots,R$. There are o_{jr} events at time t_{jr} in arm j ($j = 1,2,\dots,J$). Let $n_r = \sum_{j=1}^{J} n_{jr}$ be the number of patients at risk at the start of period r. The expected values are therefore $e_{jr} = o_r \frac{n_{jr}}{n_r}$.

The logrank test statistic comparing the number of observed events to those expected in arm *j* can therefore be written as

$$Z_j = \frac{\sum_{r=1}^{R}(o_{jr} - e_{jr})}{\sqrt{\sum_{r=1}^{R} V_r}} \sim N(0,1), \text{ under } H_0,$$

where V_r is the variance.

Since it is assumed that the Z_j are independent standard Normally distributed random variables, $\sum_{j=1}^{J} Z_j^2 \sim \chi^2_{J-1}$. The p-value can therefore be found by comparing the sum of the squared test statistics for each arm to a chi-squared distribution with degrees of freedom equal to the number of arms (*J*) minus 1.

Note that whilst this non-parametric test is useful when comparing survival patterns, it is limited in that it does not account for explanatory covariates such as differences in baseline demographics between the treatment arms. In this case, the semi-parametric Cox proportional hazards regression model[10] can be used to model the hazard function

between treatment arms after accounting for covariates, based on the assumption that the hazard remains proportional over time. The hazard ratio between two trial arms is defined below (Section 2.1.3.3). The Cox model estimates the treatment effect in terms of the hazard ratio between the treatment arms, adjusted for covariates, and calculates a test statistic and p-value to assess whether the treatments are significantly different, as described in Section 5.2.2.

2.1.3.3 Sample size and power calculation based on an exponential survival assumption

It is a generally accepted approximation that survival curves follow an exponential distribution with rate parameter λ, which is proportional to the median survival (*med*) such that $\lambda_j = \frac{ln2}{med_j}$ in arm *j*. The proportion surviving to a fixed time *t*, *S(t)*, can be written π_j, where $\pi_j = e^{-\lambda_j t}$, because the CDF of an exponential distribution (signifying the failure rate at time *t, F(t)*) is $1 - e^{-\lambda t}$, and if a proportion π have survived to time *t*, then $1 - \pi$ have failed.

The hazard ratio (*HR*), which is the ratio of the hazard rates between two trial arms ($j = 1,2$), is therefore

$$HR = \frac{\lambda_2}{\lambda_1} = \frac{med_1}{med_2} = \frac{ln(\pi_{2)}}{ln(\pi_{1)}} = \ln(\pi_2 - \pi_1).$$

The hazard ratio is assumed to be constant over time, therefore requiring that the hazard functions for each arm are proportional over time. In this case, the treatment effect, θ, is the difference in outcomes between participants randomised to each trial arm assuming the log-hazard ratio scale. A $HR > 1$ implies that survival times are shorter for arm 2, and a $HR < 1$ implies that survival times are longer for arm 2.

In the case of powering a trial to assess whether two treatment arms are significantly different, the minimum clinically relevant effect between the two treatment arms in terms of the hazard ratio, and the minimum acceptable type I and II error rates, need to be agreed. These are used to calculate the number of events (*E*) needed, as follows[11]:

$$E = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 (HR+1)^2}{(HR-1)^2},$$

where *Z* is inverse of the CDF of the standard normal distribution, α is the required two-sided type I error rate, and $1 - \beta$ is the required power.

The number of patients required is determined based on observing this number of events by time *t* in the case of an equal allocation ratio, by:

$$n = \frac{2E}{(2-\pi_1-\pi_2)}.$$

Rearranging to calculate the power for a given sample size gives

$$Z_{1-\beta} = \frac{|HR-1|}{HR+1}\sqrt{\frac{n(2-\pi_1-\pi_2)}{2}} - Z_{1-\frac{\alpha}{2}}.$$

Note that when the allocation is unequal with ratio 1:φ, with φ being the proportion randomised to the experimental arm, this equation becomes:

$$Z_{1-\beta} = \frac{|HR-1|}{\varphi HR+1}\sqrt{\frac{n\varphi}{(1+\varphi)}[(1+\varphi) - \pi_1 - \varphi\pi_2]} - Z_{(1-\frac{\alpha}{2})}.$$

2.2 Comprehensive Literature Review

2.2.1 Methods

A protocol for the literature search was written in advance to fully define the aims, methods and search strategy to be used to obtain existing literature regarding statistical methods and design considerations when adapting an ongoing trial by adding a new treatment arm. In summary, search terms were defined for the following major electronic databases: MEDLINE (Ovid), EMBASE (Ovid), Science Citation Index (Web of Science) and the Cochrane Library (Wiley), each from their dates of inception. The ProQuest database was also searched to identify further relevant grey (unpublished) material such as dissertations and theses, and conference abstracts. The search terms are provided in Appendix A, and include: Medical Subject Heading (MeSH) terms relating to ‘clinical trials’ and ‘research design methodology’, as appropriate for the database; a term relating to ‘adaptive’, ‘flexible’, ‘multi-stage’ or ‘platform’ designs, methods or trials; and a term around adding or incorporating or an additional or extra ‘treatment’, ‘arm’, ‘group’, ‘therapy’, ‘randomisation’ or ‘hypothesis’. The search was initially conducted in November 2012, and auto alerts were set up where possible to keep abreast of any further literature throughout the research period. The searches were also run periodically. In order to identify any additional relevant publications, searches were performed on references, authors and citations of directly relevant

literature, and key methodologists in the field were identified and publications reviewed. Titles and abstracts were assessed to determine whether the literature was relevant, and shortlisted results were reviewed in full.

An assessment of summary, regulatory, guidance and review documents on flexible or adaptive designs in general was manually undertaken to identify methodologies that may be relevant. Books on adaptive or flexible trial design were also included. Literature was identified with direct relevance to general guidance or reviews on adaptive designs, but not including documents relating to a particular disease or methodology. The types of adaptation and key statistical considerations discussed in each document were listed and summarised, and their relevance to adaptation when adding an arm was determined in discussions with supervisors.

The literature review aimed to include identification of practical examples of trials that had added a new treatment part-way through recruitment and within the scope of the research, as described in Section 1.4. However, these were rarely identified as the design amendment is not the primary aim of trial results publications and does not feature in the title or abstract, or else trials were ongoing and had not yet been published. In order to identify as many trials as possible, key statisticians and researchers were contacted directly, and references from relevant methodological papers were reviewed. Statisticians or researchers from each of the twenty-two UKCRC registered trials units (in 2012) were contacted, along with six prominent international researchers who were known to have an interest in adaptive designs, and two large UK funding bodies for cancer trials (CRUK and NIHR HTA). If a statistician from a trials unit did not reply, a second statistician was approached. In total, replies were received from eighteen of the trials units, five of the international researchers and both of the funding bodies. However, only CRUK was able to provide high level information on trials that had applied to be amended; the NIHR HTA declined for confidentiality reasons. In addition, relevant trials and ongoing research developments in the area continued to emerge when presenting or discussing this work at national and international conferences, workshops and seminars, which aimed to target wide audiences of trialists.

2.2.2 Literature on the methodology of adding arms

Following the methods described in Section 2.2.1, there were 55 hits from MEDLINE, of which 6 were identified for full review; 88 hits from EMBASE, of which 5 were

identified for full review; 30 hits from the Science Citation Index related to 'Statistics and Probability', of which 10 were identified for full review; 42 hits from the Cochrane library, of which 1 was identified for full review; and 11 hits from ProQuest, of which 2 dissertations were identified for full review. Literature was discarded immediately if there was clearly no mention of adding a new arm to a trial, for example because the trial has an adaptive design but incorporates something other than a new treatment arm, such as treatment selection; if relevant words were near one another, such as "surgical *treatment* can relieve pain. *Additional* benefits…"; or other aspects of the design were deemed flexible and the abstract included a key word such as 'additional' in an unrelated phrase. In total, seventeen distinct publications, abstracts or dissertations were identified for full review. Ten of these were determined not to be directly relevant because: they were based on Bayesian methodology; they discussed potential to add arms in future but with no statistical considerations discussed; or they discussed early phase dose-response, platform or response adaptive randomisation designs. Only seven publications were identified which discussed methodological considerations when adding an arm to an ongoing confirmatory trial within the scope of this research. These were reviewed in detail to assess and summarise research previously carried-out on this topic, and the recommendations or methodology discussed. It was noted that recently there have been increasing numbers of methodological publications on platform type designs, in which arms are added and dropped in a perpetual manner as part of a master protocol[12, 13], but these are primarily based on Bayesian methodology, often in exploratory rather than confirmatory studies, and therefore where this is the case they were not considered within scope. A summary of the seven identified publications now follows.

Phillips et al. (2006)[14] summarise discussion points on adaptive designs from the PSI Adaptive Design Expert Group. Within the paper there is a brief paragraph stating that it is possible to add new treatment arms, although no details or relevant considerations are provided. Three references are given, two of which are within scope: one is methodological (Hommel 2001)[15]; and another is a practical results paper (van Leth et al. 2004)[16], discussed in Section 2.2.3.The third is out of scope as it is based on exploratory dose finding endpoints[17].

Hommel[15], along with two of the other papers identified, Posch et al. (2005)[18] and Bauer et al. (2008)[19], mention adding an arm or hypothesis as being possible within a flexible framework. The primary purpose of the papers are to discuss methodology for analysing trials in order to control error rates and prevent bias when a general mid-trial

design adaptation is made at an internal interim analyses. Methods are based on adaptive combination test principles, which analyse the data within stage and then combine the information over the stages in order to control for use of the interim data informing the ongoing design. These methods are used in combination with multiple testing adjustment for multiple hypotheses. Analysis of trials over stages when adding an arm forms the basis of Chapter 5, in which these methods are reviewed and assessed in detail.

Elm et al. (2012)[20] provide a highly relevant paper on "flexible analytical methods for adding a treatment arm mid-study to an ongoing clinical trial", which considers adding an independent treatment based on external considerations. The main aim is to compare methods for analysing continuous outcome data over the trial stages before and after the amendment, accounting for potential differences in patient cohorts. The trial design has an adjusted allocation ratio so that all three arms complete recruitment at the same time with the same patient numbers, and analysis compares the new experimental treatment to all control patients rather than only those recruited concurrently. Whilst their work is highly relevant to this research, there are a number of assumptions that limit their conclusions, and may not be realistic to many trials in practice. The limitations are detailed in Section 5.3, and their work is extended in Chapter 5 to increase the generalisability of the outcomes.

Sydes et al. (2012)[21] discuss 'STAMPEDE', an ongoing multi-arm multi-stage (MAMS) randomised, controlled trial in prostate cancer, designed to be able to drop and add arms throughout the recruitment period. The publication is not written as a general guidance document, but presents trial specific methodological and practical issues for this situation. At the time of publishing, a new research arm had been added to the existing control and five experimental arms, based on the same parameters and targets designed at the outset to ensure appropriate power for the new hypothesis. The trial has a pragmatic design in which only concurrent control patients are used as comparators to patients randomised to the new treatment, and since the experimental arms are not formally compared against each other, no type I error adjustment is made for multiplicity. That is, only the PWERs are controlled and not the FWER. Since this publication, four further treatment arms have been added, and the control group has changed to be one of the original experimental therapies based on a positive result from an initial comparison within the trial[22]. Results are reported as the data reaches maturity for each planned comparison, which is prior to the close of recruitment to the new arms. The analysis methods are not discussed, but in a publication on the

outcomes from one of the first comparisons, the data is pooled across the whole recruitment period, and the analysis is “stratified by time periods defined by addition of a new research group or end in recruitment to an ongoing research group”[23]. Whilst the literature relating to this trial is very interesting and shows that adding experimental arms mid-trial can be successful in practice, it does not provide comprehensive guidance to researchers wishing to add an arm in their own situation.

Wason et al. (2012)[24] make general recommendations for MAMS trials, including a section on adding treatment arms at planned interim analyses. The example is theoretical, based on continuous outcomes, and focuses on strong FWER control due to multiple arms and analysis points, adjusting the existing group sequential stopping bounds to account for the additional hypothesis. This methodology is not applicable for amendments based on external data that were not planned at the outset, and therefore is not directly relevant. The discussion argues against only controlling the PWER rather than considering the FWER where there are multiple hypotheses because this situation is “conceptually quite different to running a series of separate trials” and strong FWER control is required for confirmatory claims. This is considered in detail in Chapters 3 and 4.

In addition to the seven publications summarised above, two text books were identified with chapters that made reference to adding an arm to an ongoing trial[25, 26]. The chapters were contributed by Hommel and Posch respectively, and contain similar ideas on the analysis methods to the publications discussed above.

2.2.3 Practical examples of ongoing trials in which an arm has been added

Following the methods presented in Section 2.2.1, 38 unique trials were identified for further review as being possibly relevant. Each of these were assessed, with trials teams contacted where appropriate, and nine trials were identified to be within scope. Reasons for exclusion included that the amendment was planned but not implemented, the trial was early phase with exploratory outcomes, Bayesian methodology was used (usually in early phase designs), the trial included a whole new randomisation, or existing arms were modified sometimes resulting in a factorial design. Of the relevant examples, most have results published in high-impact journals, some of which went on to change clinical practice, and two are still ongoing having recently added arms in

MAMS settings. All of these trials had obtained appropriate ethical and regulatory approvals.

AML16[27], STAMPEDE[21, 22] and CompARE[28] are all cancer trials with MAMS or platform type designs assessing survival as a primary outcome with intermediate outcomes for early stopping, each having added new arms. AML16 in acute myeloid leukaemia is described as a 'pick a winner' trial, in which arms could be introduced on a rolling basis at any time. It includes two phase II assessment points for each primary comparison to determine continuation of the experimental therapy to phase III. The non-intensive randomisation was open to recruitment for just over 5 years from 2006 to 2011. It included 4 experimental arms against a control arm from the outset, but two arms were dropped at interim, and one further experimental arm was then added to the randomisation 4 years into the recruitment period. Since the new arm had recently been added when the trial closed, the new assessment was carried into another trial that is currently recruiting until 2019, and the analysis of the new trial will include the relevant AML16 patients. Publications report pairwise comparisons arising from this trial, with the other comparisons in the same protocol briefly mentioned, and no multiplicity adjustment to the analysis[29]. STAMPEDE (introduced in Section 2.2.2) is still open to recruitment and has currently added five arms; 6, 7, 8 and more than 10 years after the trial opened. All original experimental arms have closed, and the control therapy has been modified. Individual comparisons are reported as the data matures, and the trial has a large number of high impact publications and has changed practice for treatment of prostate cancer (http://www.stampedetrial.org/media-section/publication-repository/). CompARE evaluated three experimental arms against one control in oropharyngeal cancer from the outset, and in 2017 one arm was dropped for futility, and another added. The statistical error rates were reported to have been controlled for both the existing and new hypotheses, and the amendment was reviewed and approved by the funder (CRUK) and drug company, however because this amendment is very recent, there is no further information available.

The 2NN trial[16] was a large international phase III trial in HIV published in the Lancet, 2004, designed to assess a binary composite outcome of treatment failure. It was initially a three-arm trial, with the new arm being a different dosing schedule of the control arm, added 5 months into recruitment. The new arm became the control in the primary comparisons, and an investigation of the original versus new control therapy was included to assess superiority of the original schedule. The overall sample size was reduced from 1350 to 1200 when the new arm was added, and the allocation ratio

amended from 1:1:1 to 1:2:2:1 with the new control and primary experimental arm having the higher proportions. Although the trial was designed to require 450 patients per arm, the final numbers were 387 to the new arm and 400, 220 and 209 to the others. Initial logistic regression models included trial stage and treatment*stage interaction as covariates, and when these were not significant, the data were pooled over the stages in all analyses without adjustment. There was no indication of the power for this interaction test or for the primary comparisons. A Bonferroni adjustment was applied for having four multiple primary hypotheses. The authors refer to the amendment as a drawback, and state that the overall efficacy estimates should be interpreted with caution, but believe that the main conclusions of the study are robust.

CATIE[30] was a double-blind, 4-arm, phase III, trial in schizophrenia published in the NEJM, 2005, with a primary endpoint of 'time to treatment discontinuation'. A 5th arm was added after 1 year of the 4 year recruitment period, due to it receiving FDA approval and therefore emerging as being of interest within the population. The randomisation continued with even allocation. Patient numbers were not increased to the trial as a whole, impacting slightly on the power for all the trial comparisons. Adjustment for multiple comparisons was planned using the Hochberg method, but did not include the new hypothesis in order not to further reduce the power. Instead the evaluation of the new treatment was a secondary comparison against concurrent comparators only, with approximately 58% power. The trial statistician commented (personal communication) that they "had a limited budget and could not add enough patients for good power, and yet it was felt by investigators if it was not added then the study might be missing an important evaluation".

SANAD[31] was an HTA-funded 4-arm non-inferiority phase III epilepsy trial published in the Lancet, 2007, with joint primary endpoints time to treatment failure and time to remission of seizures. An unplanned 5th arm was added after 19 months of a 56 month recruitment period. There was no increase in overall trial size, and the randomisation ratio remained even (1:1:1:1:1), so the new arm included fewer patients (210 compared to 378 in the other four arms). Pairwise comparisons were carried out between all trial arms, but only concurrent patients were included for the analyses of the new arm. There was no adjustment for multiple testing. The lack of power for the new comparison cast doubt on the finding that the treatment was not non-inferior due to a wide confidence interval, with the discussion stating that "the smaller numbers of patients available to the comparison reduce the statistical power and we could not conclude that they are equivalent".

N9741[32, 33] and N0147[34] were two large phase III, US regulatory colorectal cancer trials published in the JCO, 2004, and JAMA, 2012. During the trials there were a number of treatments added and dropped due to new evidence coming to light and safety concerns or futility. Each time an amendment was made, recruitment was paused, sometimes for up to a year. The trial publications report on the comparisons that remained in the trial until the end, and briefly mention that several changes were made early in the study that do not materially affect the results. In personal communication with the trial statistician it was determined that the type I error was adjusted for multiplicity due to having multiple comparisons at the end, but there was no adjustment for treatments that were dropped or for having different stages. The power was ensured to be adequate for all primary comparisons, only concurrent comparators were used, and the allocation ratio remained equal throughout. Interim analyses were included but the addition of new arms was prior to any formal analyses of efficacy. The changes were “remarkably logistically complicated” since they were registration trials involving multiple pharmaceutical companies, and the FDA were involved in discussions. However the outcome of the trials changed practice and they were very important and successful.

A pulmonary tuberculosis trial (TB trial) in India[35], published in PLOS ONE, 2013, was designed to be a three arm trial to enrol 400 patients on a 1:1:1 ratio. However, one of the experimental drugs was not available at the outset of the trial, and therefore it opened as a 1:1 randomisation with the additional experimental treatment being added after a year of recruitment with a 1:2:1 ratio to compensate for the delay in recruiting to the new experimental arm. The trial was analysed as though all arms had been included from the outset as originally planned, with no consideration for non-concurrent control patients or trial stages, and no multiple testing adjustment.

This review confirmed that there is the desire to add new experimental arms into ongoing confirmatory trials, although this was not always done without undermining the trial’s statistical validity. This suggested that further research and guidance into statistical considerations when making this type of amendment was needed.

2.2.4 General guidance on adaptive designs

The assessment of summary, regulatory, guidance and review documents on flexible or adaptive designs in general was undertaken firstly to investigate whether there was

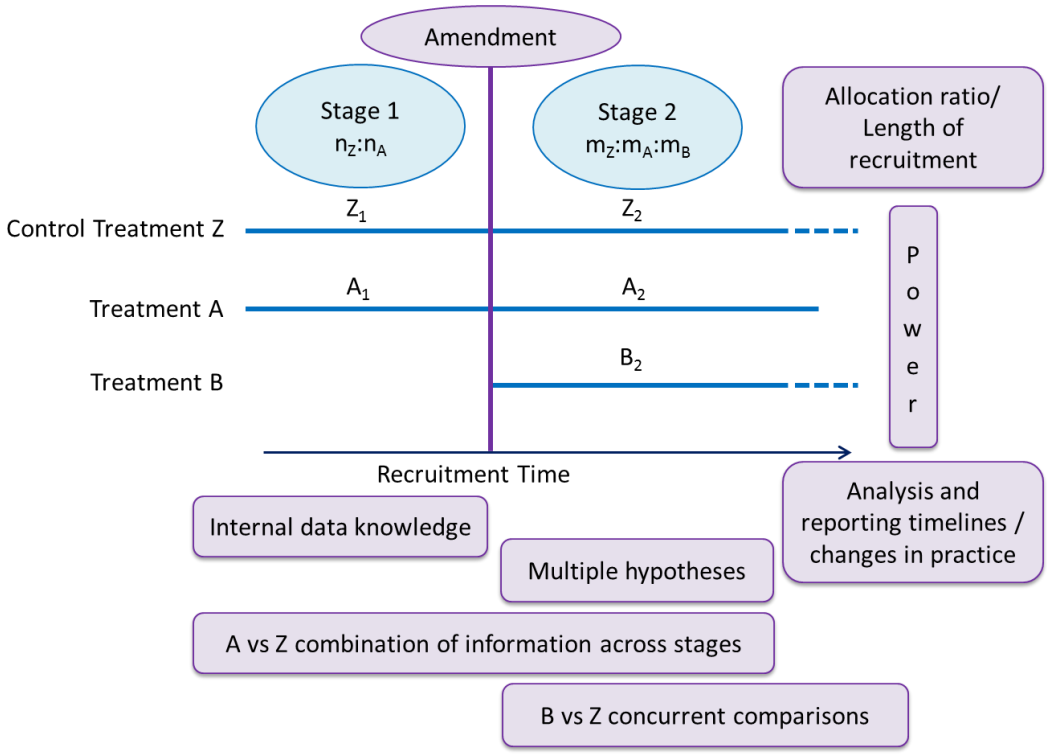
any mention of adapting a trial by adding an arm, and secondly to identify methodological themes and topics that are investigated within the field of adaptive designs in general in order to determine those that are relevant to this thesis. A MEDLINE (Ovid) search, with criteria provided in Appendix B, had 344 hits, of which 29 publications were identified for full review. In addition, citations from key papers and authors revealed other major contributions to the field. Manuscripts and texts on adaptive designs have been studied over the course of the research, which included the assessment of: two textbooks[25, 36]; regulatory documents from the FDA (Food and Drug Administration)[37, 38] and EMA (European Medicines Agency)[39, 40]; summaries from four expert adaptive designs working groups[4, 14, 41, 42]; and a number of publications summarising views on adaptive designs, of which the most relevant were considered in detail[5, 43-48].

None of these documents discussed the addition of a trial arm other than briefly mentioning that it may be possible in a flexible framework. The statistical considerations discussed within each of these documents were extracted and summarised to form a list of key statistical concerns when adapting a trial. There was a great deal of overlap, as most of the documents included the same points. The following were extracted for discussion with the supervisory team: type I error control due to combination of information across trial stages, multiple analysis points and multiple hypotheses; type II error (power); consistency of the treatment effect over trial stages; analysis methods accounting for the adaptation; estimates of treatment effects and confidence intervals; interpretability of results; introduction of statistical or operational bias; pre-specification of adaptations or methods used; definition of the patient population to whom the results apply; blinding issues; unintentional release of interim data; optimal allocation ratio; comparisons to concurrent control data; full documentation of all decisions and reporting requirements. The logistics of implementation was also a key consideration, although not necessarily statistical. Those determined to be most relevant when adding an arm within the scope of this research are included within the summary below.

2.3 Key statistical considerations when adding a treatment arm to an ongoing trial

The literature review as previously described identified a number of statistical considerations of relevance when amending ongoing clinical trials by adding a new treatment arm based on external evidence. The main considerations identified are discussed in the following sections, and are illustrated in Figure 2-1.

Figure 2-1 Illustration of a trial timeline in which an arm is added. The trial has two distinct stages and the key statistical considerations are displayed.



It was noted to what extent the statistical considerations that were identified were addressed in trials where an arm was added in practice. Of the nine trials that were identified to have added an arm, eight have reported details (Section 2.2.3). A summary of the considerations implemented within each of these trials is provided in Table 2-1 at the end of this chapter.

2.3.1 Family-wise error rate control due to multiple primary hypotheses

When a new arm is added to a trial, multiplicity concerns are introduced due to multiple primary hypotheses within the same protocol, and comparisons with a shared control group. There are conflicting views within the literature on whether strong control of the FWER is needed in this case, or whether it is adequate to control the PWER for each experimental arm versus control. Some literature[21, 49] argues that if the experimental arms would have otherwise been assessed in different protocols and are only being tested in the same trial for efficiency purposes, this is analogous to running separate trials and therefore FWER control is not necessary. However, others argue that multiplicity issues arise due to multiple use of the same control population and the efficiency of testing multiple hypotheses within the same protocol, and that strong FWER control may be a regulatory requirement for confirmatory claims[24, 37, 50, 51]. This issue is investigated in detail within Chapters 3 and 4 in order to be able to make recommendations.

Using the equations in Section 2.1.3.3, a basic Bonferroni multiple testing adjustment applied to a trial with two hypotheses, such that the significance level becomes 0.025 for each hypothesis, would decrease the power from 90% to around 84% for the individual hypotheses. Of the eight confirmatory trials that have been identified to have added an arm, the family-wise error rate was strongly controlled for multiplicity due to having more than one primary comparison in four, and was not in four (Table 2-1). This illustrates the discrepancies in practice.

2.3.2 Analysis methods to account for multiple stages

The primary statistical concern in most methodological publications discussing adaptive or flexible designs is to ensure that there is no increased error or bias due to the adaptation of the design features, creating a distinction in the stages before and after the amendment. Of the seven relevant methodological publications described above, all but the practical paper on the STAMPEDE trial[21] primarily focus on analysis methods. In all cases, the analysis is performed within each stage and a p-value combination approach is used to derive an overall outcome. Typically in adaptive designs, the amendment is based on interim data that are internal to the trial being adapted, so it is necessary to analyse the stages individually and use combination methods to derive an overall p-value in order to control the type I error rate in the final analysis[52]. In the scenario considered in this thesis, however, it is assumed that an arm

is added based on external information, and therefore the second stage isn't informed by interim data. Consequently, a p-value combination approach may not be necessary. However, changes to the trial design could affect the trial population, causing a 'stage effect'. Even if the eligibility criteria are unchanged, the new treatment could be more or less appealing in terms of efficacy and/or toxicity, and therefore attract different types of patients to the different stages causing a population drift. If data are simply pooled across the trial in the analysis, ignoring the stages, this might lead to a stage effect bias due to the treatment effects being different in each stage. Referring to Figure 2-1, stage effects would only affect the comparison for the original experimental arm (A vs Z), since the new arm B only exists in the second stage. However if multiple testing adjustment is also required this could impact the trial analysis as a whole. Analysis methods are comprehensively investigated in Chapter 5, both assuming FWER adjustment for multiple hypotheses is not necessary, and also in combination with FWER adjustment methods.

In contrast to the methodological literature primarily focusing on analysis by stage, none of the trials adding an arm in practice analysed the results by stage. Only STAMPEDE reported adjusting for stage within multivariable analyses. The 2NN trial tested for a stage effect within a multivariable analyses, but when this was not significant, simply pooled the data without adjustment in further analyses. The contrast between methodological literature and practical implementation illustrates the need for further research and guidance in this area.

2.3.3 Concurrent control data

If there is a shift in the patient population in the second stage, after the new arm has been added, the control data collected prior to the amendment may lead to different results to that collected after. For this reason, the control data collected prior to the amendment may not be an unbiased comparator for the new arm. One of the methodological papers on adding arms stipulates the use of concurrent controls[21]. The others do not discuss this directly, but by applying methods for analysing the data by stage and then combining the p-values, this is implicit. A test for heterogeneity across trial stages is unlikely to have enough power to be meaningful. A minority of literature on adaptive designs considers the use of non-concurrent controls, although these are usually in early phase trials and/or based on Bayesian methodology[12, 13] and further investigations are required to assess or adjust for a potential population drift. In general, the use of non-concurrent control data in confirmatory trials goes against the

principles of a ‘Randomised Controlled Trial’ in the same way as historical control data is not generally acceptable for a confirmatory hypothesis test, and is unlikely to be accepted as good practice. Altman (2018)[53] confirmed this with relation to adding or dropping an arm in particular, saying “between-group comparisons are valid only when restricted to participants who were randomised concurrently”. Therefore, it is recommended that only concurrent control patients are used.

Six of the eight trials used concurrently recruited control patients only. In the 2NN trial a stage effect was assessed before analysing the new hypothesis against all control patients, although power for this test was likely to be very low. The TB trial used all controls, including those recruited prior to the amendment, without further consideration.

2.3.4 Power recalculation

When a new treatment is included within a confirmatory trial, care needs to be taken that there is adequate power to assess the primary hypothesis associated with that treatment. The power to assess the original hypotheses must also be preserved. In addition, if an adjustment is made to control the FWER due to the new hypothesis, this will reduce the power for each of the individual hypotheses[26]. If the hypotheses are powered and considered independently from one another (rather than as a family informing a single claim), it would be necessary to increase the sample size for all arms in the trial in order to maintain appropriate power. See Chapter 4 for further detail.

Whilst it seems obvious that a confirmatory trial should always be appropriately powered, this wasn’t always the case in practice. Three of the eight trials were underpowered for some or all primary comparisons (Table 2-1), and all of these reported this as a limitation as it compromised the ability to report clear trial outcomes. Some of the trials reduced the sample size for all trial arms in order not to inflate the sample size for the trial overall. For example, in the CATIE trial in which a 5th arm was added 1 year into the 4 year recruitment period with no adjustment to the overall sample size, it is estimated based on the information available and the equations in Section 2.1.3.3 that the power for the original hypotheses to assess time to treatment failure decreased from 85% to approximately 80%, and the power for the new hypothesis was only 58%. None of the trials that implemented FWER control to adjust for the new hypothesis inflated the sample size to account for this.

2.3.5 Areas to improve efficiency: allocation ratio, length of recruitment and time to amendment

The allocation ratio and length of recruitment to each treatment arm could be adjusted to improve efficiency in terms of total number of patients required and time taken to answer the primary hypotheses, and these need to be carefully balanced considering the requirements for the trial.

Dunnett (1955)[54] showed that the optimal allocation to the control group in multi-arm trials is approximately the square root of the number of experimental arms, in order to minimise the total numbers of patients required. Wason et al. (2012)[24] investigated the optimal allocation ratio in MAMS trials with varying numbers of experimental arms and numbers of stages, allowing for early stopping. They found that "Although efficiency (in terms of maximum sample size) can be gained by deviating from an optimal allocation to each arm, the gain is generally fairly small". This was due to the chance of experimental arms being dropped at each stage, suggesting that optimal allocation is not necessarily straightforward where the number of treatment arms varies throughout the trial. Patient acceptability also needs to be considered, since the more attractive a trial, often perceived as being related to the higher the chance of receiving an experimental treatment, the better recruitment rates tend to be. So long as the error rates are controlled for each hypothesis based on the number of concurrent patients included in the analyses, the allocation ratio should be determined on a trial by trial basis with consideration of what is most appropriate for each particular case.

Elm et al. (2012)[20] believe that the allocation ratio should be adjusted so that all arms complete recruitment at the same time to ensure maintenance of blinding and to prevent a 'Stage 3' effect due to dropping the original arm. This may lead to very unbalanced randomisation allocations depending on when the new arm is added, which may not be desirable. Other trialists such as those who design MAMS trials, however, advocate that arms can be added or dropped throughout the trial at different times as required, which could lead to a rolling design with multiple trial stages. These trial stages may need to be accounted for in the analysis if there is a population shift due to closing a completed arm, in the same way as when an arm is added. This also creates further logistical complications due to multiple major trial amendments. These issues are not insurmountable however, and are discussed in more detail in Chapters 5 and 6.

Three trials deviated from a 1:1 allocation ratio: the 2NN trial changed from its original 1:1:1 design to recruit at 1:2:2:1 after the amendment to recruit higher numbers to the more important arms; the TB trial added the new arm with double the ratio of the others to compensate for the delay, which they did because they were not restricting comparisons to concurrent controls; and STAMPEDE recruited more to control initially because “It is more efficient to have more patients allocated to the control arm when there are more research arms co-recruiting”. Once some arms had been dropped and there were fewer experimental arms in STAMPEDE, the new comparisons were randomised with even allocation, although the allocation ratios remained constant within any given comparison. All but STAMPEDE stopped recruitment at the same time in all confirmatory arms, excluding those that were dropped early for futility at interim analyses.

Another area of efficiency when adding an arm is the time to making the amendment from opening the trial. The earlier the new arm is added, the more beneficial it is likely to be in terms of savings in patient numbers and associated costs due to sharing control patients. However, there may still be substantial savings in time and cost by amending a trial at any time over setting up a new trial. The majority of trials that added an arm did so relatively early in the trial. However, AML16 and STAMPEDE both added arms towards the end of the planned recruitment periods; AML16 carried the randomisation forward to a new trial, and STAMPEDE continues to add and drop arms on a rolling basis. In STAMPEDE it was estimated that the cost of adding a new arm was approximately 60% as much as for a separate stand-alone trial[21].

2.3.6 Changes to the control therapy

Potentially a new therapy may receive approval during the course of the trial which would make the existing control therapy inferior to standard of care and therefore unethical, and would also impact on the relevance of the trial outcomes. None of the methodological papers mention changing the control group therapy, but for very long or rolling trials where new treatments are rapidly emerging, this is likely to arise. It could happen in any long trial, but in rolling trials where new arms are still recruiting when earlier hypotheses are planned to report, this is a particular risk. Changing the control therapy would require a new power calculation and recruitment of concurrent patients between the relevant treatments. In some situations, this might be something that could be pre-empted and planned for when adding an arm by ensuring the arm is concurrently compared to both the existing and potential new control arms, if feasible,

as discussed in Chapter 6. However in many cases this might not be possible, but this risk needs to be carefully considered when determining whether to add a new arm and choosing the appropriate control treatment.

Three of the trials changed the control group therapy when adding a new arm. The 2NN trial amended the primary hypotheses so the new arm became the control group for all primary analyses. In contrast, the N9741 trial changed the control group for the whole trial to one of the existing experimental arms because of a change in the standard of care, requiring the original control arm to be dropped. This occurred a year into the trial after 300 patients had been randomised, but since the new control therapy had been present from the outset, it was concurrent to all experimental arms. The STAMPEDE trial used a different control group for the later therapies to be added than for earlier therapies because of positive results from an early comparison leading to a change in practice. However, the control group therapy did not change mid-way through the randomisation for an experimental therapy.

2.3.7 Analysis timelines

When comparisons are staggered, such as in rolling type designs where arms could be added and dropped over time, analyses can fall at different times for different comparisons. It needs to be considered how these affect the other components of the trial. For example, if one comparison reaches analysis while another is still recruiting, could the results affect the ongoing randomisation by releasing information for a partially shared control group? If so, it should be carefully considered whether it is appropriate to add an arm in this case, since it would not be ethical to delay the planned analysis and reporting for the existing hypothesis. This is somewhat less likely in trials with survival endpoints than those with shorter term endpoints because of the follow-up period, but could still be an issue. In addition, could a positive finding lead to the control therapy being superseded, as described in Section 2.3.6, and if so could this be managed?

If there is an interim analysis prior to the addition of the new arm, information from this comparison may influence the design amendment, in which case the assumption that only data external to the trial has informed the amendment may not hold. This is particularly likely if an arm is added into a seamless phase II/III trial, as is often the case in MAMS type designs. The inclusion of information internal to the trial into the design amendment is considered in Chapter 5.

In the STAMPEDE trial, new arms were added after interim analyses had taken place and internal data had been assessed by the trial team, although this was not reported to be a necessary consideration within the analysis. In addition, some arms were still open to recruitment as the first comparisons reached their final analysis and reporting points, however the later arms were only added as the original arms had almost completed recruitment so there is relatively little overlap of shared control patients and therefore analysis sets.

2.3.8 Publication of results

It is important that all publications are clear about the entire trial design, even if just focusing on results from one of the hypotheses. All trial arms and comparisons must be detailed, whether planned from the outset or added later. This is a requirement as described in the CONSORT statement for transparent reporting of trials[55].

All of the trials identified in this review discussed the wider design in results publications. It is possible that arms were added to other trials but without featuring in publications, and were therefore not picked up in this review.

2.3.9 Logistical considerations

There are many important logistical and practical considerations discussed within the literature that need to be overcome for the amendment to be feasible. They include: applications for approvals, funding and drug supply; amendments to the protocol and patient information; data management considerations such as changes to the database, Case Record Forms (CRFs), monitoring plan and processes; changes to the randomisation system; implementation and training at centres, and centre attrition; whether to continue or pause recruitment; maintaining blinding; oversight committees and their roles; contracts and negotiations, particularly if multiple pharmaceutical companies are involved; and interactions with regulators. The logistics of adapting a trial in this way may be difficult to manage, and problems at any step could increase the time it takes to implement the amendment. However, the trials identified and reviewed in this section managed to overcome these issues in order to add new treatment arms to the randomisation. In Chapter 6 we discuss the implementation of adding new arms into an ongoing trial in practice, and describe how the logistical challenges were addressed so that the amendment was achieved.

2.3.10 Summary of statistical considerations in practice

The statistical considerations that were identified within the methodological literature have been addressed to varying extents in practice, as reviewed throughout this section. Not all considerations need to be addressed in each case, as they are dependent on the nature of the trials and their objectives. When designing or critically evaluating the results of different trials, it should be determined for each trial whether the considerations are necessary for the results to be robust, or advantageous to improve efficiency or feasibility. Table 2-1 summarises whether each consideration was addressed in the trials that were identified to have added an arm, as described above. This illustrates the differences in practice between confirmatory trials that have been amended by adding a new experimental arm.

Table 2-1 Summary of statistical considerations implemented by trials when adding an arm

	Trial							
Statistical Consideration	**AML16**	**STAMP -EDE**	**2NN**	**CATIE**	**SANAD**	**N9741**	**N0147**	**TB India**
FWER control for multiplicity of primary hypotheses	✗	✗	✓	✓	✗	✓	✓	✗
Accounted for multiple stages in analysis	✗	✓	✗	✗	✗	✗	✗	✗
Used only concurrently recruited control patients	✓	✓	✗	✓	✓	✓	✓	✗
Ensured adequate power for primary hypotheses	✓	✓	✗	✗	✗	✓	✓	✓
Varied allocation ratio from even	✗	✓	✓	✗	✗	✗	✗	✓
Varied end of recruitment for existing and new arms	✗	✓	✗	✗	✗	✗	✗	✗
Changed the control therapy for the primary comparison	✗	✓	✓	✗	✗	✓	✗	✗
Amendment clearly mentioned in results publications	✓	✓	✓	✓	✓	✓	✓	✓

Key: ✓= done, ✗ = not done

2.4 Discussion

This literature review has confirmed that very few publications have addressed the topic of how to add a treatment arm to an ongoing trial, and none have done so either systematically or comprehensively. Only a small number of trials were identified to

have added arms in practice, indicating that although this type of amendment may be advantageous, it is has not been implemented widely. Of the trials that had added an arm, some failed to adequately address the statistical issues, and suffered from lack of power and difficulties in interpretability. However, it is clear that this type of amendment is desirable and advantageous, with the statistical and logistical issues seeming by no means insurmountable.

Guidance is needed to enable amendments to add new arms to existing trials to be made with robust statistical validity. In particular, current literature is contradictory on the requirement for familywise error rate control due to the inclusion of multiple hypotheses in the same trial with some shared control data; and this is comprehensively addressed in Chapters 3 and 4. Analysis methods to account for the different stages when amending a trial by adding a new treatment arm have not be adequately addressed, and therefore it is not clear to researchers how or whether to control for 'Trial Stage' in the analysis. Analysis methods accounting for the trial stage are investigated in Chapter 5. Other recommendations are clearer, such as: only using concurrent control data; ensuring adequate power for all primary hypotheses, including accounting for any multiple testing adjustment; and ensuring all results publications clearly report the entire trial design. Some design issues need to be addressed on a trial-by-trial basis at the discretion of the trial team with consideration of the trial design as a whole. These include: choosing the allocation ratio, noting that it is not necessary for all arms to complete recruitment at the same time, although any additional stages caused by closing completed arms may need to be considered in the analysis as discussed in Chapter 5; considering whether extending the trial increases the chance of a change to the standard of care within the treatment period, requiring consideration as to whether the concurrent control treatment is likely to be appropriate for the duration of the trial; and considering the effect of timelines for each planned analysis on the other hypotheses within the trial. In addition, the logistical complexities of making this type of amendment need to be discussed and planned with the wider trial team.

The aims in the following chapters are to investigate the statistical considerations that are currently unclear in the case of adding an arm to an ongoing trial. This work is used to inform the guidance and recommendation topics that were identified here, in order that they are updated and clarified in the discussion, Chapter 7.

Chapter 3
Multiple testing adjustment in multi-arm trials with a shared control group: background, literature and existing research

3.1 Background and aims

Clinical trials are designed so they have an acceptable probability of obtaining the correct answer to their primary research objective. There are two types of error that could occur: a false positive result, in which a difference is declared where one does not exist; or a false negative result, in which no difference is declared although there is truly a difference. The chance of a false positive result is required to be most stringently and carefully controlled in order for the outcomes of clinical trials to be accepted and to be able to influence practice. The convention is to set this error to be no greater than 5%, and this is usually denoted by α.

A typical confirmatory two arm trial compares an experimental therapy against the current standard within the population of interest. In this case, the probability of falsely declaring there to be a difference in efficacy between the experimental therapy and the current standard would be set so that it does not exceed 5%. That is, there is a 2.5% chance of incorrectly finding the experimental therapy to be significantly superior, and a 2.5% chance of an inferior finding. For several reasons, it is advantageous to conduct multi-arm trials, in which a number of experimental treatments are compared to the current standard. Firstly, such trials are more efficient since they use the data collected on the control group more than once so fewer patients are required. Secondly, trial set-up times and costs can be reduced over running separate trials. Finally, increasing the number of experimental arms increases the chance of finding a successful treatment[1].

When adding a new experimental arm to an ongoing trial, this will inevitably create a multi-arm trial in which the concurrent control data can be used as a comparator for multiple experimental arms, as discussed in Chapter 2. Therefore the concept of multiple testing adjustment also arises in this situation as well as in trials which have

multiple experimental arms from conception. For this reason, appropriate multiple testing adjustment in multi-arm trials, regardless of whether all the arms are recruiting concurrently or not, is a primary consideration when adding an arm to an ongoing trial[6]. Chapter 2 highlights that there are conflicting viewpoints within the literature regarding appropriate control for the chance of a false positive error when multiple experimental treatments are included in the same protocol with shared control data. In Section 3.2, various points of view on multiple-testing adjustment from within guidance documents and key authors in the field are identified and summarised. Common multiple-testing adjustment methods are reviewed in Section 3.3 so that they can be applied within this research. The factors that cause multiplicity concerns, and how well understood they are within the literature, are broken down in Section 3.4. In Section 3.5, the most relevant literature that informs the remainder of this research is described. This chapter serves as an introduction to the existing work on this topic in order to inform the novel research presented in Chapter 4, leading to informed recommendations on the requirement for multiplicity adjustment in multi-arm trials with various research goals.

Part of the work presented in this chapter has been published alongside the work from Chapter 4 in the journal *Statistical Methods for Medical Research* (Howard et al., 2018)[8]. The publication is entitled “Recommendations on multiple testing adjustment in multi-arm trials with a shared control group”. The work contained within the publication is directly attributable to myself as first author with input from the co-authors who are all part of my PhD supervisory team.

3.2 Summary of the literature

3.2.1 Aims of the literature review

There are conflicting points of view within the literature regarding multiple testing adjustment in the situation of multi-arm trials. Some authors believe the comparisons of experimental arms against control can be treated as separate trials, so that the relevant errors to control are the pairwise error rates (PWER). Others argue that the familywise error rate (FWER) needs to be strongly controlled across all hypotheses in all cases where multiple hypotheses are tested within a shared protocol. There are further arguments that some cases require FWER control across all hypotheses, whereas in other cases PWER control is adequate. With only a few exceptions, many of these arguments are based on philosophical opinions, rather than statistical theory considering the actual effects of the shared control group on the type I error rate

compared to running separate trials. The aims of the literature review conducted in this chapter are to fully understand the arguments for and against multiple testing adjustment that inform the conflicting viewpoints; to learn about the statistical considerations that guide these arguments; and to ascertain whether there is scope to build on the body of research in order to make informed recommendations for multi-arm trials in practice.

3.2.2 Literature review methods

This literature review is intended to identify the fundamental guidance available to researchers and key publications in the field on whether to adjust for multiple testing due to multiple hypotheses in a clinical trial. It is not intended to be a comprehensive summary of the literature on multiplicity adjustment in multi-arm trials. Literature was deemed to be relevant if recommendations or viewpoints were expressed on whether adjustment for multiple hypotheses is necessary. Using the search methods described here, 93 manuscripts and text books were initially identified and abstracts reviewed, 45 of these were selected for review in further detail, of which 16 were found to be relevant. Those that were relevant are broadly categorised depending on whether they always recommend strict control of the FWER in multi-arm confirmatory clinical trials (Section 3.2.3) or whether they discuss exceptions to strict control in at least some confirmatory cases (Section 3.2.4). Literature that provides informed justification on the need for adjustment is of particular interest.

3.2.2.1 Regulatory and guidance documents

Initially, regulatory and guidance documents on multiple testing in clinical trials were identified and reviewed. Key documents were authored by: the European Medicines Agency (EMA) (2017)[56]; the International Conference on Harmonisation (ICH) of Technical Requirements for the Registration of Pharmaceuticals for Human Use (1998)[51]; and Statisticians in the Pharmaceutical Industry (PSI) (Phillips et al., 2013)[57]. Guidance documents from the Food and Drug Administration (FDA) were also reviewed but were not found to address multiplicity requirements in multi-arm trials. In addition, relevant chapters of a book entitled ‘Multiple Testing Problems in Pharmaceutical Statistics’ (Dmitrienko et al., 2009)[58] were included in the review, but it was found that they do not discuss the need for adjustment in multi-arm trials.

3.2.2.2 Database search

A literature search was conducted in MEDLINE (Ovid) to identify relevant papers and key authors in the field. The following search terms were used:

1. exp Clinical Trials as Topic/

2. ((multiplicity) or (multiple adj1 testing) or (multiple adj1 comparison*)).mp.

3. ((multi* adj1 arm*) or (three adj2 arm*) or (four adj2 arm*) or (several adj2 arm*) or (multi* adj1 treat*) or (three adj2 treat*) or (four adj2 treat*) or (several adj2 treat*) or (common adj1 control*) or (shared adj1 control*) or (correlated adj1 comp*) or (correlated adj2 stat*)).mp.

4. 1 and 2 and 3

The review was carried out in February 2014 and updated throughout the research. Sixty-four results were returned, of which five were duplicates or replies to original articles. The remaining 59 abstracts were reviewed to determine whether the papers discussed the requirement for adjustment, rather than just methods for adjustment. Nine of the manuscripts were shortlisted for full review, five of which were determined to be relevant when considering the requirement for multiple testing adjustment[49, 59-62].

3.2.2.3 Grey literature review

Further references were identified during the review of relevant manuscripts, and were also obtained and reviewed. This efficient strategy led to identification of the most commonly referenced and widely cited, relevant literature. This was an iterative process, as each time a publication was reviewed it was assessed for further relevant references. This continued until new references stopped emerging. Three review or guidance papers led to the identification of the most relevant further references: Freidlin et al. (2008)[49], Wason et al. (2014)[59], and Proschan and Waclawiw (2000)[62]. Thirty further publications were identified in this way, of which 8 were relevant[24, 63-69] and are included in the summary below.

During the search, one paper was particularly interesting as it included statistical theory considering the dependence between the test statistics due to having a common control, which is key to this research (Proschan and Follman, 1995)[63]. Therefore, articles that have cited this paper were also reviewed for relevance to see how the findings were taken forward. The article has been cited seventeen times, but no

additional papers were identified that have either expanded on the methodological work or considered in detail how the findings affect the need to control the FWER.

3.2.3 Key publications for strict control of FWER in multi-arm confirmatory trials

Of the sixteen manuscripts identified, five give the clear message that they believe multiple testing adjustment to control the FWER is necessary in confirmatory multi-arm trials, regardless of the nature of the trial and its hypotheses.

The EMA (2017) 'Guideline on multiplicity issues in clinical trials'[56] draft document states that multiple-testing adjustment is likely to be required in confirmatory clinical studies 'with more than two treatment arms': "As a general rule it can be stated that control of the study-wise type I error is a minimal prerequisite for confirmatory claims".

Bender & Lange (2001) 'Adjusting for multiple testing – when and how?'[64] state that "Adjustments for multiple testing are required in confirmatory studies whenever results from multiple tests have to be combined in one final conclusion and decision". They believe this includes multi-arm trials, even though the conclusions may differ for each experimental arm, because they are obtained within a single 'experiment'. The 'final conclusion' of the trial could be that all arms are better than the standard, and this needs to be based on rigorous control of the MEER (maximum experiment-wise error rate, equivalent to control of the FWER in a strong sense). "For example, if each new treatment is significantly different from the standard treatment, the conclusion that all three treatments differ from the standard treatment should be based on adequate control of the MEER. Otherwise the type I error of the final conclusion is not under control, which means that the aim of significance testing is not achieved." In their example they talk about "three different new treatments" rather than different doses, clarifying that they believe adjustment is required regardless of the nature and aim of the hypotheses. The views in this paper seem to be based entirely on testing more than one hypothesis in a shared protocol, and do not consider the impact of the shared use of control data.

Wason et al. (2012) 'Some recommendations for multi-arm multi-stage trials'[24] include a section discussing the importance of controlling the family-wise error rate (FWER) in the strong sense, stating that the EMA guidance on multiple testing requires type I error control for confirmatory trials. They discuss the contrary opinions that adjustment in

MAMS trials is not advocated when the different arms correspond to different treatments due to the similarity to them being compared in separate trials, however they argue that a MAMS situation is “conceptually different to running a series of separate trials”, and give the analogy of testing multiple primary outcomes within a trial for which there is consensus for requiring adjustment, since each outcome could also be tested in a separate trial.

Westfall and Bretz ‘Multiplicity in Clinical Trials’ is a chapter of the Encyclopedia of Biopharmaceutical Statistics (3rd edition, 2010)[65]. They discuss the importance of adjustment for multiplicity in general, providing arguments against those who criticise the requirement for adjustment, in order that there is “a stronger standard of evidence than the unadjusted comparisons”. They believe that multiple comparison procedures are necessary to avoid errors that prevent replication of the results, which includes “declaring effects when none exist”, “declaring effects in the ‘wrong direction’” and “declaring inflated effect sizes”. They aim to dispel common controversies for non-adjustment, including: why have penalties for efficiency of assessing multiple doses; why adjust for questions asked in the same trial but not in separate trials; and why are certain tests classed as a ‘family’. They have strong views that multiple comparison procedures are always necessary, because going against this stance is “a potentially dangerous and irresponsible message”, with negative consequences “that spurious associations are published, and that inappropriate therapies are recommended”.

The Statisticians in the Pharmaceutical Industry (PSI) held an expert group discussion on multiplicity (Phillips et al., 2013)[57]. Although they do not mandate that multiplicity adjustment is always necessary, “there was consensus that any study aiming to ‘confirm’ should take into account multiplicity” and that “multiplicity adjustments need to be considered when the intention is to make a formal statement about efficacy or safety based on hypothesis tests”. They do not discuss multi-arm studies within their paper, and instead focus on multiple endpoints, analysis timepoints and multiregional developments. However the general message is that the ‘claim-wise error rate’, defined as the familywise error rate “when the families relate to multiple clinically important endpoints that need to be described in the label”, is the most important to control. Although this is discussed in terms of multiple endpoints, it could be extrapolated that families relating to multiple clinically important hypotheses within multi-arm trials that are described in a single label may also be important to control, although this is speculative.

3.2.4 Key publications discussing exceptions to strict control in some or all confirmatory cases

Of the sixteen references identified, eleven discuss exceptions to strict control being appropriate in at least some cases of confirmatory multi-arm trials.

ICH E9 ‘Statistical Principles for Clinical Trials’ (1998)[51], Section 5.6, states: “In confirmatory analyses, any aspects of multiplicity which remain after steps of this kind have been taken should be identified in the protocol; adjustment should always be considered and the details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan.” By ‘steps of this kind’ they are referring to methods to avoid or reduce multiplicity, such as choosing one of the comparisons to be primary. This is more relaxed than the EMA viewpoint, suggesting that adjustment should be considered, but if felt not to be necessary this can be justified.

Proschan and Follman (1995) ‘Multiple comparisons with control in a single experiment versus separate experiments: why do we feel differently?’[63] examine the need for a multiple comparison adjustment when treatments are compared to a shared control within the same trial, “in terms of the different distributions of the number of Type I errors and power”. This is the only publication identified to have assessed the effect of the shared control data on the probabilities of errors. The type I error is calculated based on two, three and four arm trials assuming both independent (I) and dependent (D) hypotheses, where dependent hypotheses are defined by having a shared control group. They found that the probability of making one type I error is lower in dependent trials, but the probability of making two or more errors is higher (if the control group is “bad”). So “the conditional probability of a Type I error on one comparison with control, given that a Type I error has been made on another comparison with control, is substantially increased in situation (D)”. However, they conclude that “the difference in the distributions of type I errors is relatively small when there are not too many treatments”. They do not say definitively whether they would be in favour of adjusting or not based on this evidence. They suggest that looking at the FWER and per-comparison error rate (PCER) does not tell the whole story, and in addition baseline characteristics of control patients should be compared to other trials’ data to check that whether the control group may be “bad”.

Freidlin et al. (2008) ‘Multi-arm clinical trials of new agents: some design considerations’[49] make recommendations on the need for adjustment based on the findings from Proschan and Follman (1995). They believe that the need for adjustment should be assessed by the “relatedness of clinical questions”. If the same experimental treatment is added to different backbone regimes, or for different doses of the same treatment, there is more than one chance of success so adjustment is required. However, “For the multi-arm trial application being considered here, several experimental agents share a control arm for the purpose of improving efficiency and the trial is focused on answering the efficacy question for each drug separately; the interpretation of the results of one comparison have no direct bearing on the interpretation of the others. In this situation, we believe no multiplicity adjustment is required”. This paper acknowledges the positive correlation between the individual comparisons due to the use of the same control arm, referencing the work of Proschan and Follman: “Because of this, a multi-arm trial has a lower overall probability of any false-positive result but a higher probability of making more than one false-positive conclusion (relative to separate trials). However these probability differences are small, especially when the number of experimental arms is in a practical range (two to four arms). Therefore, the fundamental issue for the purpose of multiplicity adjustment is the relatedness of clinical questions with the statistical correlation having minimal relevance”. That is, if the clinical questions are not related, they advocate that there is no need to adjust for multiple testing.

Proschan and Waclawiw (2000) ‘Practical guidelines for multiplicity adjustment in clinical trials’[62] take a middle ground to multiplicity adjustment, advocating that adjustment should be determined on a case-by-case basis. They justify this mainly on philosophical grounds, giving some scenarios and examples. The issues to be considered include “the relatedness of the questions being considered, the number of comparisons, the degree of controversy, who stands to benefit, and the nature of the study/alternative hypothesis”. The ‘relatedness’ can refer to either: the hypotheses being part of a family of experiments assessing related therapies; or statistical dependence between the test statistics due to a common control group. In terms of the number of comparisons, they argue that there is a stronger case for not adjusting when there are two comparisons with control rather than a four arm trial comparing all pairs, due to the amount of increase of the FWER. If a company stands to benefit from having more chances at success, adjustment is likely to be needed, which might not be the case for academic trials. They believe that “the burden of proof is on the investigators to defend not adjusting”. A table is provided to summarise the situations in which adjustment is necessary. For this situation where different experimental treatments are

compared to a shared control group, it reads that discretion is required, but adjustment would usually be necessary: to settle controversy; if a single entity benefits; or if there is more than once chance of declaring benefit. They mention that “an unusual control group affects all comparisons with it”, referencing Proschan and Follman’s 1995 paper, but there are no further comments about the implications of the shared use of the control data. They conclude in the discussion that “rigid rules for multiplicity adjustment are infeasible” and that “it should be the responsibility of the principal investigator to justify it [non-adjustment]”. The ideas in Proschan’s other papers in which the need for adjustment is discussed (Proschan et al., 1994[60], and Proschan, 1999[61]) are similar to those here, reiterating that the topic is controversial with different viewpoints in the literature and no definite guidance, but that there is a positive dependence between the test statistics due to the shared control data meaning that there is greater chance of more than one type I error than in independent trials.

Cook and Farewell (1996) ‘Multiplicity considerations in the design and analysis of clinical trials’[66] believe that the p-value represents strength of evidence, and should be interpreted by its value as part of the decision-making process, rather than as an absolute, artificial, cut-off for decisions. Clinical significance is also important, and statistical evidence alone is not sufficient to influence behaviour. Therefore less emphasis should be based on the exact cut-off of 0.05, but the results should be interpreted considering the questions of main interest. They believe that the concept of multiple testing adjustment is therefore non-technical. If hypotheses are defined a priori and there are a ‘reasonably small number’ then they argue that no type I error control is needed. This is because “in clinical trial designs formally based on two or more responses, multiplicity adjustments may not be necessary if marginal, or separate, test results are *interpreted marginally* and have implications in *different aspects* of the prescription of treatments”. That is, if the outcomes inform different claims of effectiveness and are interpreted separately, the ‘experimental’ type I error rate may not be meaningful. In multi-arm trials, the contrasts of primary interest should be defined in advance based on the reasons for including the treatment arms. If these are not excessive, “then it is not reasonable to impose constraints solely to control the experimental type I error rate. Each contrast can be considered individually with separate p-values or type I error rates”. However, if the main contrasts of interest are not defined a priori and all pairwise comparisons are planned, multiplicity adjustments are a reasonable price to pay for performing an excessive number of tests.

Rothman (1990) ‘No adjustments are needed for multiple comparisons’[67] offers a philosophical rationale for no adjustment based on the probability that a chance finding is in fact related to something unobserved other than chance. However, his points are abstract and do not consider adjustment in terms of observed error rates, and his rationale has not been cited in any other key literature discussing adjustment.

O’Brien (1983) ‘The appropriateness of analysis of variance and multiple-comparison procedures’[68] is a short Reader Viewpoint in which he describes common literature that recommends an ANOVA to test the hypothesis that all populations (trial arms) are identical (assuming normal distributions). If the ANOVA suggests no difference, pairwise comparisons are not recommended in general “due to the likelihood that at least one will be statistically significant by chance”, and if comparisons are made following a significant ANOVA result, a multiple testing adjustment should be used. O’Brien argues against this strategy, since researchers may want to control the comparison-wise error rate rather than the experiment-wise error rate, stating “The logic behind this view is that two investigators who have collected exactly the same data on two populations (A and B) should arrive at the same conclusions when comparing these two populations, despite the fact that additional data on other populations may have been collected by one of the investigators”. He notes that investigators are penalised for making more effort in asking more questions, and that “the investigator should be free to define the questions he wishes to address and to avoid assumptions which are unproven”. Therefore, regardless of the outcome of the ANOVA, O’Brien advocates that there are situations in which it may be appropriate to justify some pairwise comparisons without adjustment.

Wason et al (2014) ‘Correcting for multiple-testing in multi-arm trials: is it necessary and is it done’[59] review some of the contrasting views on multiple testing adjustment and assess what is done in practice. They split the requirements by trial type: exploratory; confirmatory with distinct experimental treatments; and confirmatory with arms being doses or regimes of the same treatment. Unlike Wason et al. (2012)[24], they do not mandate multiple testing correction in all circumstances. They summarise some of the different views within the literature, highlighting that conflicting viewpoints exist on this topic and providing some relevant references. A small literature review of multi-arm trials that were published in four high impact journals during 2012 was carried out, and the level of adjustment was summarised by trial phase and by whether the treatment arms are distinct or assess the same therapy. The review identified 59 trials, and showed that around half of the trials adjusted for multiplicity (49%), with 55% of

exploratory trials adjusting compared to 46% of confirmatory trials. 63% of trials with multiple doses or regimens adjusted compared to 35% of trials with distinct treatments (6/20 (30%) of confirmatory trials). The discussion includes the authors' opinions, which are based on mainly philosophical arguments, and are that: strict correction is not required in exploratory trials, although the final FWER should be reported; adjustment is required in confirmatory trials with related treatments, particularly where a single company stands to benefit; and in confirmatory trials with unrelated treatments the literature is unclear, but guidance from regulators suggests that adjustment is necessary. They conclude that more guidance from regulators is required.

Hung and Wang (2010) 'Challenges to multiple testing in clinical trials'[69] is written with regulatory trials in mind. There is no specific mention of multi-arm trials with shared control data, although some issues are relevant. Rather than mandating strong control, they discuss using common sense to define "a relevant family of hypotheses for which the type I error needs to be properly controlled". They discuss the use of a "clinical decision tree", determined in advance, to decide what aspects need to be protected from type I error inflation.

3.2.5 Discussion on the literature review

There are clearly conflicting viewpoints within the literature on the appropriate way to handle multiplicity in confirmatory multi-arm trials with shared control data, with some authors strongly of the opinion that the type I error needs to be controlled across all hypotheses, and others believing that there are at least some cases where this is not necessary. Even key regulatory guidance documents do not offer the same advice, although scientific advice can be sought from regulators on a case-by-case basis. The discrepancies within the literature leave trialists unclear on whether multiple testing adjustment is appropriate and necessary, and therefore lead to differences in practice.

The majority of the arguments are philosophical, and do not consider the effect on the FWER of having shared control data or making the adjustment compared to error rates when running separate trials. None of the papers that advocate strict control at all times give a definitive statistical justification for this view. The rationale given is generally that adjusting is the safer option to ensure that the chance of recommending an inappropriate therapy is not increased. Wason et al. (2012)[24] argue that assessing multiple hypotheses within a single protocol is equivalent to assessing multiple primary outcomes, for which adjustment is generally agreed to be necessary. However, this

analogy is not equivalent because multiple outcomes are measured on a group of patients receiving an experimental therapy, therefore increasing the chance of at least one false-positive finding with relation to that particular therapy. While this argument may have some relevance to multi-arm trials with experimental arms that test similar therapies, for example different doses or combinations of the same treatment, it does not necessarily hold for trials in which the experimental therapies are distinct. In contrast, a small number of authors do not believe that adjustment is ever required, because this is the same as would have been the case if the hypotheses had been assessed in separate protocols, and researchers are therefore being penalised for efficiency. Some of these articles have received replies and comments disagreeing with their viewpoints, further highlighting the differences of opinion on this topic.

The crux of the opinions can often be attributed to the definition of 'family' when considering whether the FWER needs to be controlled. Those that advocate adjustment at all times believe that all hypotheses belong to a family simply because they are being assessed in the same protocol. The authors that recommend adjustment in some cases but not others consider a family to be a set of hypotheses that are related in that they contribute towards a single claim of effectiveness, and believe that if this is not the case the design is essentially running different trials but under the same protocol, and therefore adjustment is not necessary. Only one publication (Proschan and Follman, 1995)[63] calculated the implications of the shared control data on the overall error rates for the trial compared to independent trials to inform their views on adjustment. However, this work is rarely referenced or considered when discussing the requirement for multiplicity adjustment in multi-arm trials. Freidlin et al. (2008)[49] include it as part their argument for not adjusting where hypotheses are not clinically related, but no other literature was found to offer a quantitative rather than philosophical perspective on adjustment. This literature review highlights the need for further research to enable definitive guidance to be produced in this area.

3.3 Multiplicity adjustment methods

In this section, some commonly implemented methods for multiple testing adjustment are introduced, so that their effects on error rates can be assessed in Chapter 4. Recall that multiplicity is the term used to describe the increased risk of a false positive conclusion (a type I error) that arises when multiple tests are carried out on a set of data. The concern is that if multiple tests are performed, each with a small chance of

error, the overall chance of error increases such that the claims made may be a consequence of an inflated rate of false positive conclusions.

3.3.1 Familywise Error Rate (FWER)

If multiple tests are performed within a clinical trial in which the null hypotheses are true, the overall chance of making a type I error across the trial as a whole increases. If the significance level, α, is set to 5% for an individual hypothesis, so that $p < 0.05$ indicates significance, there is a 95% chance that no error has been made. If a second, independent test is also performed, there is also a 95% chance that no error has been made in that test. The overall chance of no error in the trial overall is 0.95 x 0.95 = 0.9025, so the chance of some error is now increased to 9.75%. This is known as the familywise error rate (FWER), defined as the overall probability of at least one false positive conclusion anywhere within a defined set of trial hypotheses. In order for the set, or family, of trial hypotheses to influence practice, it may be required that the FWER is controlled at an equivalent level to that for a single hypothesis.

The chance of an exact number of type I errors can easily be calculated for independent hypotheses, that is, hypotheses that are tested using entirely different populations. Note that hypotheses in multi-arm trials with shared control data are not independent, and the impact of this lack of independence is assessed in Chapter 4. Each null hypothesis has a binary outcome associated with it, and therefore the probability of errors for independent hypotheses can be described using a binomial distribution. Define *Y* to be the random variable associated with the event that a type I error occurs. In the independent case, with *m* comparisons and a probability *α* of finding a significant difference, the probability of exactly *y* type I errors across the *m* comparisons (*y* = 1,…,*m*) can be expressed as

$$P(Y = y) = \binom{m}{y} \alpha^y (1-\alpha)^{m-y} ,$$

$$\Pr(Y = y) = \left(\frac{m!}{(m-y)!\, y!}\right) \alpha^y (1-\alpha)^{m-y} .$$

Since the FWER is the probability of at least one error,

$$FWER = p(Y > 0) = 1 - p(Y = 0) = 1 - (1-\alpha)^m.$$

With two independent hypotheses and $\alpha = 0.05$ for each, it can be calculated that the FWER is 0.0975 as expected. Table 3-1 summarises the probabilities of exactly *y* type I errors and the FWERs for trials with two, three and four independent hypotheses, each with $\alpha = 0.05$.

Table 3-1 Probabilities of type I errors in trials with two, three or four independent hypotheses, with $\alpha = 0.05$ for each

	Number of independent hypotheses (*m*)		
Exact number of errors (*y*)	**2**	**3**	**4**
0	0.9025	0.8574	0.81451
1	0.095	0.1354	0.17148
2	0.0025	0.0071	0.01354
3	-	0.0001	0.00048
4	-	-	0.000006
FWER	0.0975	0.1426	0.1855

The greater the number of hypotheses, the higher the chance of at least one type I error occurring. If the FWER needs to be controlled to reduce this chance to that for a single test (0.05), there are many multiplicity adjustment procedures available. Appropriate procedures must offer strong control of the FWER, implying that the FWER is controlled regardless of whether the null hypotheses are all true or not. In order to assess multiplicity adjustment in multi-arm trials, some of the common procedures that strongly control the FWER are described here. Section 3.3.2 includes procedures that are simple in that they do not take account of any correlation between the comparisons which may be present due to having a shared control group, and Section 3.3.3 includes those that do account for correlation structures due to the dependency of the shared control group.

3.3.2 Common simple multiplicity adjustment methods

Here the most commonly used multiplicity adjustment procedures that do not account for correlation between comparisons due to having a shared control group are described. Section 3.3.2.1 summarises a single-step method whilst Section 3.3.2.2 introduces stepwise closed testing methods.

3.3.2.1 Single-step Bonferroni method

The simplest multiplicity adjustment method is the nonparametric Bonferroni adjustment, which is a single-step method that is known to be conservative but applicable in all situations. The Bonferroni method states that if you are performing *m* comparisons (*j = 1,…,m)*, and your overall FWER is required to be *α*, each test should be run at a level of significance *α/m*. That is, the *α* is split equally between the number of tests being performed.

$$Reject\ H_{0j}\ if\ p_j \leq \frac{\alpha}{m}$$

For independent tests the Bonferroni method is conservative due to an approximation to simplify the formula, particularly for large numbers of hypotheses. Recall that if there are two tests with significance level set to 5%, the total chance of error is 9.75% rather than 10%. This is due to some occurrences when the errors may fall within the same pairs of trials, and so the overall chance of 'at least one' error is less than *αm*. The Sidak adjustment method is very similar but calculates the adjustment exactly ($reject\ H_{0j}\ if\ p_j \leq 1-(1-\alpha)^{\frac{1}{m}}$) so is less stringent than the Bonferroni correction, however as this is used less often and the difference between the adjusted significance levels is very small, only the Bonferroni method will be considered in this research.

For example, if there are three comparisons, with the FWER required to controlled at 0.05, each p-value would be assessed against 0.0167 by the Bonferroni adjustment. With the Sidak adjustment method the critical significance level for each comparison would be 0.0170.

3.3.2.2 Closed testing (stepwise) methods

Methods based on the 'closed testing procedure'[70] are less conservative and therefore more powerful than single-step methods, whilst still strongly controlling the FWER. The closed testing procedure is a hierarchical testing strategy, where hypotheses are tested in a pre-defined order and if a null hypothesis is not rejected then further testing stops and no further null hypotheses can be rejected. Due to this hierarchical ordering of testing, no further *α* adjustment is necessary[70]. The closed testing procedure can be applied when there are a closed family of null hypotheses. These consist of:

- Individual null hypotheses H_{0j}, *j=1,…,m*, where H_{0j} is the null hypothesis assessing experimental arm *j* versus control.
- All possible intersection hypotheses, where the notation $H_{0(12)} = H_{01} \cap H_{02}$.

Based on the Bonferroni-Holm method, the intersection hypothesis $H_{0(12\ldots m)}$ is rejected if $p_{(12\ldots m)} \leq 0.05$, where

$$p_{(12\ldots m)} = \min(1, mp_1, mp_2, \ldots, mp_m).$$

Other methods exist to calculate the intersection (such as using Sidak's formula), but differences are small and the Bonferroni-Holm method is the simplest. The closed testing principle has led to stepwise procedures to control for multiplicity, such as those by Holm and Hochberg which are commonly used and discussed here as exemplars of a step-down and step-up procedure respectively. Other stepwise and closed testing methods exist and are similarly based on some function of the raw p-values. They vary in complexity and some are slightly less conservative although the differences are small and do not affect the broad picture, so they are not included as part of this research.

Figure 3-1 Diagrammatic representation of the closed testing procedure

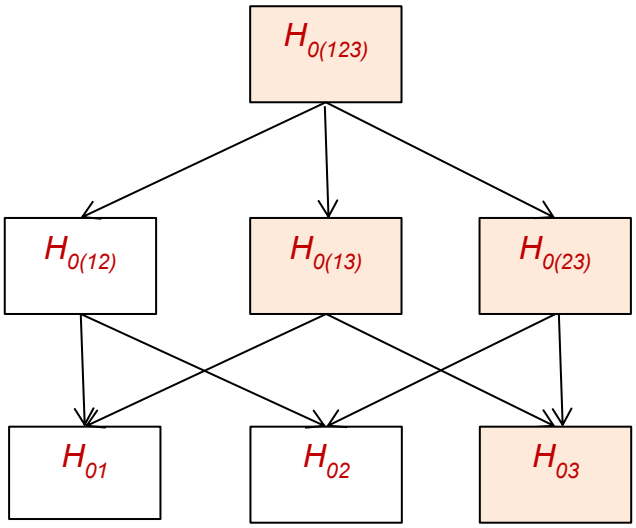


Figure 3-1 illustrates the closed testing procedure. In order to be able to reject H_{03} for example, the closed testing procedure states that every null hypothesis containing it must be rejected, i.e. $H_{0(123)}$, $H_{0(13)}$, $H_{0(23)}$, and H_{03}. That is, p_{123}, p_{13}, p_{23} and p_3 must all be less than 0.05, where $p_{123} = \min(1,3p_1, 3p_2, 3p_3)$ and $p_{13} = \min(1,2p_1, 2p_3)$ and so on. It is clear that if p_3 is less than 0.0167 (i.e. α/m), H_{03} will always be rejected regardless of the other p-values, as it would be with the Bonferroni method. However if $0.0167 < p_3 < 0.05$, it is possible that H_{03} could still be rejected if p_1 and p_2 are small, as long as one is <0.0167 and the other <0.025. This is why the method is less conservative than Bonferroni.

Holm method

The Holm method is a step-down procedure based on the closed testing principle. For *m* comparisons, the unadjusted p-values are ordered from the most to the least significant, so

$$p_1 < p_2 < \cdots < p_m.$$

Firstly, the most significant p-value is compared against *α/m*. If there is evidence to reject the first null hypothesis the next can then be assessed, and so on. Each p-value (with order *j*) is compared in order to *α/(m-j+1)*, so the stepwise procedure continues:

- The 2nd smallest p-value is compared to *α/(m-1)*
- The 3rd smallest p-value is compared to *α/(m-2)*
- The largest p-value is compared to *α*

As soon as a test fails to reject H_{0j}, no remaining null hypotheses can be rejected.

For example, in the case of three comparisons, if the p-values for H_{01}, H_{02} *and* H_{03} respectively were 0.01, 0.04 and 0.02, all null hypotheses would be rejected. This is because firstly 0.01<*α/3* i.e. 0.0167 (so from Figure 3-1 p_{123} for $H_{0(123)}$ would be 0.03), secondly 0.02<*α/2* i.e. 0.025 (so p_{12} and p_{13} would be 0.02 and p_{23} would be 0.04), and finally 0.04<*α* i.e. 0.05. Note that for this example, with the Bonferroni adjustment only H_{01} would be rejected.

Hochberg method

The Hochberg method is a step-up procedure based on the closed testing principle, addressing the p-values from least to most significant. It is more powerful than the Holm method, but is only applicable if the p-values are positively correlated, which will always be the case with shared control data (Section 4.3). For *m* comparisons, the p-values are ordered so

$$p_1 > p_2 > \cdots > p_m.$$

Firstly, the largest p-value is compared against α. If there is evidence to reject the least significant null hypothesis, all further null hypotheses are automatically rejected. If a null hypothesis cannot be rejected, the next largest is assessed, and so on. Each p-value (with order *j*) is compared in a stepwise manner to *α/j*, so:

- The 2nd largest p-value would be compared to *α/2*
- The 3nd largest p-value would be compared to *α/3*
- The smallest p-value would be compared to *α/m*

Therefore if the largest p-value is <0.05, all null hypotheses would be rejected, even if all p-values equal 0.049. Note that the Holm method requires the smallest p-value to be <*α/m* in order to proceed, so it is clear that the Hochberg method is less conservative.

For example, in the case of three comparisons, if the p-values for H_{01}, H_{02} *and* H_{03} respectively were 0.01, 0.04 and 0.03, all null hypotheses would be rejected. This is because 0.04<*α* and therefore no further testing is necessary. Note that with both the Bonferroni and Holm adjustment methods, only H_{01} would be rejected in this example.

3.3.3 Multiplicity adjustment methods that account for the lack of independence due to a shared control group

3.3.3.1 Parametric methods (Dunnett's t)

If the tests are correlated because they share control data, this increases the conservativeness of the non-parametric adjustment methods described above. It is possible to use parametric adjustment methods to exploit this correlation in order to control the FWER exactly, therefore increasing the power. Many commonly used parametric methods, such as Tukey's Honestly Significant Difference (HSD), adjust for all pairwise comparisons. For the purposes of this research, however, we are assuming that only comparisons with control are relevant for the primary hypotheses, and therefore the most commonly used, relevant parametric adjustment method in this case was proposed by Dunnett (1955)[54]. Dunnett showed that the correlation between test statistics in this case can be quantified based on the randomisation allocation ratio, as described in 4.3.1.

In simple terms, the Dunnett's t method adjusts the Bonferroni boundaries so that the probability of observing a significant result under H_0 is exactly 0.05. In order to understand how the method works, it is assumed that the data are continuous and normally distributed, although this methodology can be applied to non-normal data based on large sample approximations. The group means of normally distributed data can be compared using a one-way ANOVA, where the overall mean square (between) estimate is the sum of the squared difference between each value and the grand mean, divided by the between degrees of freedom. That is,

$$MS(between) = \frac{SS(between)}{df} = \frac{n\sum_j(\bar{X}_{.j}-\bar{X}_{..})^2}{j-1},$$

where *n* is the number per arm, *j* is the number of arms and $\bar{X}$ represents the means. The overall mean square (within) estimate is the sum of the squared differences between each value and its group mean, divided by the within degrees of freedom, that is,

$$MS(within) = \frac{SS(within)}{df} = \frac{\sum_i\sum_j(\bar{X}_{ij}-\bar{X}_{.j})^2}{j(n-1)},$$

for *i* subjects per arm. The ratio of the mean square between and within gives an F-ratio to assess overall significance based on the CDF of an F-distribution with *j-1* degrees of freedom for the numerator and *j(n-1)* degrees of freedom for the denominator. Post-hoc tests can be used to look for differences in pairwise comparisons, based on t-tests. The mean square (within) can be used to estimate the standard error required for unadjusted pairwise comparisons for the t-tests, as follows:

$$t = \frac{\bar{X}_E-\bar{X}_C}{SE},\ where\ SE = \sqrt{MS_{within}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

Parametric adjustment methods compute the required significance level of the t-statistic such that the error is controlled exactly, after accounting for any correlation. Dunnett (1955)[54] tabulated the values of *t* (t_d) to compare the experimental treatments (E) to control (C) based on different numbers of treatments and group sizes, assuming either 1 or 2-sided tests performed at 1% or 5% significance, and with equal allocation. Statistical software extends this to allow the Dunnett adjustment to be applied to any relevant scenario.

Table 2a in Dunnett (1955)[54] gives the adjusted t-statistic, t_d, for two-sided comparisons with α=0.05. Assuming a large sample size (so that the degrees of freedom are greater than 120), t_d for 2 comparisons is 2.21. Note that with 1 comparison t_d is 1.96, as would be expected. This translates to a critical significance level for comparison of 0.0271 ($2 * \left(1-\Phi(2.21)\right)$. This is clearly less conservative than the equivalent Bonferroni critical significance level of 0.025. In the case of three two-sided comparisons t_d is 2.37 which translates to a critical significance level of 0.0178 for comparison to 0.0167 with a Bonferroni adjustment.

3.3.3.2 A step-up multiple test procedure (Dunnett and Tamhane)

The procedure described by Dunnett and Tamhane (1992)[71], also known as the Adjusted Hochberg method, was discussed in Fernandez and Stone (2011)[72] as an alternative parametric closed testing adjustment method to control the FWER where tests are not independent, in order to control the error more exactly to increase the power. Similarly to Dunnett's method, the experimental treatments are compared only to the control, and not to one another. This method is similar to the step-up Hochberg method discussed in Section 3.3.2.2, except that the critical values for rejection are adjusted to account for correlation so that the final FWER is exactly 0.05. As with the Hochberg method, there are *m* comparisons, and *j* is the iterative test in the step-up procedure *j=1,…,m.* The p-values are ordered from largest to smallest, and each is compared in turn to the adjusted critical value, such that as soon as one null hypothesis is able to be rejected, testing stops and all remaining null hypotheses are also rejected.

The key to this method is determining the critical values such that the FWER is exactly 0.05. Dunnett and Tamhane (1992)[71] solved the critical constants c_i for various numbers of comparisons, amounts of correlation and sample sizes for both one and two-sided tests, and listed them in tables. The largest p-value is compared against $(2 * (1 - \Phi(c_1))$. If not significant, the next largest is compared against $(2 * (1 - \Phi(c_2))$ and so on until one null hypothesis can be rejected. Determining each c_i must be done in order starting with c_1 up to c_m. With 2 hypotheses in a 3-arm trial, assuming a large sample, two-sided tests and a 1:1:1 randomisation, Table 2 in Dunnett and Tamhane (1992)[71] can be used to find c_1 = 1.96, so $(2 * (1 - \Phi(c_1)) = 0.05$, and c_2 = 2.223 so $2 * (1 - \Phi(c_2)) = 0.0262$. The latter significance level is less conservative than the standard Hochberg significance level of 0.025.

3.3.4 Summary of multiplicity adjustment methods

There are many different options to adjust for multiple testing in order to control the FWER when assessing multiple experimental treatments against a shared control group. These can be separated into non-parametric methods that do not account for the correlation due to the shared control group, and parametric methods that do. Some methods are less conservative, whilst others may be simpler to apply and therefore tend to be more widely used in practice. Here a broad selection of different types of methods have been described, with an example of a single-step, step-up and step-down non-parametric method, and a single-step and step-up parametric method being

selected for inclusion. These methods are investigated in Section 4.4 in order to determine the effect of the different types of adjustment methods on the probabilities of type I errors. Section 4.4 also includes figures illustrating the rejection regions for each adjustment method, which is a useful aid to understand the differences between them.

In the next section, the factors that affect the need for multiplicity adjustment methods to be applied are broken down. These are then considered in turn to assess how well understood they are, in order to determine whether any further research could add to the current body of evidence on the need for multiplicity adjustment for multiple hypotheses.

3.4 Consideration of factors causing multiplicity concerns in multi-arm trials compared to independent trials

If two hypotheses are tested in separate protocols, no multiple testing adjustment is considered necessary. However if these same two hypotheses are tested within the same protocol, whether this is due to a multi-arm trial from conception or whether an experimental arm has been added, there is uncertainty as to whether adjustment is required. In this research, the aspects of multi-arm trials that might affect the family wise error rate are considered, so that their effect can be investigated.

3.4.1 Shared control data

If two experimental treatments *A* and *B* are compared against the current standard, *Z*, the hypotheses for experimental treatment *j* (*j* = *A*, *B*) can be expressed as:

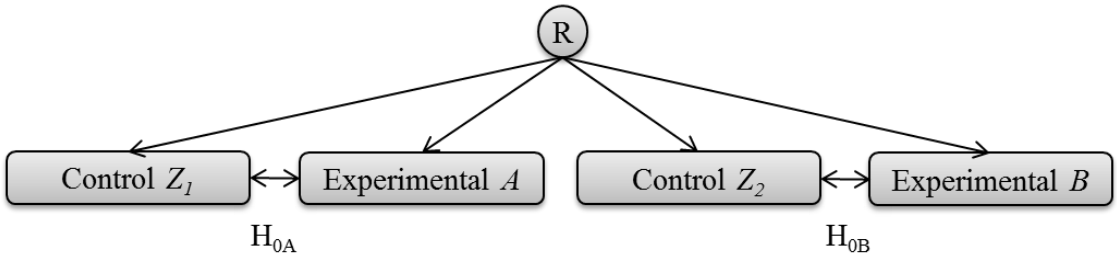
$$H_{0j}: \mu_j = \mu_Z$$

$$H_{1j}: \mu_j \neq \mu_Z$$

If the hypotheses are assessed in independent trials, it is accepted that there is no requirement to adjust for multiple testing, even if they are assessed by the same investigators and trials teams in the same centres and based on similar protocols. If they are instead assessed in the same trial but are designed to be tested in exactly the same way as they would have been in separate trials, where the data are entirely independent and non-overlapping with separate control groups Z_1 and Z_2, and the hypotheses are both powered separately and appropriately (as shown in Figure 3-2), it

would be difficult to argue for multiple testing adjustment to consider the familywise error rate across the whole protocol.

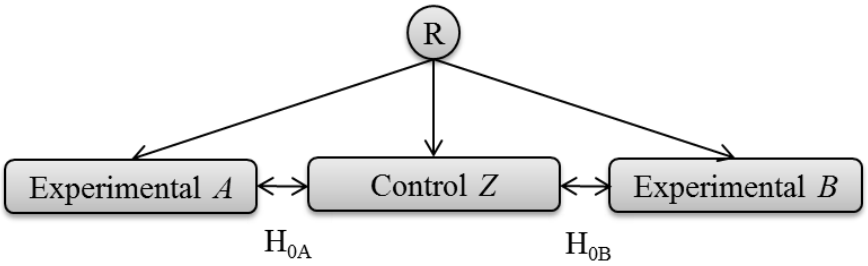
Figure 3-2: Illustration of two independent hypotheses being tested within the same protocol. The control groups are entirely separate.



The sharing of a protocol or even a randomisation system in this case does not affect the statistical probability of an error over that for independent trials. Westfall et al. (2010)[65] report that it may be plausible that multiplicity problems due to sharing a protocol could result from "selection effects" such as the method of assessment of the primary endpoint. However, this could just as easily occur in two independent trials led by the same trials team. Therefore it would seem that there is no additional reason for multiplicity concerns due to simply sharing a protocol, when separate pieces of confirmatory evidence are not required to be obtained from distinct teams. Asking more than one question independently but within the same protocol may be advantageous to reduce the burden of trial management issues such as funding applications, approvals, CRF development, staffing and set-up time. There are examples of this type of design in practice. Umbrella trials, for example, include a number of sub-study randomisations under the same protocol. These are often stratified by different eligibility requirements such as biomarker profiles, so each randomisation includes patients with different characteristics. Recent examples of trials with these designs are FOCUS4[73] and Lung-MAP[74]. Since each randomisation is independent and has its own control patients, the trials have not included a multiplicity adjustment for having multiple primary hypotheses within the protocol.

The example in Figure 3-2 is unlikely to make practical sense where the eligibility criteria and control group for both new experimental treatments are the same. Efficiency can be greatly improved by comparing both experimental arms to the same group of control patients. If the treatment difference being sought is the same, then utilising a single control group offers a saving of 25% of the trial size for an even allocation ratio (Figure 3-3).

Figure 3-3: Illustration of a multi-arm design where two hypotheses are tested within the same protocol and share the same control patients.



As in the previous example, both hypotheses are addressed separately and have both been adequately powered. Given the logic that the use of the same protocol alone does not cause multiplicity concerns over the same hypotheses being tested in independent trials, the difference is around the shared use of the control data. The comparisons are no longer independent, but are correlated based on the shared comparator group. The impact of this correlation on the chances of errors can be formally quantified, although very little literature has been published assessing the effect of shared control data on the probabilities of type I errors over those in independent trials, and this effect is rarely considered when assessing the requirement for multiplicity adjustment in multi-arm trials. Section 3.5 describes the known effects of correlation due to control data, and this work is extended in Chapter 4.

Note that when a new experimental arm is added to an ongoing trial, if only concurrently randomised control patients are used, it may be that some of the control patients differ between the comparisons as in Figure 3-2, but there may also be some overlapping patients that are used for both the original and new comparisons as in Figure 3-3. In these chapters when considering the implications of having shared control data on multiplicity, the ‘worst case’ scenario is that all control patients are used in all comparisons, and therefore this will be assumed here in order to offer the most conservative findings. If only some of the control patients are shared due to the timing of the amendment, the effects of the shared control data will be diluted. This will be considered as part of the recommendations, and also in Chapter 5 when investigating the appropriate way to analyse a trial in which an arm was added, including applying a multiple testing adjustment.

3.4.2 Including more hypotheses than would have been assessed in independent trials

Section 3.4.1 highlights that a key statistical implication of running a single multi-arm trial compared to separate trials is due to multiple use of shared control data. However, another factor that could increase the chance of a false conclusion over that for independent trials is the ability to test more hypotheses than would otherwise have been assessed. The necessity for adjustment in this case is a largely philosophical, rather than necessarily statistical, argument that is well addressed in the literature. The majority of the publications identified in the literature review in Sections 3.2.3 and 3.2.4 debate this issue, and it can be seen from the review that there is no consensus of opinion. On interpreting the literature, and ignoring the issues around having shared control data on the error rates since that is investigated separately, my opinion is that the need for adjustment depends on whether or not the hypotheses inform a single claim of effectiveness. If the hypotheses are to be interpreted independently, for example because they assess different experimental therapies, they are likely to inform entirely separate claims of effectiveness. However, if for example the trial is assessing different doses of the same experimental therapy, any success could lead to promotion of that therapy and therefore their hypotheses are likely to inform a single claim of effectiveness. This has been referred to as the 'claim-wise error rate'[57]. My argument for adjustment or not in these cases is as follows:

- ***If the hypotheses inform different claims of effectiveness, FWER control is likely to be an unnecessary penalty***

  If the hypotheses in a multi-arm trial do not inform the same claim of effectiveness because the experimental therapies are distinct, then it can be argued that the chance of a false positive error with relation to each therapy is unaffected by the inclusion of the other hypotheses. It has been argued, and it stands to reason, that the hypotheses should therefore not be interpreted as a 'family' since they do not contribute towards the same recommendation, and FWER control is an unnecessary penalty[49, 62].

- ***If the hypotheses contribute towards a single claim of effectiveness, they are likely to be considered a 'family' and therefore FWER adjustment may be required***

  If the ability to assess an increased number of primary hypotheses due to the efficiency of the multi-arm trial leads to a single therapy being assessed within a

number of experimental arms, the chance of a false positive error occurring with respect to that therapy will be increased. For example, if two different doses of a therapy are being assessed, each with a 5% probability of a false positive outcome, the overall chance of a false positive being reported from either one of those doses is increased to up to 9.75%. Due to assessing multiple hypotheses, there is a greater chance that there will be at least one false-positive error caused by a deviation in at least one of the samples from the true population, whether that is in an experimental or control arm. In this case there is general agreement in the literature that FWER control is recommended. Note that it also follows that the power in this case can be considered to be the overall chance of observing at least one true positive outcome, and this will also be increased by testing multiple hypotheses. Therefore, the penalty caused by applying the FWER adjustment may be compensated to some extent by the gain in overall power, otherwise known as the disjunctive power[75], as discussed in Section 4.6.

There is an element of common sense and logical argument required in determining whether it is necessary to control the FWER across all hypotheses in a multi-arm trial due to the increased number of hypotheses, or whether pairwise type I error control is adequate. It needs to be considered whether the hypotheses are likely to have otherwise been assessed in independent trials; whether they inform a single claim of effectiveness, perhaps in full or in part; and whether there is an associated gain in power for that claim of effectiveness. This does not contradict ICH E9 ‘Statistical Principles for Clinical Trials’[51], which states that “adjustment should always be considered, and the details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan.” Hung and Wang (2010)[69] discuss defining “a relevant family of hypotheses for which the type-I error needs to be properly controlled”, and recommend a “clinical decision tree”, determined in advance, to decide what aspects need to be protected from type I error inflation. The decision on adjustment due to assessing multiple hypotheses should be made at the design stage for each trial and documented with full justification.

3.5 The known effects of shared control data

The previous section separated out the two key causes that can affect the chances of errors in multi-arm trials compared to those in independent trials:

- The effect of correlation due to multiple use of the shared control data

- The increased chance of making a single claim of effectiveness due to testing more hypotheses involving the same therapy than would have been assessed in independent trials

As previously discussed, the second point is largely philosophical rather than necessarily statistical, and has been well addressed in the literature, albeit with varying opinions. Adjustment for this reason needs to be considered and justified on a trial-by-trial basis. However, the first point has been less well addressed and does not appear to be as widely understood, and therefore forms the main focus of this investigation into multiple testing adjustment for multiple hypotheses.

It is known that when multiple hypotheses have a shared control group, the FWER for those comparisons is lower than that for independent comparisons. This has led to the development of less conservative adjustment methods such as those discussed in Section 3.3.3. However, this phenomenon has rarely been considered in the literature addressing the need for multiple testing adjustment in multi-arm trials. In this section, the known impact of the correlation due to shared control data on the chances of type I errors is reviewed so that the reason for the reduction in the FWER is understood. The influence that this has had on the consideration of multiplicity adjustment within the literature is reviewed in order to determine what is already understood regarding the effect of the shared control data on multiplicity, and whether any further research is beneficial to help to inform the need for adjustment in multi-arm trials.

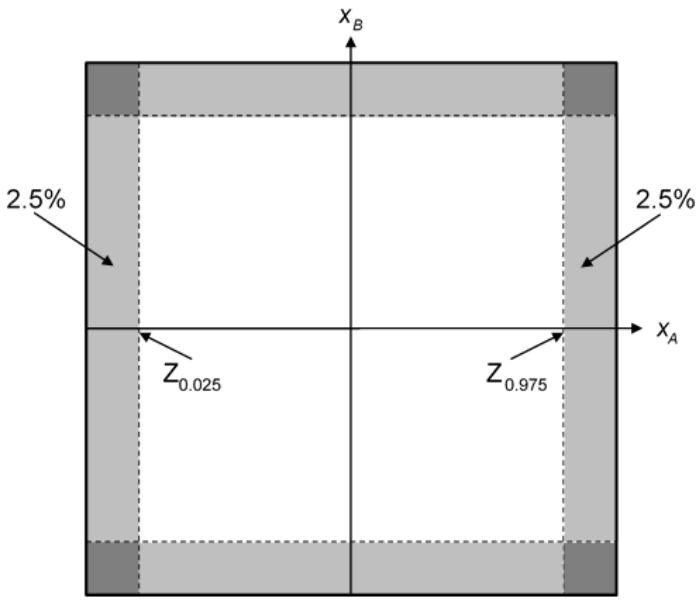
3.5.1 Fernandez & Stone "Multiplicity adjustments in trials with two correlated comparisons of interest"

Fernandez & Stone (2011)[72] was identified during the literature review described in Section 3.2.2. The paper does not question the need for multiplicity adjustment, since they take the assumption that adjustment is necessary in the case of having multiple doses of the experimental treatment, so was not relevant to the scope of the review. However, the methods used to visualise the type I error regions and the effect of the lack of independence between the test statistics are interesting and useful. The authors assess the impact of having correlated comparisons on the FWER after applying non-parametric multiple adjustment methods, and discuss parametric adjustment methods that account for this and are therefore less conservative. A 'Swiss Flag' diagram is introduced to diagrammatically represent the type I error rates, which is a helpful aid in understanding the various error regions where there are two comparisons, and picturing how the various multiple adjustment methods affect these. They also discuss

the concept that the error regions can be calculated based on the joint distribution of the comparisons having a bivariate normal distribution, and this assumption is taken forward into the research in Chapter 4.

In the manuscript, the ‘Swiss Flag’ is used to illustrate the probability of type I errors for two comparisons plotted on the same axes, with comparison A being displayed horizontally, and comparison B being displayed vertically, as shown in Figure 3-4.

Figure 3-4 ‘Swiss Flag’ illustrating the rejection regions for two independent comparisons



Note X_A *and* X_B are the test statistics assessing therapies *A* and *B* against their independent control groups, and Z_p represents the p^{th} percentile of the cumulative distribution function of a standard normal distribution. The shaded regions around the edges represent the critical regions, based on a standard hypothesis test for a normally distributed test statistic. It is explained in the Fernandez & Stone paper as follows: “For each comparison, under H_0, there is a [2.5%] probability of observing a treatment effect in favour of the experimental treatment, and a [2.5%] probability of observing a treatment effect in favour of the control arm, giving an overall [5%] significance level per comparison”. It can be seen that even in the independent case there is some ‘double counting’ in the rejection regions in the four corners. Since the FWER is the probability of ‘at least one error’, it can be calculated by the sum of the edges (4 x 0.025) minus the amount that is double counted in the corners (4 x 0.025^2), which is 0.0975, agreeing with that expected from Section 3.3.1. This explains why the

Bonferroni method (Section 3.3.2.1) is conservative, because the method adjusts based on the assumption that the total error region is 0.1 and does not account for the occasions where the errors occur within the same pair of comparisons.

Fernandez & Stone then discuss the impact of comparing both therapies to the same control, so they are not independent. They show that the correlation between comparisons (ρ) can be calculated based on the allocation ratio, such that:

$$\rho = \frac{1}{\sqrt{\left(\frac{n_Z}{n_A}+1\right)\left(\frac{n_Z}{n_B}+1\right)}},$$

where n_Z is the number of patients in the control arm and n_A and n_B are the numbers in experimental arms *A* and *B* respectively.

If the allocation ratio is 1:1:1, the correlation is 0.5. Therefore, Fernandez & Stone state "if it is reasonable to assume that the sample size is large enough so that each comparison follows a Normal distribution, then their joint distribution will follow a bivariate Normal distribution, with correlation = 0.5". They provide SAS code to calculate the probabilities for the critical regions around the edges of the Swiss Flag after accounting for the correlation, based on the assumption that the joint distribution follows a bivariate Normal. They use these methods to compare various multiple testing adjustment procedures in order to make recommendations concerning the most efficient in the case where the comparisons share a control group.

3.5.2 Proschan & Follmann "Multiple comparisons with control in a single experiment versus separate experiments: why do we feel differently?"

As noted within the literature review in Section 3.2.4, Proschan & Follmann (1995)[63] is the only publication identified to have assessed the effect of the shared control data on the probabilities of type I errors and related this to the requirement for multiple testing adjustment. They investigated the impact of the dependency of the control group on the overall type I error rates for the multi-arm trial compared to independent trials by comparing the error rates for independent trials (I) to trials with a shared control group (D). They calculate the type I errors under the theory that in the independent case (I), the number of type I errors is a binomial random variable, and in the dependent case (D), the distribution is "conditioned on the standardized sample control mean". This is

shown to lead to the following formula, using the notation of this thesis, to calculate the probability of type I errors under the dependent case:

$$\Pr(N_D = x) = \left(\frac{1}{\sqrt{2\pi}}\right) \int_{-\infty}^{\infty} \binom{m}{x} [p(z)]^x \times [1 - p(z)]^{m-x} \exp\left(-\frac{z^2}{2}\right) dz,$$

where $p(z) = 1 - \Phi(z + \sqrt{2}Z_\alpha)$, *m* is the number of comparisons and *α* is the one-sided significance level.

The probabilities of type I errors are calculated for trials with two, three and four hypotheses based on a one-sided 5% significance level, and tabulated. Note that only one-sided tests are considered for simplicity in this publication. It isn't clear how this would extend to calculate the probabilities in the common case of two-sided tests.

The authors show that the probability of making one type I error is lower when there is a positive dependence between the test statistics than in separate experiments, as expected, but that the probability of making two or more errors is higher. They calculated that in situation (D) in the case of two hypotheses with even allocation and a one-sided 5% error rate, the chance of exactly one type I error is 0.0756 and the chance of exactly two errors is 0.0122, therefore the FWER is 0.0878. Recall that in situation (I) the FWER is 0.0975 and the chance of exactly two errors is 0.0025. Therefore, whilst the FWER is smaller under (D), there is a substantial increase in the conditional probability of a type I error for one comparison if a type I error exists for the other. This is because "no matter how large [*m*] is, there is only one control group under (D); if it is "bad" then all of the comparisons are affected". They assess the power in situation (I) compared to situation (D) after applying a Dunnett adjustment for multiple testing, and conclude that is it still worthwhile to do a single experiment in many situations unless there are an 'unrealistically large' number of treatments. However, they do not quantify the effect that the Dunnett adjustment has on the increased probability of more than one type I error.

In their conclusions, the authors discuss the contentiousness of adjustment, countering the argument for adjusting only where the hypotheses are families that 'are formed of related statements' in a single experiment because "related statements are often made in separate large experiments, and we feel confident about the significant results". They conclude that "the difference in the distributions of type I errors is relatively small when there are not too many treatments", however a 'bad' control group leading to type I errors in a single experiment would reduce the chance of a contradicting result over

that had the trials been independent. They leave the overall conclusion fairly vague, without providing any conclusive guidance on the need for adjustment based on their findings regarding the distribution of errors. They suggest that it is not enough to only consider the FWER and PCER, but that baseline characteristics of control patients should be compared to those in other trials to investigate whether the control group may be 'bad'. However, this would be a subjective investigation based on the choice of: the baseline factors investigated; the number of patients involved; the availability of similar, concurrent trials data; and the general uncertainty in the comparison in terms of power of concluding a difference or not. Therefore this cannot be relied on in order to determine the need for adjustment.

Although Proschan and Follmann's work was published in 1995, it is rarely referenced or considered when discussing the requirement for multiplicity adjustment in multi-arm trials. Freidlin et al. (2008)[49] include it as part their argument for not adjusting where hypotheses are not clinically related, as described in Section 3.2.4. No other literature was found to consider the effect of the dependency of the shared control data on the requirement for multiple testing adjustment. Whilst Proschan & Follmann's work is interesting and relevant, their findings are limited and do not obviously translate into a recommendation on the need for adjustment. They only assessed one-sided tests with an even allocation ratio, and did not assess the effect of adjustment methods on the distribution of the type I errors compared to independent trials in order to fully consider their usefulness and make informed recommendations. This research is therefore reproduced and extended in Chapter 4.

3.5.3 Senn "Statistical Issues in Drug Development"

Senn's book on Statistical Issues in Drug Development (1997)[76] includes a chapter on multiplicity, with Section 10.2.8 entitled "Even when the probabilities of making at least one type I error are controlled, conditional error rates may not be". This short section highlights that in the case that the placebo (or control) results are low by chance, they are used as a comparator for all experimental treatments and therefore this will inflate the error over all comparisons. Senn demonstrates that in this case, even after applying a Bonferroni correction, the 'conditional probability' that a contrast is significant, given that previously tested contrasts are significant, is inflated due to the correlation between the comparisons caused by sharing the placebo arm.

Senn recommends that a ‘structured approach’ could be used to test doses in a pre-defined order if appropriate, or else closed testing methods could be implemented. However the effects of these methods on the conditional probability of an error is not discussed, and there are no clear recommendations on multiple testing adjustment considering the increased probability of a conditional error.

3.6 Summary

The objective of this chapter was to introduce the concept of multiple testing adjustment in multi-arm trials and the conflicting viewpoints on its necessity; understand common adjustment methods and how they are applied; consider the causes affecting multiplicity where hypotheses are assessed in multi-arm trials compared to independent trials; and provide an overview of the most relevant literature addressing the impact of having shared control data, in order to inform future work.

Published opinion is divided on the requirement for multiple testing adjustment to control the FWER, and there is no comprehensive guidance available for researchers, therefore leading to differences in practice. There are two reasons why false positive error rates may be affected in multi-arm trials compared to independent trials. The first is due to correlation between the comparisons caused by the shared use of the control data; and the second is an increased chance of making a claim of effectiveness because of an increased ability to test a family of hypotheses. Whilst the second point is widely debated, very little literature has been published considering the effect of shared control data on the probabilities of type I errors over those in independent trials, and how this impacts on the requirement for multiplicity adjustment in multi-arm trials. Proschan and Follman (1995)[63] showed that the positive correlation between the test statistics reduces the FWER over that in independent trials, but that the probability of making two or more errors is increased. However, this work is not well cited or considered in the literature discussing the requirement for adjustment.

Whether multiple testing adjustment is included or not has the potential to have a real impact on the interpretation of the trial results and their influence on practice for confirmatory trials. For example, the ALTTO trial in HER2-positive early breast cancer[77] assessed two experimental arms against a shared control in a large, confirmatory trial. The experimental arms were a sequence of trastuzumab (T) followed by lapatinib (L), or the combination of L+T, both compared to T alone in terms of disease-free survival.

Following a Bonferroni adjustment, both hypotheses were assessed against a significance level of 0.025. The p-value for the L+T vs T alone hypothesis was 0.048, which was not significant due to the Bonferroni adjustment, and therefore the trial outcome was that T remains standard of care. Whilst there are efficiencies associated with testing multiple hypotheses within the same protocol, adjusting for multiple testing can be disadvantageous for the individual hypotheses, and so it must be carefully considered in advance whether adjustment is truly necessary.

In the next chapter, the work discussed here is extended in order to fully investigate the effect of correlation due to shared control data on the different probabilities of one or more type I errors, and the usefulness of multiple testing adjustment methods on controlling these errors. This enables informed and comprehensive recommendations to be made on the need for a multiple testing adjustment in multi-arm trials with shared control data.

Chapter 4
Multiple testing adjustment in multi-arm trials with a shared control group: the effect of positive correlation between the test statistics and recommendations

4.1 Introduction

4.1.1 Background

In Chapter 3 the concept of multiplicity was introduced, including some common viewpoints for and against adjustment in the context of multi-arm trials with a shared control group. The reasons why false positive error rates may be affected in multi-arm trials compared to independent trials were identified to be related to: correlation between the comparisons due to the shared control data; and the increased chance of making a claim of effectiveness due to the ability to assess more hypotheses. Whilst the necessity for adjustment because of an increased chance of making a claim of effectiveness has been well addressed in the literature, with the opinions related to this point summarised in Section 3.4.2, very little literature was found to have been published assessing the effect of shared control data on the probabilities of type I errors over those in independent trials and the associated impact on the requirement for multiplicity adjustment. In Section 3.5, the known effects of having shared control data were introduced. In summary, Proschan and Follmann (1995)[63] and Senn (1997)[76] reported that where the comparisons are correlated due to a shared control group, the FWER is lower than that in independent trials, but the conditional probability of a second or further type I error is higher. However, these findings have not led to recommendations for multiple testing adjustment that are commonly considered within the literature. In this chapter, the work of Proschan and Follmann is extended to fully investigate the effect of correlation due to multiple use of the shared control data in order to make recommendations on multiple testing adjustment in multi-arm trials. In Section 4.2, other relevant types of type I error rate are defined in addition to the FWER, relating to the probability of more than one type I error occurring within the family. In Section 4.3 each of these error rates are assessed for multi-arm trials with shared control data and compared to those for independent hypotheses. This is extended in Section 4.4 to consider how well multiplicity adjustment methods control

the various error rates. Finally, in Section 4.5 the adjusted significance levels necessary to control the rate of multiple type I errors in favour of the experimental therapies to that in independent hypotheses are calculated for a three-arm trial. The discussion in Section 4.6 includes recommendations on adjustment based on these findings.

Note that when an experimental treatment is added to an ongoing trial and uses the existing control group as a comparator, the stage at which the arm is added becomes a multi-arm stage within the trial, regardless of the original trial design. Therefore consideration of a multiple testing adjustment is always relevant when adding an arm, whether it relates to the whole trial or only a stage of it. The findings from this chapter inform Chapter 5 on the appropriate analysis methods alongside multiple testing requirements over the stages of trials following the addition of an arm.

As noted in Chapter 3, the work from these two chapters combined has been published in the journal *Statistical Methods for Medical Research* (Howard et al., 2018)[8]. The work contained within the publication is directly attributable to myself as first author, with input from the co-authors who are all part of my PhD supervisory team.

4.1.2 Motivational examples

Multi-arm trials with different designs and varying levels of relatedness between the hypotheses may have different requirements for multiple testing adjustment. The research in this chapter aims to investigate the effect of having multiple hypotheses with shared control data on the probability of type I errors in order to make recommendations on adjustment, considering the individual trial design and aims. The following three real life examples each have multiple hypotheses, but differ in terms of how the outcomes for the hypotheses are interpreted with relation to one another and how they may affect practice. These trials are revisited in the discussion section in order to demonstrate how the recommendations could be applied in each case.

4.1.2.1 MRC COIN

The phase III MRC COIN trial[78] in previously untreated patients with colorectal cancer had three-arms and two primary hypotheses, and recruited from 2005 to 2008. The control treatment (arm *Z*) was chemotherapy with oxaliplatin and fluoropyrimidine (OxFP) given continuously. One experimental arm (arm *A*) included an additional

therapy cetuximab to OxFP, and the other (arm *B*) assessed the chemotherapy OxFP given intermittently. Patients were randomised to the three treatment arms with a 1:1:1 ratio, and the trial objective was to assess a difference in overall survival at two years for each of the comparisons, arm *A* vs *Z* and arm *B* vs *Z*.

4.1.2.2 AMAGINE-1

The phase III AMAGINE-1 trial (clinicaltrials.gov identifier: NCT01708590) was run by Amgen / AstraZeneca from 2012 to 2015. The trial assessed the safety and efficacy of brodalumab taken every two weeks via subcutaneous injection at two doses (140 mg or 210 mg) compared with placebo in patients with moderate-to-severe plaque psoriasis. The primary hypotheses concerned the efficacy of each dose of brodalumab compared to placebo, as assessed by Static Physician Global Assessment (sPGA) score and improvement in Psoriasis Area and Severity Index (PASI) at 12 weeks.

4.1.2.3 Myeloma XI+ Intensive

The Myeloma XI Intensive trial (ClinicalTrials.gov Identifier: NCT01554852) at the University of Leeds opened to recruitment in 2010, comparing the current standard therapy CTD (cyclophosphamide, thalidomide and dexamethasone) with CRD (cyclophosphamide, lenalidomide and dexamethasone) in terms of progression-free survival (PFS) in newly diagnosed patients with Multiple Myeloma. It was anticipated that recruitment would take up to four years, with the required number of events occurring within three years after the close of recruitment. During recruitment, early evidence suggested a new therapy, carfilzomib, added to the existing CRD regime (CCRD) might improve efficacy. Since it was of interest to assess CCRD as soon as possible, the follow-on Myeloma XI+ intensive trial was designed without waiting for the results of the original trial, and opened to recruitment in 2013 following on seamlessly from Myeloma XI within the same master protocol. The Myeloma XI+ trial therefore compared the experimental therapy CCRD to the current standard control CTD and the previous experimental therapy CRD at a 2:1:1 randomisation, in order to protect the trial in the case that CRD was found superior and superseded CTD as the standard therapy before the amended trial had completed and reported.

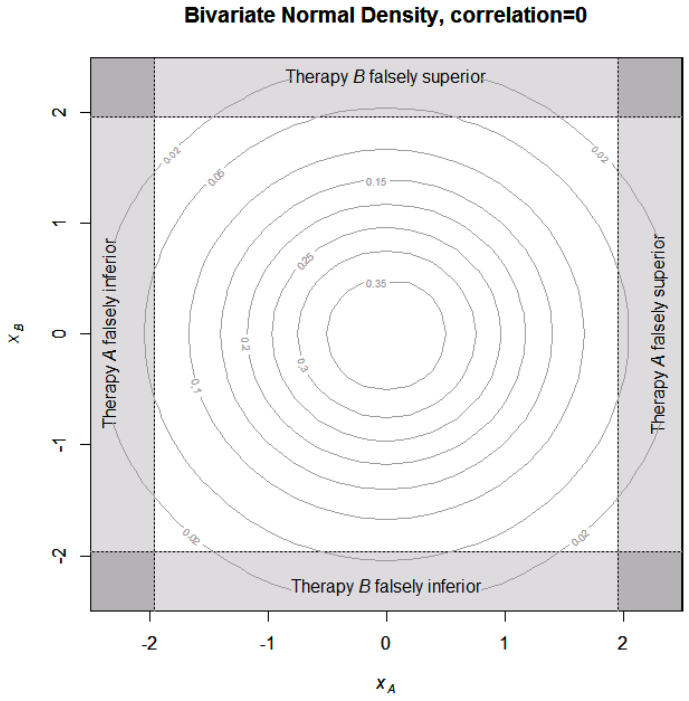
4.2 Definitions of error regions

The greatest concern in terms of multiple testing adjustment in the literature is control of the chance of at least one type I error amongst the family of hypotheses, which has been previously defined as the familywise error rate (FWER). However, it is important to understand the effect of the correlation on the totality of the error that exists when reporting outcomes from multiple hypotheses, which includes not only the chance of at least one type I error, but also the probability of more than one type I error occurring within the family. Therefore, two other types of error rates to consider when determining the effect of the correlation are now defined: the family multiple error rate (FMER) and the probability of multiple superior false positives (MSFP). These error regions are not considered in the general literature on multiple testing, but their relevance is discussed throughout this chapter and particularly in the discussion section. All three error regions are described in this section.

4.2.1 Bivariate normal density rejection regions

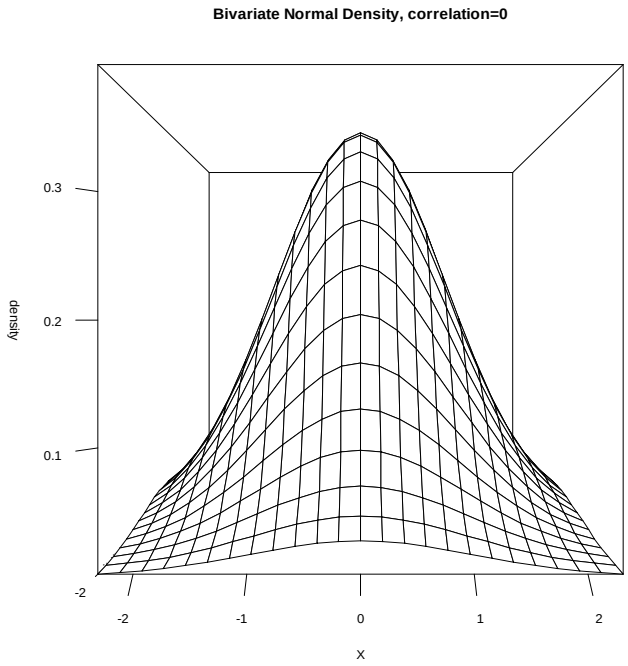
Assume there are two independent hypothesis tests, with the null hypothesis H_{0A} assessing Treatment *A* against Z_1 with standardised test statistic X_A, and the null hypothesis H_{0B} assessing Treatment *B* against Z_2 with standardised test statistic X_B, as illustrated in Figure 3-4, each with a two-sided significance level of 0.05. The test statistics for each comparison asymptotically follow a normal distribution when sample sizes are reasonable[79]. Figure 4-1 illustrates the joint density for the standardised test statistics based on the probability density function of the standard bivariate normal distribution with no correlation. The rejection regions for the hypothesis tests are the shaded areas around the outside of the square, as described by Fernandes and Stone (2011)[72] and discussed in detail in Section 3.5.1. Figure 4-2 is the equivalent perspective plot of the 3d surface, with X_A on the x-axis, X_B on the y-axis and the probability density illustrated on the z-axis. The figures were obtained using the software R, version 2.15.2.

Figure 4-1 Rejection regions for two independent comparisons plotted on orthogonal axes, with the standardised test statistic for the null hypothesis H_{0A} being displayed horizontally, and H_{0B} displayed vertically



X_A and X_B are the test statistics assessing therapies A and B against the control group. The contours represent the probability density function of the joint distribution of the standardised test statistics with no correlation.

Figure 4-2 Equivalent perspective plot of the 3d surface with the probability density illustrated on the z-axis.



As described in Section 3.5.1, the probability of falling within a given shaded region along the length of an edge in Figure 4-1 is 2.5%. That is, the probability of concluding that either therapy is either falsely inferior or falsely superior to its control therapy is 2.5%, since the overall two-sided type I error for each hypothesis is set at 5%. The darker shaded corner regions represent the probability that both hypotheses have false positive outcomes, so there are two type I errors. In order to understand how the correlation between the test statistics affects the amount of error in the different error regions in Figure 4-1, different types of false positive errors are defined below. These will first be quantified in the case of independent comparisons before exploring the case where there is shared control data in Section 4.3.

4.2.2 Familywise Error Rate (FWER)

Recall from Section 3.3.1 that the FWER is the overall probability of at least one false positive conclusion anywhere within a defined set of trial hypotheses. In the independent case with m comparisons, a probability α of finding a significant difference, and y denoting the number of type I errors (y = 1,…,m), the FWER can be calculated by

$$FWER = p(Y > 0) = 1 - p(Y = 0) = 1 - (1 - \alpha)^m.$$

Thus, with two independent comparisons and α = 0.05 for each, the FWER is 0.0975.

In Section 3.5.1 it was shown that using the joint probability density plot illustrated in Figure 4-1, in the case of two independent comparisons and a two-sided significance level of 0.05, the FWER can be calculated by the total probability that falls in the shaded region, which is $(4 * 0.025) - (4 * 0.025^2) = 0.0975$, as expected.

4.2.3 Family Multiple Error Rate (FMER)

A second type of false positive error that may be important, although rarely considered in the literature, can be defined as the chance of multiple false positive findings across the family of hypotheses, which here will be called the Family Multiple Error Rate (FMER).

Define the total error rate to be the sum of the errors for each hypothesis. With a family of two null hypotheses H_{0A} and H_{0B} respectively relating to the comparisons of

therapies *A* and *B* against control, and α=0.05 for each, the total error is 0.1. By probability theory:

$$\mathrm{P}(A) + \mathrm{P}(B) = \mathrm{P}(A \cup B) + \mathrm{P}(A \cap B),$$

$\mathrm{P}(A)$ is the probability of a type I error for the null hypothesis H_{0A},

$P(B)$ is the probability of a type I error for the null hypothesis H_{0B},

$\mathrm{P}(A \cup B)$ is the overall chance of a type I error, i.e. the FWER,

$\mathrm{P}(A \cap B)$ is the chance of two type I errors occurring from the pair of null hypotheses, which is the FMER.

In Figure 4-1, the FMER is represented by the sum of the probabilities in the four dark shaded corner regions. In the case of two fully independent hypotheses tested in two separate trials, the FMER is $4 * 0.025^2 = 0.0025$. Therefore FWER + FMER is 0.1, as expected. Note that the FMER is directly related to the conditional probability of a type I error $P(B|A)$, as discussed by Senn (1997)[76] in Section 3.5.3, since $P(B|A) = \frac{P(\mathrm{A} \cap B)}{P(A)}$. Senn describes the conditional error in terms of a trial comparing different doses of a treatment to a placebo as "the probability under the null hypothesis of concluding that a given dose is significant given that all other doses tested to date are significantly different from placebo". However, the FMER is easier to interpret in the case of a multi-arm trial due to the hypotheses not necessarily having any sensible order or not always being assessed in conjunction with one another.

In the case of three null hypotheses, H_{0A}, H_{0B} and H_{0C}, each with α=0.05, the total error is 0.15. This is equal to the probability of at least one error (FWER), the probability of at least two errors (FMER2), and the probability of three errors (FMER3), as follows:

$$\mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C)$$

$$= \mathrm{P}(A \cup B \cup C) + \mathrm{P}(A \cap B) + \mathrm{P}(A \cap C) + \mathrm{P}(B \cap C) - 2 * \mathrm{P}(A \cap B \cap C) + \mathrm{P}(A \cap B \cap C)$$

$$= \mathrm{P}(A \cup B \cup C) + \mathrm{P}(A \cap B) + \mathrm{P}(A \cap C) + \mathrm{P}(B \cap C) - \mathrm{P}(A \cap B \cap C),$$

$\mathrm{P}(A)$ is the probability of a type I error for the null hypothesis H_{0A},

$P(B)$ is the probability of a type I error for the null hypothesis H_{0B},

$P(C)$ is the probability of a type I error for the null hypothesis H_{0C},

$\mathrm{P}(A \cup B \cup C)$ is the overall chance of at least one type I error, i.e. the FWER,

$\mathrm{P}(A \cap B) + \mathrm{P}(A \cap C) + \mathrm{P}(B \cap C) - 2 * \mathrm{P}(A \cap B \cap C)$ is the chance of at least two type I errors, which is the FMER2,

$\mathrm{P}(A \cap B \cap C)$ is the chance of three type I errors occurring, which is the FMER3.

4.2.4 Multiple Superior False Positives (MSFP)

In Figure 4-1, the lower left dark shaded corner signifies both false positives falling in the rejection region in favour of the control, thus falsely declaring the experimental treatments significantly inferior in both cases (that is, multiple inferior false positive outcomes). The upper right corner signifies both false positives falling in favour of the experimental treatments, thus falsely declaring the experimental treatments significantly superior to control in both cases, which here will be called multiple superior false positive (MSFP) outcomes. The upper left and lower right corners signify one false positive favouring the control and the other the experimental treatment. Note that in the independent case with two hypotheses, the probability of MSFP outcomes is $0.025^2 = 0.000625$.

The chance of MSFP errors in particular are important because they could contribute towards a therapy being recommended for use in practice when in truth it is no better than the current standard therapy. If multiple hypotheses with a shared control group could be used as separate pieces of evidence to inform a single claim of effectiveness, for example when assessing different doses of the same therapy, the probability of multiple false conclusions of superiority (the MSFP rate) should not be inflated over that for independent studies. This is a similar issue to that discussed in two publications on regulatory strategies for one large pivotal trial in place of two smaller ones. Fisher (1999)[80] and Shun et al. (2005)[81] discuss that the overall 'positive rejection region' in a single, large trial is required to be controlled to the same level as in two smaller trials in order for both hypotheses to inform regulatory applications. This is not something that has been found to have been addressed in the literature when considering type I errors in multi arm trials, and is investigated in detail in this chapter.

4.3 The effect of the positive correlation on the error regions

In Section 3.4.1 it was discussed that in a multi-arm trial with a shared control group, as illustrated in Figure 3-3, the comparisons are not independent. If the control group sample, by chance, perform worse than the true population, there is an increased

probability for all experimental therapies of reporting a false positive outcome to conclude that they are superior. The test statistics are therefore positively correlated, since the outcomes for the control sample will affect them all in the same way. In this section, the effect of the positive correlation on the different types of error are quantified.

4.3.1 Calculating the correlation between the test statistics due to sharing control data

Recall from Section 4.2.1 that in the case of multi-arm trials with independent experimental therapies and a shared control group, the test statistics for the comparisons, X_A and X_B, can be assumed to follow standard normal distributions when sample sizes are reasonable. Follmann et al. (1994)[79] state that the test statistic for a null hypothesis "asymptotically follows the distribution of a standardized Gaussian process for a variety of common test statistics including the *t*-statistic and the log-rank statistic". Their joint distribution therefore follows a standard bivariate normal with correlation ρ_{AB}. Dunnett[54] notes that the correlation between the test statistics is directly linked to the allocation ratio, as follows

$$\rho_{AB} = \frac{1}{\sqrt{\left(\frac{n_Z}{n_A}+1\right)\left(\frac{n_Z}{n_B}+1\right)}},$$

where n_j is the sample size in arm *j (j = A, B, Z)*.

This correlation can be easily confirmed in the specific case of continuous endpoint data. Assume that the data are independent and normally distributed with means $\bar{\mu}_j$ and common variance σ^2, for arm *j*. The correlation can be written as follows:

$$\rho_{AB} = corr(\bar{\mu}_A - \bar{\mu}_Z, \bar{\mu}_B - \bar{\mu}_Z) = \frac{cov(\bar{\mu}_A - \bar{\mu}_Z, \bar{\mu}_B - \bar{\mu}_Z)}{\sqrt{var(\bar{\mu}_A - \bar{\mu}_Z) var\left(\bar{\mu}_B - \bar{\mu}_Z\right)}}.$$

Since each of the trial arms are independent from one another:

$$cov(\bar{\mu}_A - \bar{\mu}_Z, \bar{\mu}_B - \bar{\mu}_Z) = cov(\bar{\mu}_A, \bar{\mu}_B) - cov(\bar{\mu}_A, \bar{\mu}_Z) - cov(\bar{\mu}_B, \bar{\mu}_Z) + cov(\bar{\mu}_Z, \bar{\mu}_Z)$$

$$= var(\bar{\mu}_Z) = \frac{\sigma^2}{n_z},$$

and

$$var(\mu_A - \mu_Z) = var(\mu_A) + var(\mu_Z) - 2cov(\mu_A, \mu_Z) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_Z^2}{n_Z}.$$

The correlation can therefore be calculated as follows:

$$\rho_{AB} = \frac{\frac{\sigma^2}{n_Z}}{\sqrt{\left(\frac{\sigma^2}{n_A}+\frac{\sigma^2}{n_Z}\right)\left(\frac{\sigma^2}{n_B}+\frac{\sigma^2}{n_Z}\right)}} = \frac{1}{\sqrt{\left(\frac{n_Z}{n_A}+1\right)\left(\frac{n_Z}{n_B}+1\right)}}.$$

It is less straightforward to derive the correlation algebraically for endpoints with different distributions, however since the correlation is between the test statistics, which are assumed to follow standard normal distributions for reasonable sample sizes, it should be independent of the distribution of the data used. In order to verify that this is the case, error rates were calculated for simulations of three arm trials with survival endpoint data, and these were compared to the algebraic results using this assumed correlation, as shown in Section 4.3.4 below. It can be seen that the simulated results match those calculated algebraically.

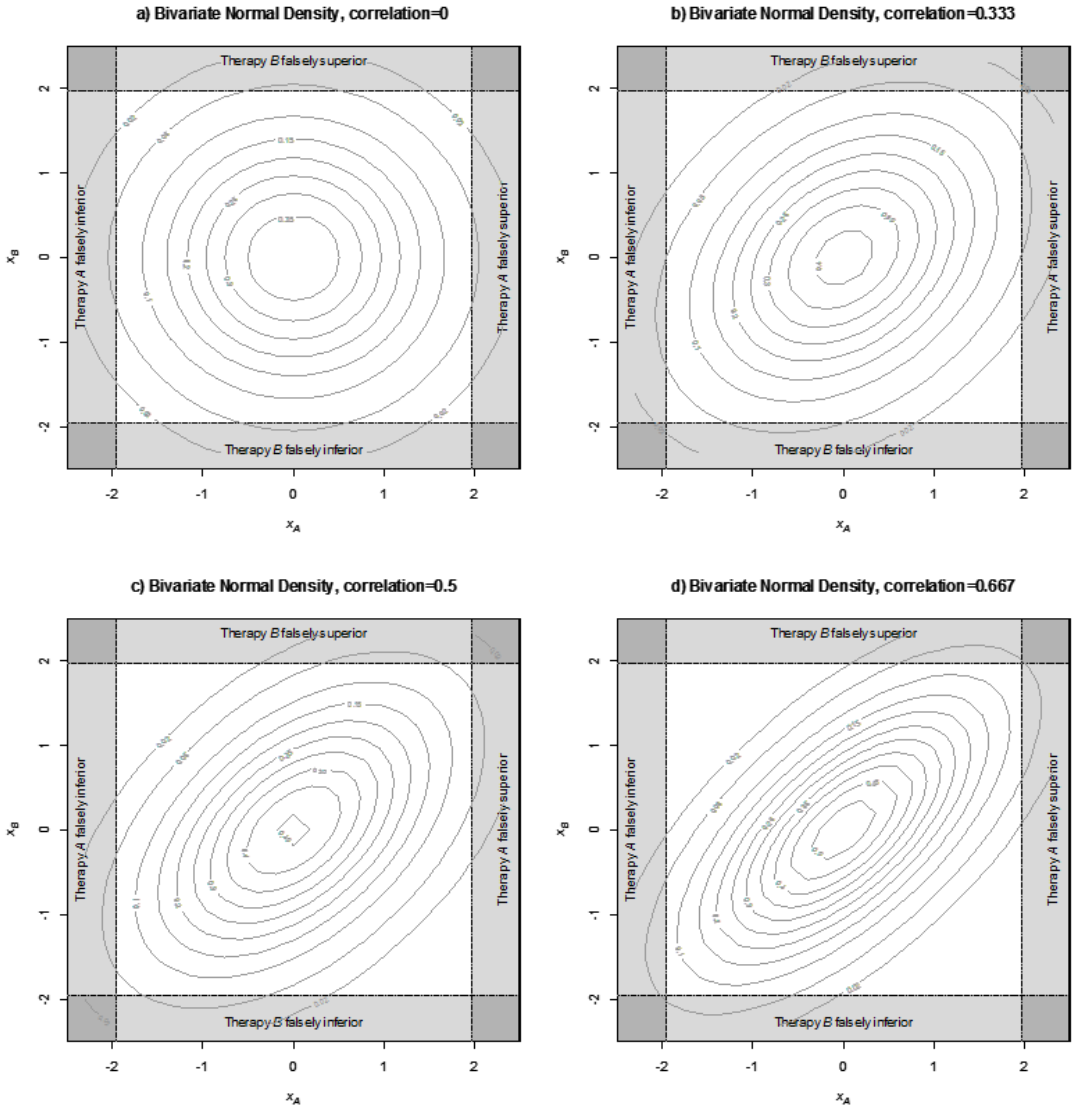
If the allocation ratio is 1:1:1, the correlation is 0.5. For an allocation of 2:1:1 in favour of control, the correlation is 0.333, and for 1:2:2 the correlation is 0.667.

4.3.2 Calculating the type I error regions assuming a multivariate normal distribution, incorporating correlation

The false positive error regions of potential interest in a multi-arm trial can be calculated based on the assumption of the joint distribution of the test statistics following a standard multivariate normal distribution. The R programs in Appendix C compute these probabilities in the case of two or three experimental therapies, respectively, allowing varying correlation in order to calculate the error density in each of the rejection regions that were defined in Section 4.2. Note that in the case of three hypotheses, the rejection regions illustrated in Figure 4-1 are instead considered within a cube, with each side representing one error, each edge two errors and each corner 3 errors. The 'pmvnorm' command calculates the multivariate normal distribution probabilities between limits for given correlation matrices, therefore outputting the probabilities of falling within the regions defined by the limits. The type I error regions are calculated by summing the errors within the appropriate rejection regions. For example, in the bivariate case, the MSFP region in the upper right corner is calculated based on the probability density that lies between $Z_{0.975}$ (i.e. the 97.5th percentile of the cumulative distribution function of a standard normal distribution, which is 1.96) and infinity for both the x and y axes, given the correlation matrix.

The effect of the correlation on the rejection regions in the case of two experimental treatments is illustrated in Figure 4-3. As the correlation increases, the proportion of error in the lower left and upper right corners, indicating false positive outcomes in the same direction for both hypotheses, also increases. That is, if the shared control group performs better or worse than expected, there is a greater chance of an error in both of the hypotheses in the same direction, as expected.

Figure 4-3 Illustration of rejection regions with varying amounts of correlation



X_A and X_B are the standardised test statistics assessing therapies A and B against the control group. The contours represent the probability density function of their joint distribution in the case of:

a) two hypotheses, each with individual control data, ρ=0

b) two hypotheses with shared control data and 2:1:1 randomisation, ρ=0.333

c) two hypotheses with shared control data and 1:1:1 randomisation, ρ=0.5

d) two hypotheses with shared control data and 1:2:2 randomisation, ρ=0.667

4.3.3 Comparison of type I error regions for multi-arm trials with a shared control group compared to independent trials

The R programs in Appendix C calculate the FWER, FMER and MSFP rates for different levels of correlation based on the allocation ratio in three-arm trials with two experimental arms and a shared control, and for four-arm trials with three experimental arms and a shared control. These probabilities are given in Table 4-1.

Table 4-1 FWER, FMER and MSFP comparisons for three and four arm trials with a shared control group and varying allocation ratios, compared to independent 1:1 randomised trials (α = 0.05 for each hypothesis)

	Independent case (Separate trials)	**Dependent case 2:1:1(:1) (2 to control)**	**Dependent case 1:1:1(:1)**	**Dependent case 1:2:2(:2) (1 to control)**
Correlation (ρ)	0	0.333	0.5	0.667
Reject H_0 for each individual hypothesis	0.050	0.050	0.050	0.050
Three-arm trial (hypotheses H_{0A} and H_{0B})				
FWER: Reject at least one H_0	0.0975	0.0946	0.0907	0.0849
FMER: Reject both H_0's (in any direction)	0.0025	0.0054	0.0093	0.0151
MSFP: Reject both H_0's in favour of treatments *A* and *B*	0.00063	0.00267	0.00462	0.00753
Four-arm trial (hypotheses H_{0A}, H_{0B} and H_{0C})				
FWER: Reject at least one H_0	0.1426	0.1348	0.1254	0.1124
FMER2: Reject at least two H_0's (in any direction)	0.0073	0.0141	0.0213	0.0301
FMER3: Reject all three H_0's (in any direction)	0.0001	0.0011	0.0032	0.0076
MSFP2: Reject at least two H_0's in favour of *A*, *B* or *C*	0.0018	0.0069	0.0107	0.0150
MSFP3: Reject all three H_0's in favour of *A*, *B* and *C*	0.00002	0.00056	0.00160	0.00378

In Section 2.3.5 it was discussed that the optimal allocation to the control group in multi-arm trials in order to minimise the total number of patients required is approximately the square root of the number of experimental arms. In this case, the correlation would be 0.414 and the FWER, FMER and MSFP results would lie between

those in the 2:1:1(:1) and 1:1:1(:1) cases. In the case of a three arm trial, the FWER is 0.0929, the FMER is 0.0071 and the MSFP rate is 0.00352, as expected.

4.3.3.1 Familywise Error Rate (FWER) comparison

The FWER is lower in all cases with shared control data than the equivalent error when assessing two independent trials. The greater the correlation, the lower the FWER. That is, the correlation between the test statistics reduces the overall probability of a type I error occurring across either of the hypotheses over the case where there is no shared control data. This agrees with the findings by Proschan and Follmann (1995)[63].

4.3.3.2 Family Multiple Error Rate (FMER) comparison

In a multi-arm trial with two hypotheses, the chance of multiple errors (FMER) has increased from 0.25% for independent trials to 0.93% in the case with even allocation, an increase of 3.7 times. The message stays the same as the number of hypotheses increases; in the case with three hypotheses and even allocation, the chance of any two errors (FMER2) has increased from 0.7% to over 2%, which is not trivial. Similar patterns of increases are found with unequal allocation ratios, and the trend across the resultant correlations from these changing allocation ratios can be clearly seen.

The increase in the FMER is due to the increased chance of an error occurring within the correlated comparisons in the same direction, often caused by a chance deviation in the outcome for the control sample from the outcome for the true population. For example, in the case of two hypotheses with shared control and equal allocation, the FMER has increased from 0.00250 in independent trials to 0.00925. The probability of two errors in the same direction has increased from 0.00125 to 0.00924, which explains almost all of the error, whilst the probability of two errors in different directions is only 0.00001. Recall from Section 4.2.3 that the total error (FWER + FMER) is fixed, thus the increased FMER explains the reduction in the FWER.

4.3.3.3 Multiple Superior False Positives (MSFP) comparison

With two hypotheses, the MSFP rate has increased from 0.06% in independent trials to 0.46% in the multi-arm case with even allocation, an increase of 7.7 times. With three hypotheses, the chance of any two superior false positive outcomes (MSFP2) has increased by nearly 6 times to over 1%, and the chance of three MSFPs (MSFP3) is

substantially greater than in the independent case, although the probability is very small at 0.16%. Again, similar patterns and trends are seen with other allocation ratios. Whilst the absolute differences are small, the relative increases are large. The effect on the MSFP rate is intuitively obvious since a chance 'bad' outcome in the control sample compared to the true population would increase the chances of false positives favouring the experimental treatment in both hypotheses, but the magnitude of this effect is now apparent, and is not trivial.

4.3.4 Validation of results

The above results are based on the assumption that test statistics follow a standard normal distribution, regardless of the distribution of the data. In addition, their joint distribution follows a standard bivariate normal with a correlation that is directly linked to the allocation ratio, as described in Section 4.3.1. Whilst these assumptions have previously been proven in the literature (Follmann et al. (1994)[79] and Dunnett (1955)[54]), the findings within this chapter are validated here in order to ensure their accuracy. Firstly, the results here are compared to those previously obtained for the case that was investigated by Proschan and Follmann, and secondly some results are reproduced using simulations.

4.3.4.1 Proschan and Follmann

Proschan and Follmann (1995)[63] used a different method to calculate the probability of type I errors where there is a dependent control group. As described in Section 3.5.2, they derived the following formula to calculate the probability of type I errors under the dependent case with equal allocation for one-sided tests

$$\Pr(N_D = x) = \left(\frac{1}{\sqrt{2\pi}}\right) \int_{-\infty}^{\infty} \binom{m}{x} [p(z)]^x \times [1 - p(z)]^{m-x} \exp\left(-\frac{z^2}{2}\right) dz,$$

where $p(z) = 1 - \Phi(z + \sqrt{2}Z_\alpha)$, m is the number of comparisons and α is the one-sided significance level. The authors calculated that in the case of two one-sided hypotheses, the chance of exactly one error is 0.076 and the chance of exactly two errors is 0.012; therefore the FWER is 0.088 and the FMER is 0.012. In addition, in the case of three hypotheses, the chance of exactly one error is 0.092, the chance of exactly two errors is 0.022 and the chance of exactly three errors is 0.005; so the FWER is 0.119, the FMER2 is 0.027 and the FMER3 is 0.005. These values were confirmed using numerical integration in SAS to solve the above integral.

In order to validate our results, the R code to calculate the rejection regions based on the multivariate normal, provided in Appendix C, was amended to the one-sided case and re-run. For both two and three hypotheses, the probabilities obtained using the R code matched those obtained by Proschan and Follmann's integral exactly (to within +/-0.001).

Note that only one-sided tests are considered by Proschan and Follmann for simplicity. It isn't clear how this would extend to calculate the probabilities in the case of two-sided tests. The multivariate normal density method presented in this chapter has the advantage that it is easily able to calculate the error rates in the case of two-sided tests as well as where the allocation ratio may deviate from 1:1:1, and is able to output the probabilities of MSFPs. However, Proschan and Follmann's method readily extends beyond three comparisons.

4.3.4.2 Simulations

Simulations were run to verify the accuracy of the results in Table 4-1, with various design assumptions and allocation ratios. Using SAS v9.4 a total of 100,000 simulations were run for each scenario described below. The simulations were set up so that there was no difference between the control and experimental arms, and each comparison was analysed to provide a p-value. The proportion of comparisons for which the p-values were ≤0.05, and therefore a type I error had occurred, were counted and are reported in Table 4-2 below.

The first scenario assumes two independent trials versus a three arm trial with a continuous, normally distributed endpoint. The trials were analysed using two-sample t-tests. With a two-sided significance level of 0.05 and 90% power to assess an effect size of 0.5, assuming a control mean of 100, assessing a difference of 5 with a common standard deviation of 10, the following sample sizes are required:

a) Two independent trials both 1:1 allocation, 86 patients per arm
b) 2:1:1 allocation, 128 patients to control and 64 to each experimental arm
c) 1:1:1 allocation, 86 patients per arm
d) 1:2:2 allocation, 64 patients to control and 128 to each experimental arm

The second scenario assumes two independent trials versus a three arm trial with a survival endpoint. The trials were analysed using log-rank tests of equal exponential

survival in two groups. With a two-sided significance level of 0.05 and 90% power to assess a hazard ratio of 0.75, assuming a 4 year recruitment and 3 year follow-up period, the following sample sizes are required:

a) Two independent trials both 1:1 allocation, 408 patients per arm for 516 events per hypothesis
b) 2:1:1 allocation, 654 patients to control and 327 to each experimental arm for 636 events per hypothesis
c) 1:1:1 allocation, 408 patients per arm for 516 events per hypothesis
d) 1:2:2 allocation, 285 patients to control and 570 to each experimental arm for 526 events per hypothesis

Whilst it would also be of interest to confirm the results for binary endpoint data, it is not possible to get an exact type I error because of the discrete nature of the binomial distribution, causing the type I error and power to 'zig-zag' as the sample size increases[82].

Table 4-2 Simulations to verify the type I error rates under different trial scenarios

	Independent case (Separate trials)	**Dependent case 2:1:1 (2 to control)**	**Dependent case 1:1:1**	**Dependent case 1:2:2 (1 to control)**
Correlation (ρ)	0	0.333	0.5	0.667
FWER: Reject at least one H_0				
Exact algebraic values	0.0975	0.0946	0.0908	0.0849
Simulated using continuous, normally distributed data	0.0969	0.0940	0.0911	0.0852
Simulated using survival data	0.0975	0.0951	0.0914	0.0859
FMER: Reject both H_0's (in any direction)				
Exact algebraic values	0.0025	0.0054	0.0093	0.0151
Simulated using continuous, normally distributed data	0.0024	0.0057	0.0094	0.0148
Simulated using survival data	0.0026	0.0054	0.0098	0.0149

The simulated outcomes match the algebraic calculations to +/- 0.001 in all cases, suggesting that both are correct. Note that due to the error around the simulated results they do not exactly match the algebraic results to four decimal places, but the results are similar enough to be supportive. The simulations confirm that the calculated type I error rates are independent of the trial design, sample size or type of endpoint, because these factors do not affect the binary probability of rejection of the null hypothesis for a given significance level. They also confirm that the correlation calculated based on the allocation ratio is correct regardless of the type of endpoint data. Note that the probabilities of MSFP errors were not calculated for the simulated trials as the error rates are so small it would not be feasible to run enough simulations to assess this accurately, and it was not felt necessary given that the FWER and FMER are correct.

4.3.5 Summary

This section describes how the correlation between the test statistics caused by the shared use of control data affects the different types of error rates in multi-arm trials. The concepts here are not new, but the effects of the correlation in terms of the chance of at least one error (FWER), the chance of multiple errors in any direction (FMER) and the chance of multiple errors in favour of the experimental treatments (MSFP) have been explicitly investigated and quantified in the case of three and four arm trials, which has not been previously reported in the literature.

Where the experimental treatments are compared against a shared control group, the FWER is reduced compared to independent trials with different control groups. However, the chance of a type I error occurring in more than one of the hypotheses (FMER) is increased, where the greater the correlation, the greater the FMER. Since it can be seen from Figure 4-3 that the positive correlation causes the probabilities of false positive outcomes in the same direction (benefitting either the control or experimental therapies in all cases) to increase, it stands to reason that the probability of MSFP errors increases by a greater proportion than the FMER. In order to determine how these findings affect the need for multiplicity adjustment, it is necessary to assess how well multiplicity adjustment methods control each of these types of error.

4.4 The effect of multiplicity adjustment methods on the type I errors

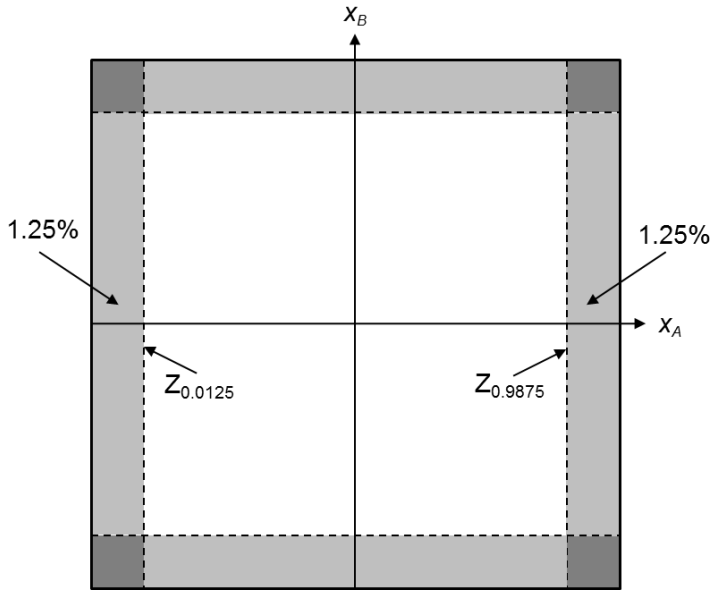
4.4.1 Calculating the type I error rates for common multiplicity adjustment procedures

In Section 3.3, some of the most commonly used methods to adjust for multiplicity were introduced. These included non-parametric methods that do not account for correlation between the comparisons as well as parametric methods that account for the lack of independence due to the shared control group. The methods considered were: Bonferroni[83], a simple, conservative adjustment method; Holm[84] and Hochberg[85], closed testing methods based on a hierarchical strategy of testing the outcomes ordered by significance; Dunnett's t[54], a parametric method that adjusts the Bonferroni boundaries to control the probability of observing a significant result under H_0 at 0.05; and Dunnett and Tamhane[71], an adjusted Hochberg step-up multiple test procedure in which the rejection levels are calculated to account for the correlation so that the final FWER is exactly 0.05. In order to investigate how well these adjustment methods control the various error rates, the FWER, FMER and MSFP, their effects are calculated for multi-arm trials with a shared control group.

4.4.1.1 Bonferroni Adjustment

The probabilities of type I errors after the Bonferroni correction has been applied can be calculated based on the probabilities of falling within the appropriate rejection regions, based on the joint probabilities for the multivariate normal distribution, as for the case where there is no adjustment described in Section 4.2. Since the Bonferroni method simply adjusts the significance level to α/m, for m hypotheses, the error regions can be calculated exactly using the R code in Appendix C, after adjusting the significance levels appropriately. In the case of two hypotheses, the overall two-sided significance level for each hypothesis would be adjusted to 2.5%, so the one-sided error region at each edge would become 1.25%. This can be understood by illustration of the rejection regions, as described by Fernandez and Stone[72] and shown in Figure 4-4. These would be superimposed on the bivariate normal density plots provided in Figure 4-3 to illustrate the probability density in the rejection regions accounting for the different levels of correlation.

Figure 4-4 Rejection regions for two comparisons when using the Bonferroni adjustment method, based on the bivariate normal density plot with the standardised test statistic for the null hypothesis H_{0A} being displayed horizontally, and H_{0B} displayed vertically



X_A and X_B are the test statistics assessing therapies A and B against the control group, and Z_p represents the p^{th} percentile of the cumulative distribution function of a standard normal distribution.

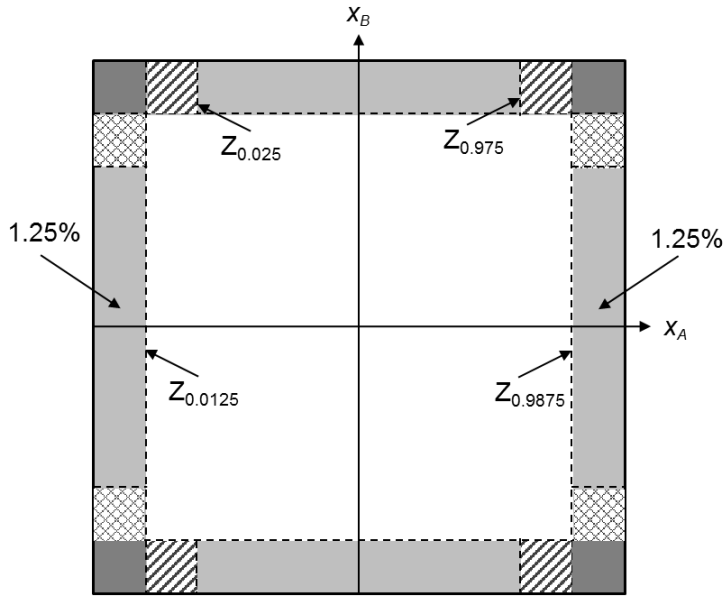
4.4.1.2 Holm Adjustment

The Holm method (Section 3.3.2.2) is a step-down procedure in which the p-values are ordered from smallest to largest, and each p-value (with order position *j, j=1,…,m*) is compared to $\alpha/(m-j+1)$ until a null hypothesis cannot be rejected. In the case of two experimental treatments being compared to control and a FWER set at 5%, the more significant null hypothesis would be rejected if $p_1 \leq 0.025$, and if that is the case, the larger p-value would be assessed and the null hypothesis rejected if $p_2 \leq 0.05$.

The FWER when using the Holm method is the same as it is with the Bonferroni adjustment. This is because the first step is to reject the most significant null hypothesis if the p-value is less than α/m, as is the case with the Bonferroni adjustment, and so the probability of 'at least one' error is equivalent with both the methods. The FMER, however, will be larger than with Bonferroni because the next steps compare the p-values to $\alpha/(m-j+1)$, which is always larger than α/m for $j \geq 2$, and therefore allows a greater chance of multiple errors.

In the case of two hypotheses, this can be understood by illustration of the rejection regions based on the bivariate normal density plots shown in Figure 4-3. The Holm rejection regions are illustrated in Figure 4-5. The FWER is any shaded region around the outside of the square, which is the same as with the Bonferroni adjustment. However, the larger test statistic is now compared to $\alpha = 0.05$, rather than 0.025 as for Bonferroni, so the FMER region is increased to include the striped and checked shaded areas. If hypothesis A is more significant and $p_A \leq 0.025$, the checked areas show the extended rejection area for hypothesis B. If hypothesis B is more significant and $p_B \leq 0.025$, the striped areas show the extended rejection area for hypothesis A. Therefore the total FMER region is the areas of the L-shaped corners, and the probability of MSFP outcomes is illustrated by the top right L-shaped area. The probabilities of the different type I errors can therefore be calculated by amending the R program in Appendix C to calculate the probability densities in the different regions. The increase in the FMER does not affect the FWER since this extended area only relates to the least significant p-value, and therefore the total error (the FWER+FMER) will be higher than with the Bonferroni adjustment, which is why this method is less conservative.

Figure 4-5 Rejection regions for two comparisons when using the Holm adjustment method, based on the bivariate normal density plot with the standardised test statistic for the null hypothesis H_{0A} being displayed horizontally, and H_{0B} displayed vertically



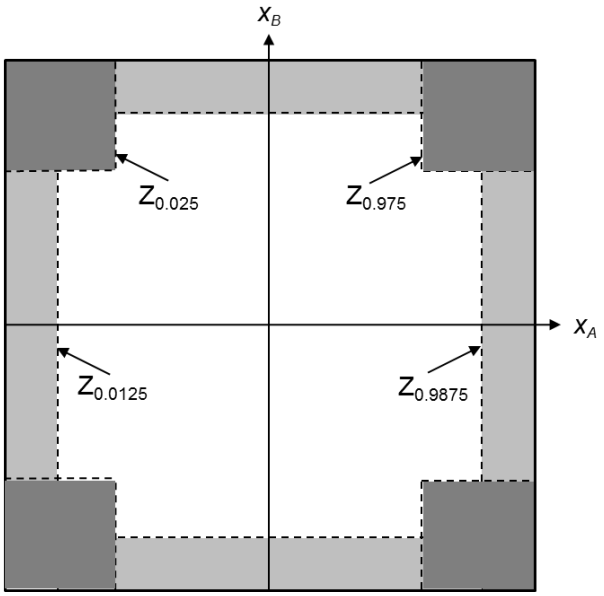
X_A and X_B are the test statistics assessing therapies A and B against the control group, and Z_p represents the p^{th} percentile of the cumulative distribution function of a standard normal distribution.

4.4.1.3 Hochberg Adjustment

The Hochberg method (Section 3.3.2.2) is a step-up procedure in which the p-values are ordered from largest to smallest, and each p-value (with order *j*) is compared to α/j until a null hypothesis is rejected, at which time all further null hypotheses are automatically rejected. In the case of two experimental treatments being compared to control, and a FWER set at 5%, all null hypotheses would be rejected if $p_1 \leq 0.05$, and if that is not the case the more significant would be assessed and rejected if $p_2 \leq 0.025$.

In the case of two hypotheses, this can be understood by illustration of the Hochberg rejection regions as described by Fernandez and Stone[72] and shown in Figure 4-6. The probabilities of the error rates can be calculated by amending the R program in Appendix C to calculate the probability densities in the different regions. The FWER is the probability density in any of the shaded regions, since at least one error is observed if either two-sided p-value is ≤ 0.025 or both are ≤ 0.05. The FMER is the probability density in the darker shaded corner regions, which is the probability of both p-values ≤ 0.05; and the probability of MSFP outcomes is illustrated by the top right darker shaded corner region. Note that the FWER region here is larger than that with the Bonferroni or Holm adjustment methods, and in the case of two independent hypotheses so with no correlation it will equal exactly 0.05 because the area of the double counted corner regions has essentially been replaced.

Figure 4-6 Rejection regions for two comparisons when using the Hochberg adjustment method, based on the bivariate normal density plot with the standardised test statistic for the null hypothesis H_{0A} being displayed horizontally, and H_{0B} displayed vertically



X_A and X_B are the test statistics assessing therapies A and B against the control group, and Z_p represents the p^{th} percentile of the cumulative distribution function of a standard normal distribution.

4.4.1.4 Dunnett's t

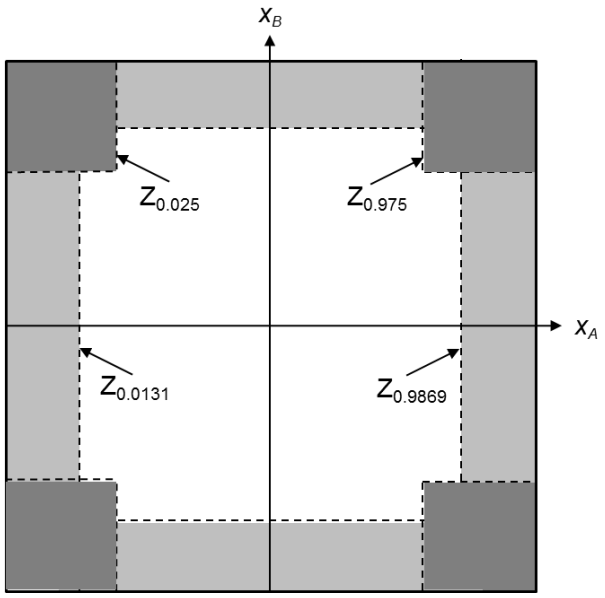
Dunnett's parametric method (Section 3.3.3.1) calculates the exact t-statistic, t_d, required to control the FWER at exactly the required level after accounting for any correlation between the test statistics. These can be read from tables in Dunnett (1955)[54]. With two hypotheses, two-sided tests and equal allocation, and assuming a large sample size (d.f. > 120), t_d is 2.21, which translates to a significance level for comparison of 0.0271. The error rates can therefore be calculated in the same way as they were for the Bonferroni adjustment, using the R code in Appendix C after adjusting the significance levels appropriately. Whereas with the Bonferroni method the one-sided errors were set at 1.25%, with Dunnett's method they are slightly more relaxed at 1.355% so that the rejection regions around the edges are wider than they are in Figure 4-4.

4.4.1.5 Dunnett and Tamhane

The Dunnett and Tamhane procedure (Section 3.3.3.2) is similar to the Hochberg adjustment described in Section 4.4.1.3, but calculates adjusted critical values to control the FWER exactly after accounting for any correlation. The critical values can

be read from tables from Dunnett and Tamhane (1992)[71]. With two hypotheses, two-sided tests and equal allocation, and assuming a large sample (greater than 30) per comparison, both null hypotheses would be rejected if the larger p-value $p_1 \leq 0.05$, and if that is not the case the smaller p-value would be assessed and the null hypothesis rejected if $p_2 \leq 0.0262$. It can be seen from Figure 4-7 that the areas representing the FMER and MSFP are unchanged from the unadjusted Hochberg method, but the FWER area is increased because the edges are wider. The probabilities of the different type I errors can be calculated by amending the R program in Appendix C to calculate the probability densities in the different regions in exactly the same way as for the Hochberg adjustment method.

Figure 4-7 Rejection regions for two comparisons when using the Dunnett and Tamhane adjustment method, based on the bivariate normal density plot with the standardised test statistic for the null hypothesis H_{0A} being displayed horizontally, and H_{0B} displayed vertically



X_A and X_B are the test statistics assessing therapies A and B against the control group, and Z_p represents the p^{th} percentile of the cumulative distribution function of a standard normal distribution.

4.4.2 The effects of applying adjustment methods to the various error rates

The error rates after applying the various adjustment methods were calculated as described in Section 4.4.1 using the software R. In addition, the results (excluding the MSFP rate due to the required number of accurate decimal places) were verified based on 100,000 simulations of the example trial with a survival endpoint described in Section 4.3.4.2 using SAS v9.4. The probabilities matched to +/-0.001 in all cases.

Table 4-3 summarises the effects of applying the adjustment methods on the various error rates, using the example of a three-arm trial with 1:1:1 allocation in which the two experimental arms are compared to a shared control group. The two-sided α is set at 0.05 for each unadjusted comparison, and the FWER is required to be controlled at 0.05 when adjustment methods are used.

Table 4-3 FWER, FMER and MSFP comparisons for three arm trials with two hypotheses (α=0.05 for each), a shared control group and even allocation ratio, after applying various multiple testing adjustments

	Indepen-dent case	Dependent case, 1:1:1 allocation					
		Un-adjusted	Bonf-erroni	Holm	Hoch-berg	Dunnett's t	Adjusted Hoch-berg
Reject H_0 for each individual hypothesis	0.0500	0.0500	0.0250	0.0271	0.0286	0.0271	0.0296
FWER: Reject at least one H_0	0.0975	0.0908	0.0465	0.0465	0.0480	0.0502*	0.0500
FMER: Reject both H_0's (in any direction)	0.0025	0.0093	0.0035	0.0077	0.0093	0.0039	0.0093
MSFP: Reject both H_0's in favour of treatments *A* and *B*	0.00063	0.00462	0.00176	0.00385	0.00462	0.00197	0.00462

**Note that FWER with the Dunnett's t adjustment is 0.0502 rather than exactly 0.0500 due to rounding in Dunnett's table, which provides t-statistics to 2 decimal places.*

4.4.2.1 FWER

All adjustment methods control the FWER at 0.050 or less, as expected, with the range being between 0.047 and 0.050. In all cases, the chance of rejecting the null hypothesis for each individual comparison has taken a penalty compared to running independent trials, with the probabilities ranging from 0.025 to 0.030. The Dunnett's t and Adjusted Hochberg methods account for the effect of the correlation due to the shared control data on the FWER and so control the FWER at exactly 0.050. The least conservative methods for the individual hypotheses are the Hochberg and Adjusted Hochberg methods, with Bonferroni being most conservative.

4.4.2.2 FMER

Although all adjustment methods control the probability of falsely rejecting at least one hypothesis (the FWER) as required, no method fully controls the chance of multiple errors occurring within the same set of hypotheses to be what it would have been if the hypotheses had been assessed in independent trials. The FMER with two independent trials is 0.0025, and in the multi-arm case with two hypotheses and no adjustment it increases to 0.0093. It can be seen that the FMER after multiple testing adjustment ranges from 0.0035 to 0.0093. With the Bonferroni and Dunnett's t methods, the probabilities of multiple errors are reduced towards those in independent trials, but the Holm, Hochberg and Adjusted Hochberg methods based on the closed testing principle offer very little or no protection of the FMER over no adjustment. The first step of a step-up procedure is to reject all null hypotheses if the least significant hypothesis has $p < 0.05$, so it can easily be seen why this is the case.

4.4.2.3 MSFP

Since the adjustment methods do not control the FMER, they also do not offer full protection against the chance of MSFP outcomes. After applying the Bonferroni and Dunnett's t corrections, the chance of two superior false positive errors is still inflated by approximately 3 times over that with independent trials (0.0018 and 0.0020 compared to 0.0006), and following the Hochberg adjustments the probability is increased by over 7 times to 0.0046, the same as with no adjustment.

The above results highlight that multiple testing adjustment methods only control the probability of the overall FWER to that for a single hypothesis. They do not offer effective control over the chance of multiple false positive errors, which is the type of error that is increased over that had the hypotheses been assessed in independent trials. The impact of these findings on the need for adjustment in multi-arm trials is discussed in Section 4.6.

4.5 Adjustment of the significance level to control the MSFP

In Section 4.3.3 it was shown that in the case where two superior hypotheses may both be used to jointly inform a claim of effectiveness, the overall chance of both having a false positive outcome in favour of the experimental treatments (MSFP) is inflated in a multi-arm trial over that chance occurring in independent trials. In addition, Section

4.4.2 confirmed that applying multiple testing correction methods does not reduce the chance of MSFP outcomes to the same level as in two independent trials. However, there may be cases where it is beneficial to control the MSFP rate. The FDA guidance on 'Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products'[86] suggests that it is feasible for multiple hypotheses from within a single study to be accepted as evidence of effectiveness if the trial is designed appropriately. In addition, Fisher[80] discusses "one large, well-designed, multicentre study as an alternative to the usual FDA paradigm". In this paper, he considers one large study with double the patient numbers in place of two independent studies in order to inform regulatory applications. Although this scenario is different from a multi-arm trial, his argument that the strength of evidence is the same as for two studies if the "probability of two statistically significant positive trials when there is no treatment effect" is set to be equivalent to that in two studies (i.e. 0.025 x 0.025 = 0.000625), is also relevant in the multi-arm situation. That is, the evidence can be considered as strong as that obtained from two separate trials as long as the overall probability of multiple conclusions of superiority (MSFP) is not inflated over that for independent studies.

The example of the AMAGINE-1 trial described in Section 4.1.2.2 assesses two doses of an experimental treatment against placebo. If these doses were investigated in independent trials, both trial outcomes may be used to inform a claim of effectiveness, but the penalty for assessing these within a multi-arm trial in terms of inflation of the MSFP rate has not been investigated or quantified.

4.5.1 Significance levels to control the MSFP rate in the case of two hypotheses with a concurrent shared control group

In two independent trials, the chance of two superior false positive outcomes is 0.000625 (Section 4.2.4). Since the joint distribution can be described using a bivariate normal (Section 4.2.1), this can be used to obtain the exact significance level that returns a probability of 0.000625. This principle is similar to the work of Follmann et al. (1994)[79], which relies on the multivariate normal assumption of the test statistics to estimate critical values that strongly protect the type I error rate in the case of multi-armed trials with interim looks. The R code to calculate the significance level required to control the probability of two MSFP errors based on a multi-arm trial with concurrently recruited arms to that for independent trials is provided in Appendix C.3.

In the 1:1:1 case, the significance level required to protect the MSFP rate at 0.000625 is 0.0118. In the 2:1:1 case it is 0.0195, and in the 1:2:2 case it is 0.0069, as shown in Table 4-4. Note that if the correlation is set to 0, the significance level returned is 0.05 as expected, as it would be in two independent trials. If two hypotheses are assessed in a multi-arm trial with a concurrent shared control group, and are to be used to jointly inform a claim of effectiveness; in order to control the probability of two superior false positive outcomes to the level in independent trials, the p-values for both hypotheses are required to be less than these adjusted significance levels. It can be seen that with this level of control, the FWER is reduced to much lower than 5%. In the 1:1:1 case, when the significance level is reduced to 0.0118 to protect the MSFP rate, the FWER decreases to 2.24%.

Table 4-4 Adjusted significance levels to control the chance of a MSFP error in a three-arm trial to that for two independent 1:1 randomised trials

	Independent case No adjustment	**Dependent case No adjustment**	**Dependent case 2:1:1 α = 0.0195**	**Dependent case 1:1:1 α = 0.0118**	**Dependent case 1:2:2 α = 0.0069**
Reject H_0 for each individual comparison	0.050	0.050	0.0195	0.0118	0.0069
FWER: Reject at least one H_0	0.0975	0.0908	0.0377	0.0224	0.0125
MSFP: Reject both H_0's in favour of treatments *A* and *B*	0.000625	0.00462	0.000623*	0.000624*	0.000628*

**Note that MSFP rates are not exactly 0.000625 due to rounding of the significance levels to 4 decimal places.*

If an arm is added part way through recruitment so there is less overlap of the shared control patients, the adjustment required to control the MSFP rate will be less stringent. This would need to be calculated on a case by case basis depending on the amount of overlap and the resulting level of correlation between the test statistics for the two hypotheses.

4.5.2 The effect of MSFP control on the power and sample size

If a trial is designed to allow two superior outcomes to be used as evidence to inform a single claim of effectiveness by controlling the MSFP rate, the power is required to be

maintained for each hypothesis as it would for independent trials, requiring an increased sample size. As an example, take a confirmatory trial with a survival primary endpoint and analysis based on the log-rank test for equality assuming an exponential survival distribution. The estimated median survival in the control group is 36 months, and a clinically relevant difference would be an improvement to 48 months (HR=0.75). In a two-arm trial, with a recruitment period of 48 months and an additional 36 month follow-up period, 408 patients are required per arm (1:1) to achieve 516 events for 90% power with a two-sided type I error rate of 5%. If there are two experimental arms of interest in the population, a three-arm trial may be considered rather than two independent trials. In the scenario of running independent trials, the total sample size for the two trials assuming 1:1 allocation would be 1632, and in the multi-arm trial this is reduced to 1224 with no adjustment. The following results were obtained from the software Stata (StataCorp, v13.1) by setting the stpower command. Adjusting the significance level to 0.0118 to control the chance of MSFP outcomes reduces the power from 90% to 77%, and to account for this loss in power the sample size would need to be increased to 1680 (with 708 events required per hypothesis), which makes the multi-arm trial slightly larger than running two independent trials. For comparison, with a Bonferroni adjustment the power is reduced to 84%, requiring 1443 participants (with 610 events required per hypothesis) for 90% power.

The greater the dependence of the control data, the more impact this will have on reducing the significance level and therefore increasing the sample size. For a 2:1:1 randomisation with 90% power and no adjustment, 1308 patients would be required (654 to control and 327 per experimental arm). Adjusting the significance level to 0.0195 to control the MSFP rate requires 1628 participants for 90% power (814 to control and 407 per experimental arm), which is fewer patients than would be needed for the 1:1:1 case with MSFP adjustment. In the 1:2:2 case with 90% power and no adjustment, 1425 patients would be required (285 to control and 570 per experimental arm). Adjusting the significance level to 0.0069 to control the MSFP rate requires 2155 participants for 90% power (431 to control and 862 per experimental arm).

Even though the sample size may be slightly larger in the adjusted multi-arm case compared to independent trials following MSFP adjustment (an increase of 3% in this example with even allocation), there may still be benefits to running a multi-arm trial in terms of reducing the total number of patients receiving the control therapy as well as the time and cost of only needing to set-up and run a single trial.

4.6 Discussion

4.6.1 FWER adjustment due to shared control data

When an experimental treatment is added to an ongoing trial, the stage of the trial after the amendment will have multiple primary comparisons based on a shared control group. Therefore consideration of a multiple testing adjustment is an essential requirement when adding an arm, as described in Chapter 2. In addition, multi-arm trials in general are efficient and can therefore be advantageous over running independent trials, thus this research has wider implications than solely for trials in which treatment arms are added. Currently there are conflicting views in the literature on how to appropriately control the probability of a false positive error. A lack of proper control of the FWER could lead to an unacceptable chance of an ineffective treatment being recommended to be taken forward into practice; but unnecessary control of the FWER could affect the efficiency of a trial, requiring increased patient numbers and resources. FWER adjustment without increasing the sample size to maintain power could lead to a superior treatment being denied. Each of these scenarios raises ethical concerns.

One reason that false positive error rates may be affected in multi-arm trials compared to independent trials is due to correlation between the test statistics caused by the shared use of the control data. It is a common misconception that the FWER is increased due to sharing control data compared to that in independent trials. When considering the designs illustrated in Figure 3-2 and Figure 3-3, some might assume that the overall FWER for the family of hypotheses, H_{0A} and H_{0B}, would be larger in Figure 3-3 where there is a common control group. However, it has been confirmed here that the FWER is in fact smaller in Figure 3-3 than in Figure 3-2. The common control group instead has the effect of increasing the chances of more than one false positive outcome within the family of hypotheses; although FWER adjustment methods do not specifically control for this. Adjustment methods reduce the probability of multiple errors to varying extents, but none to the same levels as in independent trials. Closed testing methods, in fact, exploit the FWER being defined by the chance of 'at least one' error, by allowing a higher chance of multiple errors in order to reduce their conservativeness. In this case, if there is a bad shared control group by chance, these methods do very little to prevent a subsequent false positive finding given that a type I error has already been made in the case of two hypotheses, yet they are acceptable and recommended in the literature on multiple testing. This suggests that FWER

adjustment is only required to control the probability of at least one error, and not to reduce the inflated probability of multiple errors.

In summary, since the FWER is not inflated, and FWER adjustment methods do not aim to offer control over the inflated chance of multiple errors, it is difficult to justify FWER adjustment in trials with multiple hypotheses because of the shared use of a control group. These findings remain valid in the case of imbalanced randomisations, and also where there are more than two experimental therapies, although in the latter case the chances of multiple errors become increasingly difficult to consider depending on whether the interest would be in controlling the probability of at least two errors or at least three errors and so on. This is considered further as part of the discussion in Chapter 7.

Note that if the experimental therapies are competing against each other for approval in the trial population, the correlation due to the shared control group in a multi-arm trial is an advantage. The reasoning is as follows, if in two independent trials one of the control samples performs worse than the true population, the associated experimental group has an increased chance of being significant and taken forward. However, in the equivalent multi-arm case, the lack of comparability due to comparisons to different control samples is removed. It is more likely that efficacious experimental therapies would be considered against each other by decision makers directly without the influence of variations in the control samples.

4.6.2 FWER adjustment due to assessing multiple hypotheses

The other factor concerning the necessity for FWER adjustment is whether assessing multiple hypotheses within a multi-arm trial increases the chance of making a single claim of effectiveness. Phillips et al. 2013[57] describe this as the ‘claim-wise error rate’, defined as “the family-wise error rate when the families relate to multiple clinically important endpoints that need to be described in the label”, with consensus from the PSI expert group that this is “probably the most important attribute to control”. In order to determine whether the increased chance of a familywise type I error is relevant, it is helpful to also consider whether it makes sense to have a ‘claim-wise’ power for the trial, which is the probability of correctly rejecting the null hypothesis in at least one comparison across the trial, leading to a claim of effectiveness. This has been described in the literature as the ‘minimal’[87] or ‘disjunctive’[75] power, and is the appropriate power to consider alongside FWER adjustment. If it does not make sense

to consider the claim-wise power for the trial as a whole because the outcomes for the hypotheses do not inform the same claim of effectiveness, for example if the experimental arms are assessing different therapies, it also follows that control of the FWER is unnecessary. In Chapter 5 it is shown using simulations that with two hypotheses and 80% power for each independently, following a Bonferroni adjustment the power for each hypothesis drops to 72%, but the power that at least one null hypothesis is rejected in the case that both treatments are actually superior is 89.5%. Therefore, the penalty caused by applying the Bonferroni adjustment may be compensated to some extent by the claim-wise power. Note that in Figure 3-2 and Figure 3-3, both designs have an increased chance of an incorrect claim due to testing two hypotheses, and so the need for FWER adjustment due to more hypotheses being included than would have otherwise been assessed in separate protocols is present in both cases.

4.6.3 Multiple superior false positive adjustment

The increased risk of multiple errors within the family of hypotheses due to the shared control group could be important if superiority in more than one of the hypotheses were to contribute as separate pieces of evidence towards a claim of effectiveness. In the case of two hypotheses, FWER adjustment methods are not stringent enough, and the probability of multiple superior false positive (MSFP) outcomes needs to be controlled for the evidence to be equivalent to that obtained from two independent trials. In this situation, the significance level adjustment proposed in Section 4.5.1 to control the chance of MSFP outcomes in a three-arm trial in order for the evidence to be equivalent to that obtained from two independent trials can be used. It is possible that controlling for MSFPs could lead to a trial that is as large or larger than independent trials, and this should be considered during the design stages. The overall probability of MSFP outcomes is low even where there is shared control data, and unless the outcomes are to be used as multiple pieces of evidence towards the same claim of effectiveness, adjustment is unlikely to be necessary. In the case of more than two hypotheses, it would need to be considered whether having an inflated chance of two or more superior false positives out of the set of hypotheses is important. It is unlikely to be necessary to control the chance of all of the hypotheses being falsely superior as this will become very small, but it might be beneficial to control the probability of at least two superior false-positive errors, if these could inform the same claim of effectiveness. Whilst adjusting the FWER might somewhat control this probability, since this is not the aim of this adjustment it may not control it appropriately, depending on the number of hypotheses and the adjustment method. This would benefit from

further consideration and is an interesting area for further work, as discussed in Chapter 7.

The concept of controlling the probability of multiple errors was also recently discussed by Jaki and Parry (2016)[88]. They introduce a metric E(V), the expected number of false rejections, which is equal to the FWER + FMER in the notation of this thesis. They discuss that instead of controlling the FWER, it may be more relevant to control the expected number of false claims (EFC) in circumstances where the chance of more than one wrong rejection will "result in a consequential decision" to make a claim of effectiveness. They note that "no method exists that explicitly considers claims" and go on to suggest a method to control the EFC in some situations. They note that using closed testing methods to control the FWER gives a "false sense of security" in terms of the EFC, agreeing with our findings. The conclusions of this manuscript agree with the recommendations in this thesis that multiple testing adjustment should be determined based on the requirements for the resulting claim of effectiveness, as summarised in Section 4.6.6.

4.6.4 Multiple testing adjustment considerations when adding an arm

In trials where arms are added and only concurrent control data is used, there is likely to be a combination of shared and independent control data for the hypotheses. Since having shared control data does not inflate the FWER, the need for adjustment to control the probability of at least one type I error is no different than it is for trials in which all the control data is concurrent. The efficiency of sharing a protocol and the ability to test more hypotheses leading to a claim of effectiveness than in independent trials is unaffected by the time at which the arm is added. Therefore, the requirement for FWER adjustment is the same as for standard multi-arm trials. Parmar et al. (2017)[22] discuss adjustment with relation to the Stampede trial, and believe that if there is relatively little overlap in the control groups, there is less need to adjust the FWER. However, this assumes that shared control data inflates the FWER, which is not the case. If there is a lower proportion of shared control data, however, the probability of MSFP errors will not be inflated to the same extent, and would therefore require a less stringent adjustment to the critical value, if necessary. In Chapter 5, FWER adjustment alongside analysis methods when an arm is added is considered.

4.6.5 Motivational examples

Returning to the trials introduced in Section 4.1.2; recall the MRC COIN trial[78] (Section 4.1.2.1) in which OxFP is present in all treatment arms. Since one primary hypothesis addresses the addition of the experimental therapy cetuximab to OxFP and the other addresses a reduction in duration of OxFP therapy, these do not contribute towards the same claim of effectiveness. In this case, the chance of a false positive outcome for either the addition of cetuximab or the change in the dosing schedule is not increased by the presence of the other hypothesis. Similarly, the power to make a claim of effectiveness based on either hypothesis is not inflated by the presence of the other. Since the correlation between comparisons due to the shared control group does not increase the FWER, adjusting to control the FWER in this case is therefore an unnecessary penalty.

On the other hand, the AMAGINE-1 trial introduced in Section 4.1.2.2 assesses two doses of brodalumab compared to placebo. Since a rejection of the primary null hypothesis for either comparison could lead to a claim of effectiveness for brodalumab, there are two chances for a false positive result with respect to that claim. In this case, there is general agreement in the literature that FWER control is recommended[59] since the type I error rate can be considered for the claim of effectiveness as a whole, rather than for each individual hypothesis. The power for each comparison will be reduced from 90% to around 84% following a multiple testing correction without inflation of the sample size, but the claim-wise power for superiority of either dose is relevant, and will be increased above 90% if both doses are truly superior. If the outcomes from the AMAGINE-1 trial were to be used as two separate claims of effectiveness for brodalumab, a more stringent adjustment would be required for the evidence to be equivalent to that obtained from two independent trials, with the significance level reduced to 0.0118. In this case, the sample size to maintain 90% power would need to be increased from 660 to 900, which is similar to the 880 patients that would be needed in two independent trials.

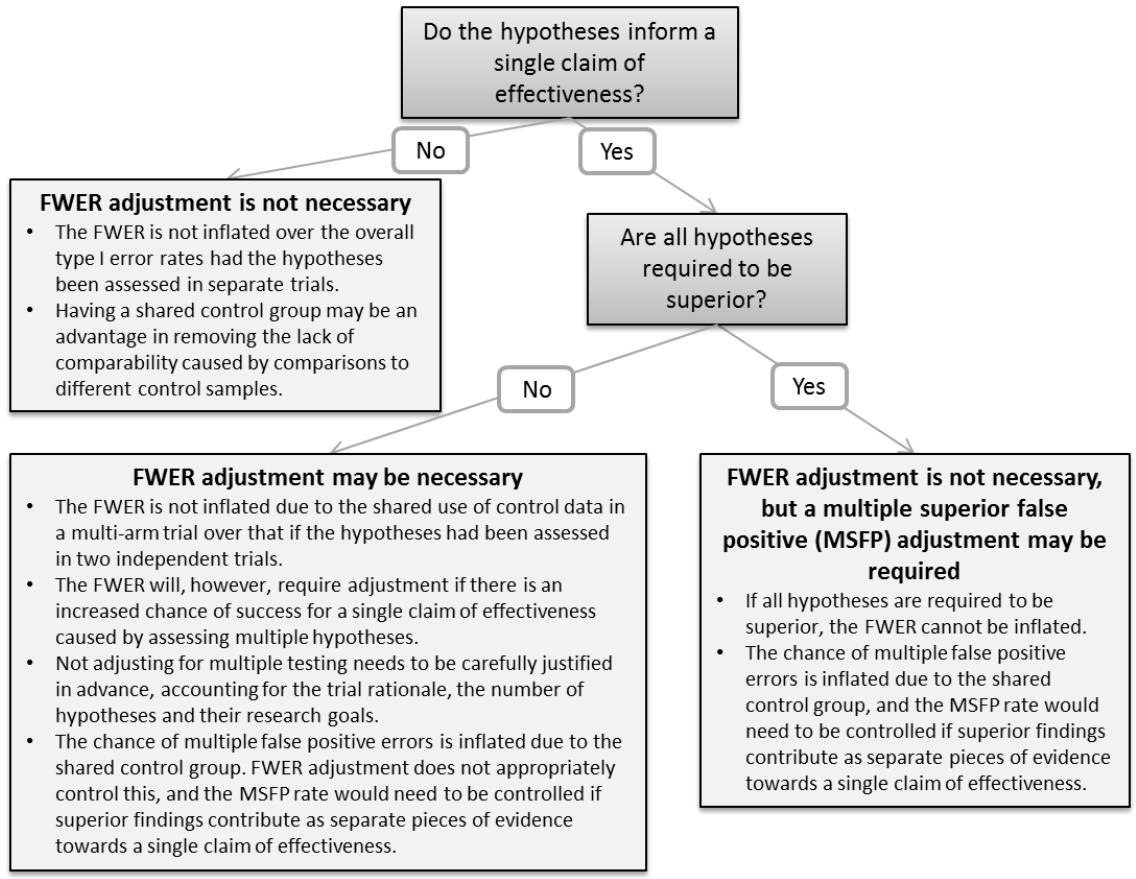
In the Myeloma XI+ Intensive trial introduced in Section 4.1.2.3, the four-drug regime CCRD is compared against the current standard control CTD, as well as the previously assessed three-drug regime CRD. Since CCRD will only be recommended for approval if it is better than both CTD and CRD, both hypotheses are required to be significant in order to recommend CCRD for use in practice. Here there is only one chance for an

overall false positive outcome for the trial, so the chance of 'at least one' error cannot be inflated, and therefore no type I error adjustment is necessary.

4.6.6 Decision diagram

The implications of running a multi-arm trial with shared control data on various types of false positive error rates have been formalised here, considering the effects of multiple testing adjustment methods, in order to make informed recommendations on the requirement for adjustment. A flow diagram to aid the determination of the requirement for a multiple testing adjustment in a multi-arm trial is provided in Figure 4-8. The decisions on the need for error control with respect to the interpretation of the trial results should be agreed and documented in advance in the protocol and statistical analysis plan. Care should always be taken in reporting and interpretation if more than one hypothesis within a multi-arm trial with shared control group is positive.

Figure 4-8 Decision diagram to determine the requirement for a multiple testing adjustment in multi-arm trials



Chapter 5
Analysis methods when adding an arm to an ongoing trial

5.1 Introduction

5.1.1 Rationale

When a new treatment arm is added to an ongoing trial, there is a change to the original study design. How to account for this type of adaptation has rarely been considered in the literature, and it not clear how to analyse the data appropriately to ensure that bias is not introduced, the type I error rate is not inflated, the power is maximised and the results are meaningful.

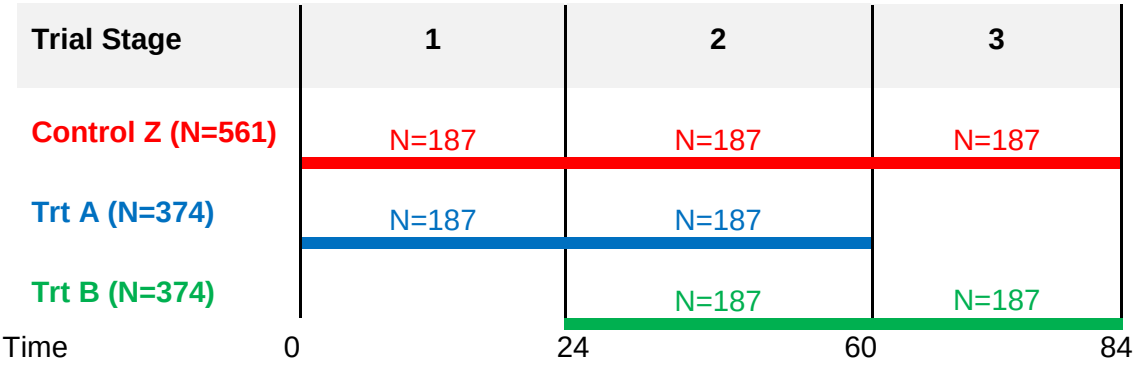
In order to protect trials from bias due to design adaptation, adaptive designs literature usually recommends conducting an adjusted analysis to account for interim analyses that use internal trial data to inform the amendment. However, the situation considered here differs because the amendment to add a new treatment arm is likely to be based on emerging evidence that is external to the trial which is being adapted. Similarly to adaptive designs based on internal data, the amendment might have consequences for the hypotheses of interest within each stage, causing a stage effect. In addition, although the evidence to add an arm is assumed to have arisen externally, data internal to the trial may inform the decision to make the amendment or its design characteristics. Depending on the nature of the experimental therapies, multiplicity adjustment may or may not be necessary. In this chapter we aim to investigate various analysis methods in order to make recommendations for researchers when planning the analysis of a trial they wish to amend by adding a new treatment arm. Firstly the sources of potential bias when analysing a trial that is adapted by adding an arm are considered alongside current literature on this topic in order to inform the necessary research within this chapter. Two adaptive and two non-adaptive analysis methods are introduced Section 5.2, and in Section 5.3 literature that has previously considered these analysis methods in trials amended by adding an arm is reviewed and limitations discussed. A simulation study is undertaken to compare the error rates following analysis using the four methods in Section 5.4, and these are considered alongside

multiplicity adjustment in Section 5.5. The findings are discussed in Section 5.6 and a summary diagram of the recommendations is provided.

5.1.2 Scenario

This work is motivated by the design of the FLAIR trial in CLL, which is introduced in Section 1.3 and discussed in detail in Chapter 6, although the scenario presented here is simplified. Consider a randomised, parallel arm clinical trial assessing a null hypothesis H_{0A}, comparing Treatment *A* against Treatment *Z*, the control. The trial is powered to assess progression-free survival (PFS) in 748 participants (374 per arm) to observe 379 events in order to assess a superiority hazard ratio of 0.75 with an overall 5% significance level and 80% power assuming a 4 year recruitment and 4 year follow-up period and allowing for a 5% dropout rate. During recruitment, Treatment *B* is identified as extremely promising from an early phase trial for the population being assessed, and it is desirable to compare this treatment in the same population to the same control group. In order to maximise resources, Treatment *B* is added to the trial, so that patients are randomised between the control treatment (*Z*) and the two experimental treatments (*A* and *B*). The new primary comparison assessing null hypothesis H_{0B} will be between Treatment *B* and the control Treatment *Z*, and pairwise comparisons between treatments *A* and *B* are not planned. This reflects what would likely have been the case if Treatment *B* was assessed in a new trial against the current standard Treatment *Z*. Hypothesis *B* is powered to assess the same improvement in PFS as the Treatment *A* to Treatment *Z* comparison, also requiring 748 participants to be randomised to treatments *Z* and *B* concurrently. Randomisation allocation will be even during each stage (1:1 for two arms and 1:1:1 for three). It is planned to stop recruitment to Treatment *A* once the required numbers have been randomised to that comparison, with the trial then becoming Treatment *B* vs *Z* in Stage 3. Note that this is not in any way related to the efficacy of Treatment *A*, and no efficacy analysis will take place at this time because the survival data will not yet be mature. Therefore, the overlap between concurrent recruitment to treatments *A* and *B* will be comparatively small. The total sample size will be less than the 1496 that would have been required in separate trials, depending on when the new arm is added. For example, if the new treatment was added halfway through the planned recruitment, the total sample size would be 1309, as illustrated in Figure 5-1:

Figure 5-1 FLAIR randomisation by treatment arm and trial stage



It is assumed that the rate of recruitment remains constant and that the two hypotheses will be analysed at different times, when there are enough events in the concurrently randomised arms. In the case where the new arm is added halfway through recruitment, the 4 year recruitment period to assess hypothesis *A* is increased to 5 years because of the additional arm. However, the follow-up period to reach the same number of events will decrease from 4 to 3.25 years, so the total time to reach the primary outcome for hypothesis *A* will only be increased by 3 months overall, from 8 to 8.25 years.

The focus of this work is on the analysis of Hypothesis *A* over stages 1 and 2 due to the addition of Treatment *B*. However, the findings relating to potential stage effects and multiple testing adjustment will also be relevant to the analysis of Hypothesis *B* over stages 2 and 3. Whilst the effect of dropping a treatment arm is widely researched and published, this type of adaptation is usually following an interim analysis with stopping or selection criteria, so the situation of stopping an arm simply because it has completed recruitment and without analysis of interim data is an unusual feature of a platform (or MAMS) trial.

5.1.3 Considerations for the use of adaptive analysis approaches

Wassmer and Brannath (2016)[36] introduce a common basic principle in flexible designs known as the Conditional Invariance Principle. They define this as follows: “Think of a trial with two sequential stages, where design characteristics of the second stage are chosen at an interim analysis based on data from the first stage as well as external information. The design of the first stage is pre-fixed and remains unaltered. Assume further that the first and second design stage data are from independent cohorts of

patients. If the trial continues to the second stage, let T_2 be the statistics for H_0 from the second cohort recruited after the interim analysis. Due to the data-driven choice of design features, the null distribution of T_2 will in general depend on the interim data. However, we can often transform T_2 in a way that the conditional null distribution of T_2 given the interim data and the second stage design equals a fixed pre-specified null distribution, and hence is *invariant* with respect to the interim data and mid-trial adaptations." Depending on whether the Conditional Invariance Principle is relevant or not when adding an arm, it may be necessary to use specific methods of analysis which we will refer to as 'adaptive analysis methods'. In the context of the work of this thesis, the approach taken is p-value combination methods, as described in Section 5.2.3.

The key statements in the above definition in the context of adding a treatment arm have been underlined. It is not clear whether adaptive analysis methods are required in the case of adding an arm, because there is no adaptation to the existing Hypothesis *A*, only to the protocol and randomisation, and it is assumed that the amendment is primarily informed by information that is not obtained from within the trial. Wassmer (personal communication) commented "If no (and really no) information from the current trial was used for the decision to add a new treatment, an analysis that considers the treatment as if it was recruited from the beginning of the trial needs no p-value combination method. It should be clarified with the agency (if necessary), if this point of view is accepted from a regulatory point of view". This suggests it may therefore be appropriate to pool the data over stages if the amendment to add a new treatment arm is only informed by evidence that is external to the trial. However, if any analysis of interim data internal to the trial has informed the decision to add the new hypothesis or its design characteristics, this may need to be accounted for in the final analysis using adaptive analysis methods such as p-value combination over the stages.

Adaptive analysis methods could also be beneficial to account for any stage effects caused by the amendment. If the addition of the new treatment arm changes the characteristics of the trial in some way, this could alter the treatment effect in the second stage. For example, a change in the eligibility criteria might be necessary, shifting the patient population. Even without a formal change to eligibility, if, for example, the new arm has perceived increased toxicity, frailer patients might be discouraged and therefore the population would shift towards being younger and fitter, which could improve the outcomes in the second stage leading to a stage effect.

Conversely, very promising phase II results published for the new therapy could encourage different patients into the trial. In these cases, simply pooling the data could bias the results, but a multivariable analysis adjusting for stage might be appropriate as opposed to necessarily requiring p-value combination methods. In addition, if it is determined that multiplicity adjustment is required due to assessing multiple hypotheses in the same protocol, it is not known which of the analysis methods optimise the power alongside adjustment in the case where the hypotheses are not concurrent. In this chapter, the various analysis methods will be reviewed and the error rates compared, including after incorporating adjustment for multiple testing.

5.1.4 Sources of error rate inflation in adaptive designs

P-value combination methods alone do not control for other sources of multiplicity such as repeated hypothesis testing in the case where there is an interim analysis, or for assessing multiple hypotheses within the same trial, but these adjustments can be made in conjunction if appropriate for the trial design. Standard adjustment techniques can be planned into the design as described by Maurer et al. (2010)[89] using the following table:

Table 5-1 Sources and control of type I error rate inflation in adaptive designs (adapted from Maurer et al. 2010)

Sources of potential error rate inflation	Techniques for error rate control
Repeated hypothesis testing with early rejection of null hypothesis at an interim analysis.	Classical group sequential designs, e.g., designs based on the α-spending approach.
Adaptation of design and analysis features with combination of information across trial stages.	Combination of p-values, e.g., the inverse normal method, Fisher's combination test; conditional error function approaches; adjustment for known adaptation rule.
Multiple hypothesis testing, e.g., comparing multiple experimental treatments with a control.	Classical multiple testing methods, e.g., the closure principle or Bonferroni method.

Rejecting a null hypothesis at an interim analysis is outside the scope of this research since we are only investigating the addition of a new therapy based on emerging external evidence, and not any formal interim analyses. However, the second and third rows are both relevant to consider when adding an arm: the addition of an arm is an adaptation of the design by definition; and adding a new hypothesis will always imply that multiple hypotheses are being tested within the same trial. In the previous chapter it was shown that multiplicity adjustment for multiple hypotheses is necessary in some situations but not others, so both scenarios are included in the investigations. Where

more than one source of error rate inflation is possible, the adjustment methods to control each one need to be applied in combination.

5.1.5 Literature on analysis methods when adding an arm

Recall from Section 2.2.2 that there are very few papers that discuss analysis methodology when adding a treatment arm to an ongoing trial. Hommel (2001)[15], Posch et al. (2005)[18] and Bauer (2008)[19] mention adding an arm or hypothesis as being possible within a flexible framework. In all cases the arm is assumed to be added at an interim analysis, such that it is necessary to account for the Conditional Invariance Principle, and p-value combination methods are recommended to achieve this. Wason et al. (2012)[24] assume the treatment is added at a pre-planned interim analysis and focus on strong family-wise error rate control due to having multiple arms, adjusting the existing group sequential stopping bounds to account for the additional hypothesis rather than considering analysis by trial stage. This methodology would not be applicable when adding an arm at an unplanned time based on external data. Elm et al. (2012)[20] is the only publication found to have considered analysis methods when adding an independent treatment based on external considerations, comparing p-value combination methods to those where the data are pooled over the stages. Whilst this research is very relevant, there are a number of specific assumptions which prevent the extrapolation of their results to make recommendations to researchers adding an arm more generally. This work is reviewed in detail below in Section 5.3.

Of the eight confirmatory trials that were identified to have added new treatment arms for which details were available Section 2.2.3, none analysed the data by stage using adaptive methods. Instead, they pooled the data over the stages, with one trial accounting for stage within a multivariable analysis but the others not. One trial tested for intra-stage correlation within the analysis, although the power for that test would have been low, and then pooled the data over the stages without adjustment. None of the arms were added based directly on interim trial data; all were identified through external evidence. However, some trials had reported looking at interim data before the new arm was added, particularly those with multi-arm multi-stage designs that are designed specifically with the intention of dropping and adding arms throughout the recruitment period based on a series of interim analyses. No interim data were reported to have informed the design amendment.

In summary, very little literature has considered analysis methods when a trial is adapted by adding a new treatment arm, and the appropriate analysis methods to ensure statistical validity when adding a treatment arm are currently unclear.

5.2 Analysis methods investigated

5.2.1 Introduction

As discussed in Section 5.1.3, it may be necessary to use adaptive analysis approaches in order to satisfy the Conditional Invariance Principle. The p-value combination approach, where the trial stages are analysed separately and p-values subsequently combined, is a convenient and appropriate method to achieve this in the situation of adding an arm. This is because it is based on stagewise rather than cumulative test statistics, which are not required to be calculated in advance of the final analysis and are not based on a decision arising from an interim analysis. The approach is simple and flexible to implement, and can be planned at the time of the amendment rather than necessarily at the outset of the trial.

If adaptive analysis approaches are not required, the analysis could simply pool the data over the stages as if there had been no amendment, or else pool over stages but include ‘trial stage’ as a covariate in a multivariable model. Each of these methods will be considered when analysing the original hypothesis having added a new treatment arm part way through recruitment, and their effect on the type I error and power will be investigated under various scenarios. The aim is to aid researchers in understanding the impact of each analysis method so they can determine the appropriate method for their particular situation.

Once Treatment *B* is included in the randomisation in Stage 2, there are two hypotheses being assessed within the same protocol and using some of the same control group, therefore multiple testing adjustment for multiple hypotheses needs to be considered. The need for adjustment in this case was discussed in detail in Chapter 4, and it was determined that if the hypotheses assess therapies that inform different claims of effectiveness, no adjustment is necessary. Therefore it is first assumed that this is the case. In Section 5.5.4 this assumption is changed and multiplicity adjustment methods are investigated alongside the analysis methods.

5.2.2 Pooled, non-adaptive analyses

The scenario in Section 5.1.2 has a time-to-event primary endpoint assessing the two-sided null hypothesis H_{0A} that there is no significant difference between Treatment *A* and Control Treatment *Z* in terms of Progression Free Survival. Treatment *B* is added part way through recruitment, and no interim data are looked at prior to the amendment. The change to the randomisation does not directly affect hypothesis *A* as patients are still being randomised 1:1 between treatments *Z* and *A* with no amendments to the design for those treatment arms, although the rate of recruitment is likely to reduce due to the inclusion of the additional arm in the randomisation, and the amendment could cause a stage effect if there is an accompanying change to the population.

In survival analysis where proportional hazards are assumed, the hazard function, which is the risk of progression at time *t* for a patient *i*, can be written as

$$h_i(t) = \exp\{\beta_1 x_{1i} + \ldots + \beta_l x_{li}\}\, h_0(t),$$

where $h_i(t)$ is the hazard function for the i^{th} patient (i = 1,…,n), and $h_0(t)$ is the baseline hazard function; X is an indicator variable with value *x*, such that for explanatory variable l, $x_{li} = 0$ if it is the first level (such as the control treatment) and $x_{li} = 1$ if it is the second level (such as the experimental treatment); and $\exp(\beta_l)$ is the hazard ratio of $x_l = 1$ relative to $x_l = 0$, adjusted for the other covariates.

In the simple pooled analysis where there are no covariates, the above can be reduced to include only the treatment effect ($l = \theta$) as an explanatory variable

$$h_i(t) = \exp\{\beta_\theta x_{\theta i}\}\, h_0(t).$$

In the multivariable model adjusting for stage, the hazard function includes trial stage ($l = k$) as a covariate

$$h_i(t) = \exp\{\beta_\theta x_{\theta i} + \beta_k x_{ki}\}\, h_0(t).$$

The multivariable model tests the extent that each of the explanatory variables affect the hazard function, adjusted for the other covariates, with the hypotheses for each set to $H_0: \beta_l = 0$ vs $H_1: \beta_l \neq 0$. Cox proportional hazards regression model uses the likelihood ratio test which calculates the maximised partial log likelihood ($log\hat{L}$) with and without each explanatory variable in turn, and compares the differences in values

of the $-2Log\hat{L}$ between the two models for each explanatory variable to χ^2_{q-1}, as they have an approximately chi-squared distribution with $q-1$ degrees of freedom under H_0, where q is the number of levels for the variable. In this way, the model can output a p-value for the treatment effect and treatment effect estimate with confidence intervals after adjusting for the trial stage, and vice-versa.

5.2.3 P-value combination, adaptive analyses

The principle of p-value combination methods are that p-values are calculated separately for each stage from the independent cohorts of patients, and then a combination function is applied to allow testing across the trial as a whole. This controls for the adaptation of the design, according to the Conditional Invariance Principle (Section 5.1.3). An advantage of combination methods is that p-values can be calculated for each stage using any frequentist analysis method, regardless of the type of design or endpoint. A disadvantage is that outputting a treatment effect estimate and confidence interval is not straightforward.

Let H_0 be the null hypothesis to be tested over the whole trial. Define p_1 and p_2 to be the p-values from the first and second stage respectively, which are assumed to be independent and uniformly distributed on [0,1] under H_0 so that they satisfy the *p-clud* criterion (Brannath et al. 2002[90]). The outcome of the final analysis is based on a combination function of the p–values from each stage, $C(p_1, p_2)$, which is continuous and monotonically increasing. H_0 is rejected if $C(p_1, p_2) < c_\alpha$, where c_α is the appropriate critical value, depending on the combination method. The two most commonly discussed combination functions in the literature are Fisher's combined probability method[52, 91] and the weighted inverse normal combination function[92].

Note that when p-values are combined, it is necessary to use the results from one-sided tests to avoid the possibility of combining p-values for which the treatment effect is in the opposite direction. However, it is often appropriate to design trials with a two-sided null hypothesis to allow for the possibility that the experimental treatment is actually worse than control, which is an interesting and relevant finding. As noted by Whitehead[93], it is possible to combine the one-sided p-values for each tail and use those to determine the equivalent two-sided overall p-value by doubling the most significant of the one-sided p-values. This method is used when calculating the overall p-values throughout this chapter.

5.2.3.1 Fisher's combined probability method

Based on meta-analysis methodology by Fisher (1932)[91], the evidence from the two stages can be combined by the product of the independent p-values, $p_1 p_2$. Recall the *p-clud* criterion that under H_0 the p-values are independent and uniformly distributed, so $p_k \sim U(0,1)$ for stage *k* (*k* = 1,2). Fisher derived a proof that the natural logarithm of the product of *k* random variables with uniform distributions form a χ^2 distribution with 2*k* degrees of freedom, such that

$$-2\ln(p_1 p_2) \sim \chi_4^2(1-\alpha).$$

That is, $C(p_1, p_2) = p_1 p_2$ and $c_\alpha = e^{-\frac{1}{2}\chi_4^2(1-\alpha)}$. With a one-sided $\alpha = 0.025$, there is evidence to reject H_0 if $p_1 p_2 \leq 0.0038$.

5.2.3.2 Weighted inverse normal method

Fisher's method above includes the p-values from each stage with equal weighting, regardless of the amount of contributing information. With the inverse normal method proposed by Lehmacher and Wassmer (1999)[92], the weights can be varied, as long as they are chosen in advance of the trial, and are usually set to mirror the anticipated amount of information contributing from within the stages. This is optimal in terms of the combined p-value approximating the p-value based on a pooled analysis.

Let Z_k be the normal Z-value for stage *k*, such that $Z_k = \Phi^{-1}(1-p_k)$, $k = 1, \ldots, K$. The weighted inverse normal test rejects the null hypothesis if $\sum_{k=1}^{K} w_k Z_k > \Phi^{-1}(1-\alpha)$, where the weights w_k are set so that $\sum_{k=1}^{K} w_k^2 = 1$ and the α is one-sided. For *K*=2,

$$C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1-p_1) + w_2 \Phi^{-1}(1-p_2)],$$

where Φ is the CDF of the normal distribution function. There is evidence to reject H_0 if $C(p_1, p_2) \leq 0.025$.

The weights are usually set so that they are based on the size of the stages,

$$w_k = \sqrt{n_k / n_1 + n_2}, (k = 1,2)$$

where n_k reflects the amount of information from stage *k*. For continuous or binary endpoint data, n_k is determined by the sample size.

Note that whilst weighted methods are frequently used when combining p-values, Becker (1994)[94] commented that "Since the *p* value depends on the sample size for which it is calculated, for any given effect size or outcome a larger study will have a smaller *p* than a smaller study. Giving larger studies even more weight would be inappropriate and might adversely affect the power of the methods." Therefore, both combination approaches, with and without weighting, will be investigated.

5.2.3.3 Combining p-values from time-to-event outcome data

Time-to-event outcomes in trials assessing survival add complexity because events continue to occur over time and the p-value is therefore dependent on when the analysis is carried out. The necessary assumption of the p-values being independent for each stage is easily held by defining the stages according to when the patients were randomised, so for Hypothesis *A* patients recruited before the amendment contribute to stage 1, and those after contribute to stage 2, regardless of when any events occur. In addition, it is assumed that the final analysis will be carried out for a hypothesis when the pre-specified number of events have been observed from patients across the relevant stages, and the data from the relevant trial arms within all stages will be analysed concurrently. The time to analysis may be increased by the addition of the new arm if it affects the recruitment rate, but this alone would not affect the power and type I error rates as long as the analysis is triggered by the occurrence of the planned number of events. Since the p-values for each stage are independent and uniformly distributed under their respective null hypotheses so *p-clud* is satisfied, and are calculated at a pre-specified point based on number of events which is unaltered by the addition of the new arm, p-value combination methods are appropriate. Note that if there is an interim analysis on the time-to-event endpoint, the data will be immature compared to that at the final analysis. This needs to be taken into consideration when planning the design, but is out of scope here as our assumption is that the amendment is not based on interim data. Jenkins et al. (2011)[95] have discussed this scenario.

In weighted combination methods, the weighting should be set to be the anticipated amount of information gained from each stage. In time-to-event trials, patients recruited in the first stage are followed up for longer and have the opportunity to record more events, and therefore the weighting should reflect that. Jenkins et al. (2011)[95] state "Ideally, the weights w_1 and w_2 would be chosen to be proportional to the square roots of the numbers of [survival] events observed during each stage. The combination test

would then have the attractive property of yielding, approximately, the usual test statistic from a single combined analysis." In order not to introduce bias, the weights need to be pre-determined, and therefore can be estimated based on the expected number of events from patients in each stage assuming they follow an exponential distribution, the same assumption that is used to power the trial. In this case, the e_k in the following formula would be determined by the estimated number of events to be observed from stage *k* patients at the time of the planned analysis hence giving weights

$$w_k = \sqrt{e_k / e_1 + e_2}, (k = 1,2).$$

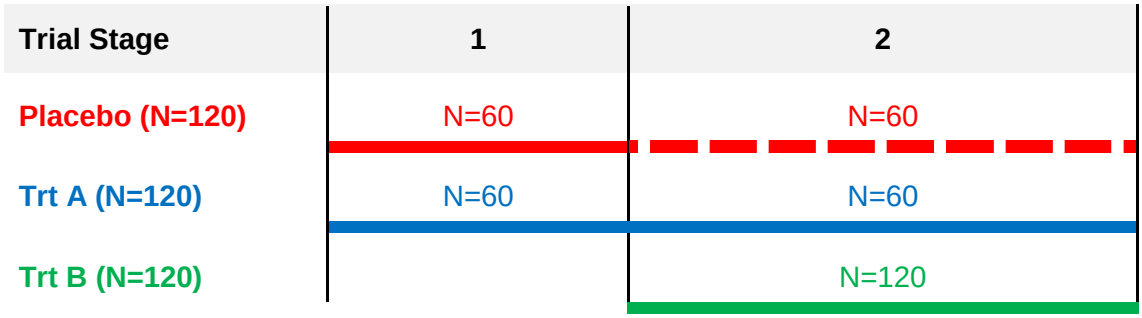
These weightings would be used in the final analysis, regardless of the actual number of events observed. Deviations in the number of events observed from each stage to those expected could affect the test statistic so it would move away from the pooled test statistic, which could have an impact on the observed error rates. However as long as the misspecification is not unreasonable, this is still likely to be closer to optimal than an unweighted combination method.

5.3 Review of existing research on analysis methods when adding a treatment arm based on external evidence

Elm et al. (2012)[20] are the only authors identified to have previously investigated analysis methods specifically when adapting a trial by adding a treatment arm. They share the assumption that the evidence for adding a treatment arm is external to the trial being amended and no interim analyses are planned. Their work was motivated by the possibility of adding an arm to a phase III trial in Parkinson's disease with a continuous, normally distributed primary endpoint. They assessed bias between the four analysis methods reviewed in Sections 5.2.2 and 5.2.3 in terms of type I error and power, by applying them to simulated scenarios in which an arm is added at varying times and with various degrees of stage effect. Pairwise comparisons were assumed between each experimental arm and the placebo arm. The findings were that when there is a stage effect, there is a familywise type I error bias and a loss of power for both comparisons when the data are simply pooled, but particularly for the new comparison. In the reported scenario, the linear model was the most powerful, but the adaptive approach was thought appropriate if interim data are involved.

As noted previously, whilst their work is highly relevant to this current research, there are a number of considerations that limit their conclusions, and may not be realistic to many trials in practice. To illustrate this, an example of their randomisation scheme is shown in Figure 5-2, where the new arm is added halfway through the original recruitment target. The numbers in each stage are determined by when the amendment occurs, which is varied in the simulations to 10%, 30% and 50% of the planned recruitment for the original trial.

Figure 5-2 Hypothetical Parkinson's disease trial randomisation by treatment arm and trial stage



The allocation ratio is adjusted so that all three arms complete recruitment at the same time to achieve the same patient numbers, and the comparisons for Treatment B are against the entire placebo arm recruited over stages 1 and 2. The trial was designed in this way for blinding purposes, since the control group are receiving placebos for both experimental treatments in the second stage, and to avoid having a third stage with different characteristics again. This violates our assumption that control patients should be recruited concurrently, and therefore will increase the bias in the case of a stage effect for the Treatment B comparison using the pooled methods[53]. They, in fact, point out that Parkinson's studies have shown that "changes in clinical practice over time have a major impact on outcome measures", therefore recommending that the amendment is made as early in the trial as possible. Whilst this strategy reduces the overall number of patients allocated to placebo, the potential bias in the event of a stage effect seems to negate this benefit and could affect the interpretability of the outcomes. The adaptive methods will naturally overcome the bias caused by having non-concurrent controls because the p-values are calculated separately from patients within each stage. However, since "H_B is tested using just the stage 2 data via $p_{2,B}$ (since there is no data for treatment B in stage 1)", there will be reduced power for the Treatment B comparison with adaptive methods. Where the new treatment is added 50% through recruitment, the combined p-value will be based on only 60 control and 120 experimental patients, rather than the 240 patients needed in total.

In addition, Elm et al. make the assumption that the familywise error rate needs to be controlled due to having multiple experimental arms compared to the same control group, and used closed testing methods in combination with adaptive procedures in order to adjust within stage. The requirement for and use of multiple testing adjustment procedures are addressed in Section 5.5, where this and alternative methods are discussed, and it is suggested that this method of adjusting within stage may not be appropriate and that adjustment should instead be across the trial as a whole. This is firstly because the FWER is not inflated only within specific stages, and secondly because Stagewise methods give unexpected results in the case where arms are added (Section 5.5.2). Since all of their simulations incorporate this adjustment method, the results are muddied between the effect of the analysis method and that of the multiplicity adjustment, so that they cannot be extrapolated to cases with different multiple testing assumptions. If multiple testing adjustment is not deemed necessary, or if adjustment is made across the whole trial rather than within stage, it is not clear from their work which analysis method is appropriate. The use of closed testing methods requires that both hypotheses are analysed at the same time, and the type I error rates concern the hypotheses jointly. For this reason only the familywise error rate is presented, but not the type I error rates for the two separate comparisons, which may be the appropriate error to consider if the hypotheses are considered individually and without requiring adjustment. Using closed testing methods, the effect on the error rates of the non-concurrent controls for Hypothesis *B* also affects the overall recommendations for Hypothesis *A*, inflating the FWER and reducing the power for both hypotheses using the pooled analysis method if there is a stage effect. Finally, their results are restricted to continuous endpoint data, and we consider time-to-event data but confirm that the findings are similar for other types of endpoints.

5.4 Simulation study

5.4.1 Aims and assumptions

The aim of the simulation study is to compare the type I error and power for testing Hypothesis *A* based on the four different analysis methods and under various assumptions following the addition of Treatment *B*. It is assumed that the amendment is made without reference to internal trial data. The simulations assess the impact of stage effects caused by the addition of Treatment *B* on the error rates, whilst varying the time of the amendment and the allocation ratio. The null hypothesis $H_{0A}: \theta_A = 0$ is tested against the two-sided alternative hypothesis $H_{1A}: \theta_A \neq 0$, where θ_A is the treatment effect for treatment *A* compared to *Z*. The focus of the simulations are on the

analysis of Hypothesis *A* because that is the hypothesis affected by the potential stage effect caused by the addition of Treatment *B* (Section 5.1.3). Hypothesis *B* is not affected by the addition of any other new arms, but since the design includes a third stage in which Treatment *A* is dropped because it has completed recruitment, there could be another stage effect affecting Hypothesis *B* that is similarly not caused by an interim analysis. The results addressing the impact of a stage effect on the analysis of Hypothesis *A* will also be valid for Hypothesis *B*, and this is considered in the recommendations.

The simulations are based on the FLAIR trial design described in Section 5.1.2. Each scenario is assessed by simulating 100,000 trials. It is assumed that the survival patterns follow an exponential distribution with scale parameter $\lambda_j = \frac{log2}{med_j}$, where med_j is the median survival time in each arm j ($j = Z, A$) as described in Section 2.1.3.3. In order to assess the type I error the median survival times are set to be the same in each arm, so they both equal the estimated survival for Control Treatment *Z*. To assess the power they are set to equal the survival estimates used for the sample size calculation, so they vary by the clinically relevant difference. It is assumed that the 5% participant dropout is spread evenly, so that the uninflated sample size is relevant (354 participants per arm). Survival times are censored at the median follow-up for each stage k ($k = 1,2$). Whilst this differs to practice, where survival times are instead censored at the data last seen prior to the analysis, the results will be equivalent because of the memoryless property of the exponential distribution. The censored survival times are analysed using the Cox proportional hazards model (Section 5.2.2), comparing experimental treatment *A* to control treatment *Z*, and outputting a p-value for the treatment effect for each of the simulated trials. The error rates are obtained by counting the number of trials in which H_{0A} is rejected.

For each scenario, the data are analysed based on the four methods under investigation: pooled without adjustment (POOLED); pooled and adjusting for stage using a multivariable model (MULTI); adaptive method analysing within stage and then combining p-values using Fisher's combination method (FISHER); adaptive method analysing within stage and then combining p-values using the weighted inverse normal combination method (INVN) with weights set by the number of events anticipated from patients in each stage.

Simulations assess the error rates under different assumptions of stage effect in Stage 2, whilst also varying the time to the addition of the new arm. The stage effect is assumed to alter the median survival by the same relative amount in both arms, as shown in Table 5-2. The Stage 1 median survival estimates are 4.5 and 6 years for treatments *Z* and *A* respectively, so a 0.89 stage effect in Stage 2 decreases the Stage 2 estimates to 4 and 5.33 years, and a 1.11 stage effect increases the Stage 2 estimates to 5 and 6.67 years. Various Stage 2 estimates are included in the simulations to assess the impact on the error rates. It is assumed that the rate of recruitment remains constant over the trial, and that Hypothesis *A* is analysed when there are enough events in the concurrently randomised arms. The analysis timepoint is event driven, but the simulations are based on a fixed length of follow-up, so it is necessary to calculate the follow-up for each scenario to ensure that 379 events would be observed across the two arms if the median survivals are equal to those set at the design stage. The length of follow-up differs in each scenario because the additional arm affects the rate of recruitment into the existing arms, therefore affecting the time to reach the primary outcome, and in addition when there is a stage effect it changes the rate of occurrence of events. For example, with a 0.89 stage effect in the second stage, the overall median follow-up needed to reach the required number of events is reduced from 6 years to 5.743 years, calculated by solving the CDF of the hyperexponential distribution to find the combined median of the exponential distributions for each stage. In the case where the new arm is added halfway through recruitment with 1:1 allocation, the 4 year recruitment period to assess hypothesis A is increased to 5 years. The follow-up period required after recruitment for Hypothesis A (*fupA*) is therefore reduced to 3 years, because $0.5 * \left({}^{2}/_{2} + 3 + fupA\right) + 0.5 * \left({}^{3}/_{2} + fupA\right) = 5.743$, so the median follow-up for stage 1 patients is 83.9 months, and for stage 2 patients it is 53.9 months.

Whilst the follow-up time is varied for each simulated scenario involving a stage effect to maintain the appropriate power, the weighting based on number of events for the weighted inverse normal combination method is not varied. This reflects practice, because the weightings are determined at the time of the amendment based on the initial protocol assumptions and do not change if the number of events observed are not as expected. In the example of a 1:1 allocation with the arm being added 50% of the way through recruitment, ignoring the stage effect, the number of expected events in each stage will be 216 from stage 1 patients compared to 163 from stage 2 patients if they follow the exponential model.

5.4.2 Assessing error rates with varying stage effects for different amendment time points

Simulations are used to compare the type I error and power for the four analysis methods based on different degrees of stage effect, see Table 5-2 below. The time to the amendment is varied to 25%, 50% and 75% of the way through recruitment to the original trial, with a 1:1 allocation ratio. It is assumed that any stage effect affects both arms in the same way in the second stage, so that the hazard ratio remains constant across the stages. It is also assumed that no multiplicity adjustment is necessary, although this is addressed in Section 5.5. Therefore the median survival for Treatment *B* is not important, only its effect on the median survival for Treatments *A* and *Z*.

Table 5-2 Median survival and follow-up times to assess a Stage 2 stage effect, assuming varying times to amendment

Scenario	Stage Effect (HR = 0.75)	Stage 2 Median Survival Trt Z (and Trt A for type I error) (months)	Stage 2 Median Survival Trt A (for power) (months)	Overall Median follow-up (months)	Median follow-up Stage 1 (months)	Median follow-up Stage 2 (months)
Time to amendment = 25%. No. events from patients in: Stage 1 = 117; Stage 2 = 265.						
No difference from Stage 1	0.000	54	72	72.0	96.8	63.8
Reasonable decrease	0.889	48	64	67.4	92.2	59.2
Reasonable increase	1.111	60	80	77.9	105.8	72.8
Large decrease	0.444	24	32	40.2	65.0	32.0
Large increase	1.555	84	112	99.2	124.0	91.0
Time to amendment = 50%. No. events from patients in: Stage 1 = 216; Stage 2 = 163.						
No difference from Stage 1	0.000	54	72	72.0	87.0	57.0
Reasonable decrease	0.889	48	64	68.9	83.9	53.9
Reasonable increase	1.111	60	80	75.9	90.9	60.9
Large decrease	0.444	24	32	48.4	63.4	33.4
Large increase	1.555	84	112	88.8	103.8	73.8

Time to amendment = 75%. No. events from patients in: Stage 1 = 306; Stage 2 = 76.						
No difference from Stage 1	0.000	54	72	72.0	78.8	51.8
Reasonable decrease	0.889	48	64	70.4	77.5	50.5
Reasonable increase	1.111	60	80	73.9	80.7	53.7
Large decrease	0.444	24	32	59.0	65.8	38.8
Large increase	1.555	84	112	79.7	86.5	59.5

The simulations programs were written in SAS v9.4 based on the above assumptions, and each scenario was run 100,000 times. Table 5-3 provides the percentages of trials in which the null hypothesis was rejected (two-sided $p \leq 0.05$) where the medians for both arms were set to equal the same value, showing the type I error rates following each analysis method.

Table 5-4 provides the percentages of trials in which the null hypothesis was rejected where the medians for the arms were set to be different as described in Table 5-2, showing the power following each analysis method.

Table 5-3 Two-sided type I error results for Hypothesis *A* from 100,000 simulations per scenario

Scenario	Stage Effect	POOLED	MULTI	FISHER	Weighted INVN
Time to amendment = 25%					
No difference from Stage 1	None	5.02	5.02	4.98	4.99
Reasonable decrease	0.889	5.02	5.05	5.00	5.02
Reasonable increase	1.111	4.89	4.90	4.85	4.88
Large decrease	0.444	4.39	4.98	4.88	4.95
Large increase	1.555	4.68	4.94	4.97	4.90

Time to amendment = 50%					
No difference from Stage 1	None	5.01	5.03	4.95	4.99
Reasonable decrease	0.889	4.92	4.94	4.82	4.89
Reasonable increase	1.111	4.99	5.02	4.96	4.95
Large decrease	0.444	4.41	5.13	5.03	5.10
Large increase	1.555	4.78	5.05	5.04	5.00
Time to amendment = 75%					
No difference from Stage 1	None	4.92	4.92	4.85	4.90
Reasonable decrease	0.889	5.06	5.08	4.91	5.03
Reasonable increase	1.111	4.94	4.98	4.84	4.93
Large decrease	0.444	4.35	4.89	4.84	4.89
Large increase	1.555	4.76	4.95	4.78	4.90

The 95% confidence interval on a type I error of 5% for 100,000 simulations is 4.87% to 5.14%. In one case the error is 5.13% where the time to amendment is 50%, but in the equivalent scenarios with time to amendment of 25% and 75% it is less than 5%, so this is most likely a chance result rather than a true inflation. Therefore the two-sided type I error does not exceed the required 5%, outside the expected margin of error due to the accuracy of the simulations, in any of the scenarios. In the pooled analysis without adjustment, the type I error drops to around 4.3-4.4% where there is a large decrease in survival in the second stage, and to around 4.7-4.8% where there is a large increase. This is regardless of when the amendment is made.

Table 5-4 Power results for Hypothesis *A* from 100,000 simulations per scenario

Scenario	Stage Effect	POOLED	MULTI	FISHER	Weighted INVN
Time to amendment = 25%					
No difference from Stage 1	None	80.85	80.88	77.87	80.81
Reasonable decrease	0.889	81.38	81.44	78.38	81.35
Reasonable increase	1.111	81.97	82.03	79.17	81.93
Large decrease	0.444	79.05	80.55	76.95	80.34
Large increase	1.555	80.25	81.04	78.13	80.86
Time to amendment = 50%					
No difference from Stage 1	None	80.70	80.72	78.20	80.63
Reasonable decrease	0.889	80.58	80.63	78.06	80.57
Reasonable increase	1.111	80.67	80.73	78.20	80.68
Large decrease	0.444	77.76	79.83	77.33	79.38
Large increase	1.555	79.84	80.54	77.95	80.41
Time to amendment = 75%					
No difference from Stage 1	None	80.93	80.95	77.11	80.86
Reasonable decrease	0.889	81.12	81.19	77.57	81.08
Reasonable increase	1.111	80.97	81.05	77.16	80.92
Large decrease	0.444	78.85	80.78	77.77	80.20
Large increase	1.555	80.66	81.08	76.77	80.90

Take for example the scenario in which an arm is added halfway through recruitment and the population shifts so there is an unrealistically large decrease in the median survival for both treatment arms in the second stage. In Stage 1 the median survivals are 54 months for control and 72 months for experimental treatment A, assuming a clinically relevant difference, but in Stage 2 the median survivals decrease to 24 months and 32 months respectively. The hazard ratio of 0.75 remains unchanged for the treatment effect across the stages. Using the MULTI and weighted INVN analysis methods, the power to detect the treatment effect remains above 79%, but with the

POOLED analysis it decreases to 77.8%, and to 77.3% with the FISHER combination method.

The 95% confidence interval on a power of 80% for 100,000 simulations is 77.52% to 82.48%. Varying the time to amendment has no noticeable impact on the power in any of the scenarios. In all cases, the power using Fisher's combination method is a couple of percent lower than with the other methods, and outside of the error margin for the simulations. Where there are large stage effects, the power with the pooled analysis method with no adjustment is up to 2% lower than with the multivariable analysis adjusting for trial stage, although where the stage effect is set to be more realistic, the pooled analysis appears to perform as well as the multivariable analysis in terms of power.

These simulations suggest that none of the four analysis methods would be unsuitable, since none inflate the type I error and the difference in power is relatively small. However, a small difference in power can still lead to a large impact on the sample size and therefore it is advantageous to use the methods that lead to the highest power in all cases. There is no advantage in terms of error rates of using combination methodology over the pooled methods if there is a stage effect. If it is felt appropriate to pool the data over the stages without using combination methodology, the multivariable method adjusting for trial stage is more powerful than the unadjusted method where there is a stage effect, so this method would be recommended. If combination methodology is determined to be necessary, the weighted inverse normal method is more powerful in all cases than Fisher's combination method.

Simulations shown in Table 5-5 were also conducted investigating the effect of the allocation ratio on the error rates. The simulations included 1:2 and 2:1 allocation ratios for the original hypothesis, and assumed that the arm was added at 50% of the way through recruitment to the original trial, based on the same design as previously and with the same allocation ratio to Treatment B as to Treatment A. With no stage effect, the error rates for each of the four analysis methods have the same patterns as seen above, where the type I error is around 5% in all cases and the power around 80%, with Fisher's combination method leading to a slightly lower power than for the other methods. Based on the above results, there is no reason to believe that varying the stage effect alongside the allocation ratio would have any different effect on the patterns of error results for the different analysis methods.

Table 5-5 Two sided Type I Error and Power results for Hypothesis *A* when varying the allocation ratio, based on 100,000 simulations per scenario

Allocation Ratio	POOLED	MULTI	FISHER	Weighted INVN
Time to amendment = 50%, no stage effect, two-sided Type I Error results				
2:1	4.94	4.96	4.94	4.90
1:2	5.06	5.07	4.99	5.04
Time to amendment = 50%, no stage effect, Power results				
2:1	80.28	80.30	77.71	79.82
1:2	79.66	79.67	77.46	79.58

5.4.3 Validation of the results using scenarios with different outcome measures

In order to validate the findings and recommendations on the appropriate analysis method when the addition of an arm causes a stage effect, the simulations were reproduced using different scenarios. The first assumes a three arm trial with a continuous, normally distributed endpoint, and the second a binary endpoint. The aim is to investigate whether the findings are robust using different examples, and to investigate any differences when the analysis methods to produce the p-values differ.

5.4.3.1 Normally distributed endpoint data

The scenario is similar to that in Section 4.3.4.2, and assumes a three arm trial with a continuous, normally distributed endpoint, analysed using two-sample t-tests. With a two-sided significance level of 0.05 and 80% power to assess an effect size of 0.5, assuming a control mean of 100, assessing a difference of 5 with a common standard deviation of 10, 64 patients are required per arm with 1:1:1 randomisation. The common standard deviation is varied in Stage 2 along with the means so that the effect size in each stage remains at 0.5.

Table 5-6 Means and standard deviations to assess a Stage 2 stage effect

Scenario	Stage Effect (effect size = 0.5)	Stage 2 mean Trt Z (and Trt A for type I error)	Stage 2 mean Trt A (for power)	Common SD to maintain effect size
No difference from Stage 1	None	100	105	10
Reasonable decrease	0.889	88.9	93.3	8.8
Reasonable increase	1.111	111.1	116.7	11.2
Large decrease	0.444	44.4	46.6	4.4
Large increase	1.555	155.5	163.3	15.6

The simulations were run using SAS v9.4, and each scenario was run 100,000 times. Table 5-7 provides the percentage of trials in which the null hypothesis was rejected (two-sided $p \leq 0.05$) where the means for both arms were set to equal the same value, showing the type I error rates following each analysis method. Only the case where the arm is added halfway through recruitment is shown. Table 5-8 provides the percentages of trials in which the null hypothesis was rejected where the means for the arms were set to be different as described in Table 5-6, showing the power following each analysis method.

Table 5-7 Type I error results for Hypothesis *A* with normally distributed outcomes from 100,000 simulations per scenario

Scenario	Stage Effect	POOLED	MULTI	FISHER	Weighted INVN
Time to amendment = 50%					
No difference from Stage 1	None	5.07	5.07	5.08	5.08
Reasonable decrease	0.889	2.29	5.02	4.99	5.03
Reasonable increase	1.111	2.73	5.13	5.13	5.08
Large decrease	0.444	0	5.00	5.04	5.04
Large increase	1.555	0	4.91	4.93	4.91

The type I error remains around 5%, within the expected error margin, in all cases apart from following the POOLED analysis method. This is discussed below.

Table 5-8 Power results for Hypothesis *A* with normally distributed outcomes from 100,000 simulations per scenario

Scenario	Stage Effect	POOLED	MULTI	FISHER	Weighted INVN
Time to amendment = 25%					
No difference from Stage 1	None	80.13	80.14	76.48	79.94
Reasonable decrease	0.889	71.64	80.11	76.51	80.09
Reasonable increase	1.111	74.25	80.11	76.50	80.08
Large decrease	0.444	0.00	73.93	76.54	80.04
Large increase	1.555	12.86	79.05	76.68	80.15
Time to amendment = 50%					
No difference from Stage 1	None	80.03	80.05	77.37	79.85
Reasonable decrease	0.889	69.30	79.95	77.49	79.93
Reasonable increase	1.111	71.82	80.16	77.48	80.14
Large decrease	0.444	0.00	74.38	77.26	79.81
Large increase	1.555	2.34	78.29	77.29	79.92
Time to amendment = 75%					
No difference from Stage 1	None	80.30	80.29	76.53	80.16
Reasonable decrease	0.889	72.76	80.16	76.48	80.13
Reasonable increase	1.111	73.38	80.04	76.42	79.90
Large decrease	0.444	0.09	76.85	76.44	79.98
Large increase	1.555	3.19	78.58	76.54	80.12

The power is generally slightly lower with Fisher's combination analysis method than the multivariable and weighted inverse normal combination methods, agreeing with the findings in the survival case. In addition, the time to amendment does not affect the findings. In contrast to the survival case discussed previously, the POOLED results show that if there is any stage effect, pooled methods without adjustment are not appropriate. This is because the standard deviation after pooling the data across stages becomes very large, therefore reducing the effect size. In this case, even with only a reasonable stage effect, the power is reduced by a large amount with the

POOLED method. Therefore stage should always be accounted for in the analysis, either within a multivariable analysis, or using combination methods. Note that these results were obtained using a t-test and validated using a regression, and were identical. The power is slightly lower using the multivariable analysis in the case of a large stage effect, although the power is good if the stage effect is not unrealistic.

5.4.3.2 Binary endpoint data

It is also of interest to investigate the analysis methods in the example of a binary outcome measure. The scenario here assumes a three arm trial with a binary endpoint, analysed using logistic regression. With a two-sided significance level of 0.05 and 80% power to assess an odds ratio of 1.667, assuming a control proportion of 0.5 (experimental proportion of 0.625). 246 patients are required per arm with 1:1:1 randomisation.

Table 5-9 Binary proportions to assess a Stage 2 stage effect

Scenario	Stage Effect (OR fixed at 1.667)	Stage 2 proportion Trt Z (and Trt A for type I error)	Stage 2 proportion Trt A (for power)
No difference from Stage 1	None	0.5	0.625
Reasonable decrease	0.889	0.444	0.571
Reasonable increase	1.111	0.556	0.676
Large decrease	0.444	0.222	0.323
Large increase	1.555	0.778	0.854

The simulations were run using SAS v9.4, and each scenario was run 100,000 times. Table 5-10 provides the percentage of trials in which the null hypothesis was rejected (two-sided $p \leq 0.05$) where the proportions for both arms were set to equal the same value, showing the type I error rates following each analysis method. Table 5-11 provides the percentages of trials in which the null hypothesis was rejected where the proportions for the arms were set to be different as described in Table 5-9, giving the power following each analysis method. Only the case where the arm is added halfway through recruitment is shown as the results based on different times to amendment showed the same pattern.

Table 5-10 Type I error results for Hypothesis *A* with binary outcome measures, from 100,000 simulations per scenario

Scenario	Stage Effect	POOLED	MULTI	FISHER	Weighted INVN
Time to amendment = 50%					
No difference from Stage 1	None	5.28	5.28	4.99	5.28
Reasonable decrease	0.889	5.27	5.27	4.97	5.27
Reasonable increase	1.111	5.18	5.18	4.89	5.18
Large decrease	0.444	4.05	5.04	4.85	4.97
Large increase	1.555	4.04	5.00	4.87	4.91

The type I error is slightly inflated above 5%, even in the case with no stage effect. This is likely to be due to the discrete nature of the binomial distribution, causing a slight zig zag effect in the error rates.

Table 5-11 Power results for Hypothesis *A* with binary outcome measures, from 100,000 simulations per scenario

Scenario	Stage Effect	POOLED	MULTI	FISHER	Weighted INVN
Time to amendment = 50%					
No difference from Stage 1	None	80.16	80.16	77.75	80.16
Reasonable decrease	0.889	80.75	80.75	77.90	80.75
Reasonable increase	1.111	78.78	78.85	76.86	78.90
Large decrease	0.444	72.92	76.20	73.38	75.94
Large increase	1.555	68.05	70.82	67.68	70.07

In the binary case, the power is reduced with all analysis methods if there is a stage effect. This is expected because even though the odds ratio is unchanged, the sample size needs to be larger the further the estimated proportions are from 0.5. The general message is the same as for the survival case in that the power is always lower using the FISHER analysis method compared to the weighted INVN method, and that the POOLED method performs worse than the MULTI method if there is a large stage effect.

5.4.4 Assessing the effect of a treatment*stage interaction

A treatment*stage interaction is possible if the population shifts due to the amendment but the outcomes within the treatment arms are affected differently. If this is the case, it may not be appropriate to combine data over the stages at all. This is a similar issue to that in factorial designs, where if the treatments interact with one another it could have a severe impact on the ability to analyse the main effects, or cross-over trials when considering a carry-over effect.

Simulations assessed the impact of an extreme case in which there is no difference between treatment arms in the first stage but the difference becomes clinically relevant in the second stage. The amendment was based on the FLAIR design as above, was assumed to be made halfway through recruitment, and the randomisation allocation was even between the arms. Firstly, the power to detect an interaction was assessed using a multivariable model, described in Section 5.2.2, as follows

$$h_i(t) = \exp\{\beta_\theta x_{\theta i} + \beta_k x_{ki} + \beta_{\theta k} x_{\theta i} x_{ki}\}\, h_0(t),$$

where $h_i(t)$ is the hazard function for the i^{th} patient (i = 1,…,n), and $h_0(t)$ is the baseline hazard function; X is an indicator variable with value x, such that for explanatory variable l, $x_{li} = 0$ if it is the first level (such as the control treatment) and $x_{li} = 1$ if it is the second level (such as the experimental treatment); and $\exp(\beta_l)$ is the hazard ratio of $x_l = 1$ relative to $x_l = 0$, adjusted for the other covariates. $l = \theta$ represents the treatment effect, $l = k$ represents the trial stage, and $l = \theta k$ represents the treatment*stage interaction.

In this example, based on 100,000 simulations, the power to detect the treatment*stage interaction is less than 29%. Therefore a non-significant result cannot be assumed to imply that no interaction exists. The strategy implemented in the confirmatory 2NN trial[16] (Section 2.2.3) when an arm was added was to use a multivariable logistic regression model to assess stage and treatment*stage interaction, and then justify a pooled analysis based on the lack of significance. With the low power to detect treatment*stage interactions, this strategy cannot be recommended.

The role of homogeneity testing in adaptive trials is discussed by Gallo and Chuang-Stein (2009)[96], who feel that formal statistical methods to assess a treatment*stage interaction should not be performed, stating “For example, what significance level should be used? If we use the familiar 5%, this might be construed as too stringent,

since the study was presumably not designed to have high power for detecting meaningful levels of interaction. On the other hand, if a higher significance level is used, one would run a high risk of obtaining a false-positive signal of interaction. This could incorrectly invalidate the trial results, falsely leading to the trial being declared inconclusive". Instead, it should be considered whether the adaptation to the trial is clinically likely to introduce a bias, and how this might affect the patient characteristics or outcome measures. This can then be assessed as part of the final analysis.

Simulations next assessed the impact of the treatment*stage interaction on the probabilities of rejecting the null hypothesis for the treatment effect using the four analysis methods. Note that the multivariable model only includes the explanatory variables treatment and stage, and not the interaction term.

Table 5-12 Treatment effect rejection rates based on 100,000 simulations, where a treatment*stage interaction exists

Scenario	Stage 1 median survival estimates (HR=1)	Stage 2 median survival estimates (HR=0.75)	POOLED	MULTI	FISHER	Weighted INVN
No difference in Stage 1; CRD in Stage 2	Trt A: 4.5y Trt Z: 4.5y	Trt A: 4.5y Trt Z: 6.0y	20.72	20.28	31.10	23.73

It is not obvious how a 20-31% chance of rejection of H_0 should be interpreted since this is not a meaningful statistic. It illustrates that if there is a large treatment*stage interaction, the analysis should not be performed over the trial as a whole using any of the methods. Therefore, before an arm is added, it should be considered whether a treatment*stage interaction is likely. If it is felt clinically that adding an arm could alter the population so that the treatment effect in one group is affected differently from the other group, the trial could be compromised by the amendment, and adding an arm is not recommended.

5.5 Multiple testing adjustment for multiple hypotheses

5.5.1 Theory behind the appropriate FWER adjustment

When an experimental arm is added to an ongoing trial in order to compare it to the current control, the trial will by definition include multiple comparisons. We previously

discussed (Chapter 4) that the FWER in this situation is not inflated over that in independent trials due to comparisons with shared control data, in fact it is reduced, but FWER adjustment might be required because the efficiency of sharing a protocol may allow more hypotheses to be tested, leading to more opportunities for success. For example, if experimental Treatment *B* is added to a trial just as Treatment *A* completes recruitment there is no shared control data but many advantages of sharing the same trial are still present such as: the protocol, approvals and database only require amendments, which reduces the set-up time and cost; the trials team exists and operating procedures are in place; and the centres are already set-up for recruitment. In this case, arguments for multiplicity adjustment around multiple opportunities for success within the same protocol, as discussed in Section 4.6 will hold. If Treatment *B* is added prior to Treatment *A* completing recruitment, the only additional efficiency is that fewer patients are required to be randomised to control (Treatment *Z*) overall, so fewer patients are required to address both hypotheses than for independent trials. Therefore, since the FWER is not inflated specifically because of the use of shared control data, it is proposed that if multiple testing adjustment is deemed necessary, adjustment should not only apply in the stages where the experimental therapies overlap, but should be considered over the protocol as a whole.

In contrast, the majority of literature regarding multiple testing adjustment for multiple hypotheses in adaptive designs includes adjustment within the stages where the experimental treatments overlap. For example, when discussing FWER when adding arms, Parmar et al.[22] sum up the common belief that “the implications for the FWER are dependent on the proportion of overlap in control arm patients between the research arms. If there was overlap of only one patient in the control groups for two different research arms, this would be (almost) like doing two independent trials with one common patient. In this instance, there would be no practical change to the type 1 error for these two comparisons”. They believe that FWER control is not needed when they have added arms with little overlap because the correlation between the test statistics is low “emphasising the lack of overlap and their relative independence”. What they have not acknowledged here is that any overlap has the effect of reducing, rather than increasing, the overall FWER for the trial, and therefore this is not the appropriate driver for adjustment.

This section reviews and compares multiplicity adjustment methods in adaptive designs in order to make recommendations when adding an arm and for multi-arm multi-stage designs more generally.

5.5.2 Application of the closed test procedure to adaptive designs

The most common method of adjustment for multiple hypotheses in adaptive designs literature is based on the closed testing principle, for example as discussed in key publications on adaptive designs such as Bauer and Kieser (1999)[97], Bretz et al. (2006 and 2009)[98, 99], Friede et al. (2011)[100], Keiser, Bauer and Lechmacher (1999)[101], Maurer, Branson and Posch (2010)[89] and Posch et al. (2005)[18]. Closed testing methods are based on the hierarchical testing theory that if null hypotheses are rejected in sequential order there is no inflation to the overall type I error, and were introduced in Section 3.3.2.2. It was shown that for each primary comparison, the methods first reject the null hypotheses for the intersections containing combinations of that comparison before assessing the main effect (see Figure 3-1).

Recall that there are different ways to calculate the intersection hypotheses, with the simplest based on the Bonferroni test, in which the overall p-value for testing the intersection hypothesis $H_{0(AB)} = H_{0A} \cap H_{0B}$ is calculated by $p_{AB} = \min(1, mp_A, mp_B)$, where *m* refers to the number of hypotheses being tested. Other tests are available that are marginally less conservative, such as those by Simes or Sidak, but as long as the chosen method to calculate the intersections is consistent, the overall conclusions are not affected.

In adaptive designs generally, the stagewise intersection hypotheses are first calculated and then combined in order to reject the overall intersection null hypothesis. That is for each stage *k*, $p_{k,AB} = \min(1, mp_{k,A}, mp_{k,B})$, where *m* is the number of hypotheses in that stage. If a stage includes only one hypothesis, say H_{0A} because the new hypothesis has not yet been added, $p_{k,AB} = p_{k,A}$. If the combined stagewise intersection p-values suggest rejection of the null hypothesis at level α, the combined individual p-values can then be tested at level α, otherwise the individual null hypotheses cannot be rejected. This is illustrated in Table 5-13, using a design similar to the FLAIR trial, and is labelled the 'Stagewise' intersection.

Another way to control for multiplicity in an adaptive design could be to apply standard adjustment methods to the p-values that have already been combined over the stages, similarly to a non-adaptive trial. If the p-values are first combined for each hypothesis, a Bonferroni or closed testing adjustment could be applied to the final single-stage p-

values. The Bonferroni adjustment simply compares the p-values for each hypothesis, p_A and p_B, to α/m. The closed testing intersection is illustrated in Table 5-13, and is labelled the 'Overall' intersection. Note that if a 'pooled' analysis method is carried out, as described in Section 5.2.2, the Overall method is the only type of adjustment possible as intra-stage p-values are not calculated.

Table 5-13 Illustration of the Stagewise and Overall methods to calculate the intersection hypothesis to apply closed testing for multiplicity adjustment within a p-value combination analysis

	Stage 1	Stage 2	Stage 3	P-value Combination (calculated)
H_{0A*}	$p_{1,A}$	$p_{2,A}$	-	$p_A=C(p_{1,A}, p_{2,A})$
H_{0B*}	-	$p_{2,B}$	$p_{3,B}$	$p_B=C(p_{2,B}, p_{3,B})$
Intersection $H_{0(AB)}$ (calculated)	$p_{1,AB} = p_{1,A}$	$p_{2,AB} = min(1, 2p_{2,A}, 2p_{2,B})$	$p_{3,AB} = p_{3,B}$	**1. Overall** $p_{AB} = min(1, 2p_A, 2p_B)$ **2. Stagewise** $p_{AB} = C(p_{1,A}, p_{2,AB}, p_{3,B})$

H_{0j} is rejected at level α if $p_{AB} \leq \alpha$ and $p_j \leq \alpha$, j=A,B.

In the 'Overall' method of adjustment (1.), p_{AB} is calculated by

$$p_{AB} = C(p_{1,A}, p_{2,A}) \cap C(p_{2,B}, p_{3,B}) = p_A \cap p_B = min(1, 2p_A, 2p_B).$$

Note that if a treatment is not present in a stage, the weighting for that stage will equal 0, so $p_{3,A}$ and $p_{1,B}$ are not present in the combination functions. If Fisher's combined probability method is used, the missing p-values are conservatively set to 1 so they don't affect the product term.

In the 'Stagewise' method of adjustment (2.), using the weighted inverse normal combination function,

$$p_{AB} = C(p_{1,AB}, p_{2,AB}, p_{3,AB}) = C(p_{1,A}, p_{2,AB}, p_{3,B}) =$$

$$\sqrt{e_1/e_1 + e_2 + e_3} * \phi^{-1}(1 - p_{1,A}) + \sqrt{e_2/e_1 + e_2 + e_3} * \phi^{-1}(1 - p_{2,AB}) +$$

$$\sqrt{e_3/e_1 + e_2 + e_3} * \phi^{-1}(1 - p_{3,B}),$$

where e_k is the number of events expected from patients who were randomised within stage k ($k=1,2,3$), and $p_{2,AB} = min(1, 2p_{2,A}, 2p_{2,B})$.

Posch et al.[18] give an example of applying the Stagewise method, and confirm that in the case where there are two stages and both treatments are present in each stage

"using the Bonferroni test for the stagewise tests for H_{12} the corresponding combination tests is given by [$C(\min(1,2p_{1,1}, 2p_{1,2}), \min(1,2p_{2,1}, 2p_{2,2}))$] and obviously differs from the overall Bonferroni test applied to the pooled data from both stages, even in the normal case when using the inverse normal combination function." However, the magnitude of these differences on the error rates are not discussed. The effects of the Stagewise and Overall methods for calculating the intersection hypothesis or otherwise adjusting for multiplicity are investigated here using simulations.

5.5.3 Comparison of the adjustment methods

Throughout this chapter so far both the Fisher and the weighted INVN methods for p-value combination have been compared. As we have shown that the weighted INVN method is more powerful, and it is generally accepted as better than Fisher's, this section will only include the weighted INVN combination technique. For comparison purposes, the multivariable (MV) method adjusting for stage will also be included as the optimal non-adaptive analysis method.

Simulations for both the pooled multivariable analysis adjusting for stage and the adaptive weighted INVN analysis are used to compare the error rates in the scenarios of:

- No multiplicity adjustment.
- Overall Bonferroni adjustment applied to the final p-values.
- Closed testing adjustment based on the 'Overall' method to calculate the intersection p-value ($p_{AB} = min(1, 2p_A, 2p_B)$).
- Closed testing adjustment based on the 'Stagewise' method to calculate the intersection p-value ($p_{AB} = C(p_{1,A}, p_{2,AB}, p_{3,B})$), possible for the adaptive method only.

The simulations are again based on the FLAIR trial design with three stages, as illustrated in Section 5.1.2, and each scenario is assessed by simulating 100,000 trials. The assumptions are that the new arm is added halfway through recruitment at a 1:1:1 ratio, recruitment remains at a constant pace, there is no stage effect, there is no heterogeneity of treatment effect over the stages (treatment*stage interaction), and Treatment *A* is dropped once its recruitment target has been met (this is not based on looking at interim data). It is necessary to assume that each hypothesis is analysed when the required number of events are reached, in order not to inflate the power. Therefore H_{0A} will be ready to be analysed before H_{0B}. However to apply the closed testing method they need to be assessed at the same time, which could be a major limitation in some adaptive designs.

Due to the nature of closed testing methods, the significance of one hypothesis will affect the error rates for the other. The following scenarios are included in the simulations:

1. No difference between any of the treatment effects: $\theta_A = \theta_B = 0$.
2. Treatment A is not different to control but treatment B is: $\theta_B \neq \theta_A = 0$.
3. Treatment B is not different to control but treatment A is: $\theta_A \neq \theta_B = 0$.
4. Both experimental treatments are different to control: $\theta_A \text{ and } \theta_B \neq 0$.

In each case, if there is no experimental treatment effect, the survival estimate for the experimental arm is set to equal that for the control, and if there is an experimental treatment effect, the survival estimate is set to the clinically relevant improvement.

The weights for the INVN combination method are determined based on the number of events expected for each treatment in each stage. The median follow-up is 87 and 57 months for the first and second stages respectively for each arm, so the number of events from patients in the first of the stages for each hypothesis can be calculated to be 216, compared to 163 from the second, assuming they follow an exponential distribution. In Stage 2, some of the control group patients overlap for both hypotheses, leading to 93 of the events contributing to both hypotheses within that stage. Therefore the total events occurring from all patients within each stage are 216, 286 and 163 respectively. Table 5-14 gives the results from the simulations comparing the adjustment methods under various assumptions.

Table 5-14 Results of 100,000 simulations to assess the probabilities of rejection with the multivariate and inverse normal combination analysis methods, comparing different multiplicity adjustment techniques

	Percent H_0 rejected							
	$\theta_A = \theta_B = 0$		$\theta_B \neq \theta_A = 0$		$\theta_A \neq \theta_B = 0$		θ_A and $\theta_B \neq 0$	
	MV	INVN	MV	INVN	MV	INVN	MV	INVN
No multiplicity adjustment								
Reject H_{0A}	5.09	5.03	4.97	4.91	80.50	80.42	80.34	80.27
Reject H_{0B}	5.09	5.06	80.59	80.52	5.05	5.00	80.37	80.29
Reject at least one H_0	9.74	9.67	81.69	81.59	81.64	81.56	94.38	94.32
Overall Bonferroni adjustment								
Reject H_{0A}	2.56	2.54	2.48	2.44	71.81	71.68	71.48	71.36
Reject H_{0B}	2.59	2.56	71.90	71.79	2.50	2.49	71.74	71.64
Reject at least one H_0	5.02	4.97	72.64	72.52	72.57	72.44	89.52	89.41
Overall closed testing adjustment								
Reject H_{0A}	2.66	2.63	4.21	3.51	71.99	71.73	77.45	77.33
Reject H_{0B}	2.70	2.65	72.09	71.85	4.27	3.56	77.60	77.49
Reject at least one H_0	5.02	4.97	72.64	72.52	72.57	72.44	89.52	89.41
Stagewise closed testing adjustment								
Reject H_{0A}		1.96		2.49		57.31		79.10
Reject H_{0B}		1.81		54.08		2.47		79.27
Reject at least one H_0		3.39		54.34		57.51		92.14

MV=multivariable analysis method

INVN=weighted inverse normal combination analysis method

As expected, the multivariable and weighted inverse normal combination analysis methods give similar results. With no adjustment the approximate two-sided type I errors for individual hypotheses are 5%, the FWER is 9.75%, and the power for each individual hypothesis is 80%. The 'disjunctive power' to reject at least one of the null hypotheses if both contribute to the same claim of effectiveness approximately 81% in the case where only one of the treatments is truly superior, to 94% when both treatments are superior. With the Overall adjustment methods, the FWER is strongly controlled at the 5% level, with the power for individual hypotheses reduced to approximately 72%, or 77.5% if both experimental treatments differ from control in the

closed testing case, and the ‘disjunctive power’ is 72% in the case where only one of the treatments is truly superior, and 90% when both treatments are superior. Note that the Bonferroni and Overall closed testing adjustment methods give the same probabilities of rejecting at least one hull hypothesis, as expected as illustrated by Figure 4-4 and Figure 4-5 and described in Section 4.4.1.2, because both methods reject the most significant null hypothesis if the p-value is less than α/m. The difference between the methods relates to how the least significant null hypothesis is assessed given that the most significant (and therefore ‘at least one’) has been rejected.

The results based on the Stagewise method of adjustment are surprising. The FWER is controlled at the 5% level, in fact it is less than 3.5%; and if both experimental treatments are different from control the power for each hypothesis is 79%, so hardly reduced from that with no adjustment. However, if one experimental treatment is different and the other is not, the power to reject the single null hypothesis suffers a large penalty. This can be explained using the worked example in Table 5-15, in which with no adjustment experimental treatment *A* is no different to control, but experimental treatment *B* is significantly different. With an Overall adjustment method, H_{0B} can be easily rejected. However, using the Stagewise method, the impact of having to include the stage 1 results based on treatment *A* alone in the calculation for the intersection is large, causing the intersection null hypothesis to fail to be rejected. Posch et al.[18] acknowledge this when discussing adding new treatments: “Note that if p_1 is large, this is a serious penalty for the rejection of the new null hypothesis. This is the price to be paid for the great flexibility provided by the adaptive design”. This phenomenon is exaggerated when adding an arm or with multi-arm multi-stage designs in particular over other types of adaptation. If an arm is dropped based on interim data, the second stage would only include the better arm from the first stage, therefore the poor performance of the dropped arm would not contribute to the intersection p-value for either stage. Similarly if the adaptation did not involve adding or dropping arms, the poorer arm would not influence any of the intersection p-values. However in designs where arms are added, or where arms are stopped because they have completed recruitment rather than based on performance, the Stagewise adjustment method could lead to a very large disadvantage in the case that the experimental treatments perform differently. For example in a dose trial, if less active smaller doses are assessed first and more active larger doses are added later, the power to declare significance for a larger active dose is heavily penalised.

Table 5-15 Worked example of the Stagewise and Overall closed testing adjustments within a p-value combination analysis when Hypothesis *A* is not significant but Hypothesis *B* is significant

	Stage 1	Stage 2	Stage 3	P-value Combination (calculated)
Number of events (for weighting the INVN combination)	$e_{1,A} = 216$	$e_{2,A} = 163$ $e_{2,B} = 216$ $e_{2,AB} = 286$	$e_{3,B} = 163$	
H_{0A}	$p_{1,A} = 0.823$	$p_{2,A} = 0.705$	-	$p_A = 0.854$
H_{0B}	-	$p_{2,B} = 0.112$	$p_{3,B} = 0.021$	$p_B = 0.012$
Intersection $H_{0(AB)}$ (calculated)	$p_{1,AB} = 0.823$	$p_{2,AB} = 0.224$	$p_{3,AB} = 0.021$	**1. Overall** $p_{AB} = 0.024$ **2. Stagewise** $p_{AB} = 0.165$

The example in Table 5-16 illustrates why the power based on the Stagewise adjustment method is not much lower than that when there is no adjustment where $\theta_A \text{ and } \theta_B \neq 0$. In this example both of the individual combined p-values are greater than 0.04, so the null hypotheses would not be rejected after adjustment by the Overall Bonferroni or closed testing methods. However, the Stagewise intersection p-value is only 0.032 so the null intersection hypothesis can be rejected, allowing the individual p-values to be compared to 0.05. Only Stage 2 has taken a penalty for having multiple arms, so the effects of the adjustment are diluted.

Table 5-16 Worked example of the Stagewise and Overall closed testing adjustments within a p-value combination analysis when both hypotheses are significant with no adjustment

	Stage 1	Stage 2	Stage 3	P-value Combination (calculated)
Number of events (for weighting the INVN combination)	$e_{1,A} = 216$	$e_{2,A} = 163$ $e_{2,B} = 216$ $e_{2,AB} = 286$	$e_{3,B} = 163$	
H_{0A}	$p_{1,A} = 0.124$	$p_{2,A} = 0.093$	-	$p_A = 0.041$
H_{0B}	-	$p_{2,B} = 0.132$	$p_{3,B} = 0.107$	$p_B = 0.049$
Intersection $H_{0(AB)}$ (calculated)	$p_{1,AB} = 0.124$	$p_{2,AB} = 0.186$	$p_{3,AB} = 0.107$	**1. Overall** $p_{AB} = 0.082$ **2. Stagewise** $p_{AB} = 0.032$

The results from Table 5-14 have been validated by being reproduced using the example from Section 5.4.3.1 based on normally distributed outcome data. Table 5-17

shows the simulated results based on the example with normally distributed outcome data, and it can be seen that the error rates are very similar to those with survival outcome data above.

Table 5-17 Results of 100,000 simulations based on normally distributed outcome data to assess the probabilities of rejection with the multivariate and inverse normal combination analysis methods, comparing different multiplicity adjustment techniques

	Percent H_0 rejected							
	$\theta_A = \theta_B = 0$		$\theta_B \neq \theta_A = 0$		$\theta_A \neq \theta_B = 0$		θ_A and $\theta_B \neq 0$	
	MV	**INVN**	**MV**	**INVN**	**MV**	**INVN**	**MV**	**INVN**
No multiplicity adjustment								
Reject H_{0A}	5.10	5.09	4.93	4.95	80.15	79.95	80.19	80.05
Reject H_{0B}	4.99	5.00	80.17	79.95	4.98	4.97	79.98	79.84
Reject at least one H_0	9.65	9.68	81.34	81.14	81.31	81.10	93.99	93.92
Overall Bonferroni adjustment								
Reject H_{0A}	2.56	2.54	2.47	2.47	71.32	71.13	71.21	71.11
Reject H_{0B}	2.54	2.54	71.27	71.05	2.47	2.50	70.96	70.84
Reject at least one H_0	4.97	4.96	72.10	71.89	72.12	71.94	88.58	88.54
Overall closed testing adjustment								
Reject H_{0A}	2.66	2.64	4.18	3.54	71.54	71.18	77.05	76.96
Reject H_{0B}	2.65	2.65	71.47	71.12	4.21	3.58	76.78	76.64
Reject at least one H_0	4.97	4.96	72.10	71.89	72.12	71.94	88.58	88.54
Stagewise closed testing adjustment								
Reject H_{0A}		1.94		2.48		54.98		78.85
Reject H_{0B}		1.90		54.90		2.47		78.58
Reject at least one H_0		3.44		55.10		55.16		91.46

MV=multivariable analysis method

INVN=weighted inverse normal combination analysis method

Similarly to Table 5-14, the power for individual hypotheses has suffered from a large loss of power using the Stagewise adjustment method in the case that one treatment is significantly superior and the other is not. This is of the same magnitude as the example based on survival endpoint data.

5.5.4 Discussion on multiplicity adjustment when adapting a trial by adding new treatment arms

If multiplicity adjustment is required in an adaptive trial in which different stages include different arms, adjustment should be across the trial as a whole and not just the stages in which the treatments overlap. This is because the FWER is not inflated due to the treatments recruiting concurrently, but only due to the ability to test multiple hypotheses leading to an increased probability of multiple false positive outcomes across the protocol. This contradicts the majority of methodological literature on adaptive designs, which advocates adjustment within stage. Bretz et al.[98] compared the power of Stagewise and Overall (single-stage) adjustment methods in the cases where an arm was and was not dropped at interim, using an adaptive Dunnett adjustment rather than closed testing. They concluded that the Stagewise strategy is actually slightly less powerful than the Overall strategy when both treatments are present in both stages. In the case of selecting one treatment at interim, however, they further note that “the single-stage tests pay too high a price for multiplicity and thus perform inferior to the adaptive combination tests”. This is expected because adjustment is only in one of the stages, so the total level of adjustment will be less, and the selected treatment will be present in all stages to contribute to the final combined p-value. However, whilst Stagewise adjustment is more powerful, that does not necessarily mean it is appropriate, and for each trial design the type of adjustment needs to be fully considered with respect to what it is aiming to control. A full comparison of Stagewise and Overall adjustment in adaptive designs more generally would be an interesting area for further work.

Where trials are amended by adding arms, or dropping arms for a reason other than their performance, Stagewise methods lead to very low power in the case where the experimental treatments perform differently from one another. This is because the combination p-value for the intersection includes results from stages that do not contain the experimental treatment of interest. The smaller the overlap for the concurrent recruitment, the greater the impact of the other stages and the larger the potential penalty. In these types of design, Stagewise adjustment seems to be clearly inappropriate and is not recommended.

When trials are adapted by adding and dropping arms in different stages, such as in the FLAIR trial (illustrated in Figure 5-1), it might be the case that the hypotheses reach

their final analysis triggers at different times. This differs from the majority of adaptive designs, where the primary analyses are at the end of the trial regardless of the adaptations made along the way. It isn't adequate to only adjust for multiplicity based on the information up to the time when the analysis takes place, the adjustment should be over all hypotheses that contribute to the same claim of effectiveness across the trial as a whole. In order to apply most adjustment methods, such as closed testing methods, it is required that the final analyses for all hypotheses take place at the same time. This would make adding arms much less attractive and likely unethical if it delays the final analysis for the earlier hypotheses. However, for the crude Overall Bonferroni or Sidak adjustment methods, the only requirement is that the number of hypotheses is known. Therefore, although the methods are slightly more conservative, a Bonferroni or Sidak adjustment is likely to be the best option in multi-arm multi-stage type trials where multiplicity adjustment is deemed necessary.

5.6 Discussion and Summary

5.6.1 Discussion

Methods of analysis after the addition of a new treatment arm to an ongoing trial have been considered in this chapter. Adaptive p-value combination methods to analyse trials of this type were considered for three reasons. Firstly, key methodological literature which discusses the addition of a new treatment arm to an ongoing trial describes the analysis based on this methodology (Section 5.1.5). Secondly, this type of amendment could cause a stage effect due to a shift in the population, and it isn't clear how best to conduct the analysis to account for this. Finally, where multiplicity adjustment is necessary in multi-stage designs, closed testing methods are often applied alongside p-value combination methods in order to adjust within stage, and this is assumed to be advantageous over adjusting the final, overall p-values where the stages include different hypotheses[98].

In flexible designs literature generally, adaptive analysis methods are required, such as p-value combination over the stages, to satisfy the Conditional Invariance Principle (Section 5.1.3). Adapting a trial to add a new hypothesis, however, is very different to adapting an existing hypothesis based on interim data, and in this case it has not been considered whether adaptive analysis methods are necessary. It is assumed that the existing hypothesis (H_{0A}) is not being amended at all, only the protocol and randomisation are affected. In a simple case, the amendment to add the new arm

might be informed entirely by data external to the trial and with no internal analysis having taken place. The Conditional Invariance Principle is therefore not relevant because the test statistics in the second stage for either hypotheses have not been informed by data from the first stage, and therefore it seems appropriate to conduct a pooled final analysis. If there is an interim analysis for H_{0A} this could affect the decision to add H_{0B} or its design characteristics, for example if outcomes for a low dose suggest that investigation of a higher dose would be beneficial, or if the interim treatment effect informs the new power calculation. However, this does not affect any design aspects of H_{0A} itself in Stage 2. Therefore, for H_{0A}, the second stage test statistics remain independent of the first-stage data. The second stage test statistics for H_{0B} could have been informed by the interim data for H_{0A}, but these data do not contribute to the H_{0B} analysis since only concurrent controls are used. Therefore, it seems reasonable that adaptive analysis methods are not required for H_{0A} due to amending the protocol to add a new treatment arm. Similarly, if Treatment A is stopped at the end of Stage 2 because it has completed its planned recruitment, adaptive analysis methods are unlikely to be necessary for H_{0B} over Stages 2 and 3. Adaptive analysis methods would be required if some amendments to the design of the existing hypothesis based on interim data were also made during the protocol amendment to add Treatment *B*, or if the randomisation allocation between Treatments *Z* and *A* was changed alongside the amendment[53].

Simulations showed that where there is a stage effect, a multivariable analysis adjusting for trial stage performs as well as the adaptive weighted inverse normal p-value combination method. The multivariable analysis has the advantage that statistics other than the p-value, such as the effect size and confidence intervals, are more readily available. In addition, the stage effect is also assessed within the model, which can help interpretation of results. Whilst p-value combination methods are necessary to avoid bias in the case when the adaptation was informed by interim data, they only output a p-value and it is difficult to clinically interpret the results by assessing the treatment effect estimate and confidence interval. Therefore where it is appropriate to use a multivariable analysis, this is likely to be the preferred method.

In all simulations, simply pooling the data without adjusting for trial stage performs somewhat worse than the adjusted multivariable analysis if there is a large stage effect. In the case of normally distributed outcome data, pooling the data without adjustment performs much worse, even with only a reasonable stage effect. It is therefore

recommended that where a pooled analysis method is used, a stage covariate is incorporated in a multivariable model.

In trials in which arms are added and multiplicity adjustment is deemed necessary, it is not appropriate to use a ‘Stagewise’ adjustment method which adjusts the within stage p-values before combining across the stages. This is because, firstly, the FWER is not inflated within stages due to the shared control data, but is inflated over the trial as a whole due to the ability to test multiple hypotheses, so adjustment should be across the whole protocol. Secondly, ‘Stagewise’ adjustment methods do not perform as expected in trials in which arms are added, or dropped for a reason other than futility, therefore leaving some stages with only the poorer performing experimental treatment. An ‘Overall’ adjustment method is therefore recommended to be applied to the final p-values for each hypothesis.

In summary, since adaptive analysis methods are not required for the original hypothesis due to amending the protocol by adding a new treatment arm, p-value combination methods do not perform better than an overall multivariable analysis adjusting for Stage if there is a stage effect, and because multiplicity adjustment is recommended over the protocol as a whole rather than within Stage, there is no benefit to using p-value combination methods in this case. It is appropriate to conduct a pooled final analysis, and a multivariable analysis adjusting for trial stage before and after the arm is added is recommended. This recommendation also holds for H_{0B} in the case where Treatment *A* completes recruitment, therefore creating Stages 2 and 3.

If there is an interim analysis for H_{0A} that informs the decision to add the new arm, whether it is planned or unplanned, there might be perceived investigator bias in the situation of trying to ‘save’ a failing trial. This will have the effect of inflating the FWER across the protocol, in which case multiplicity adjustment could be necessary due to multiple chances of success. However, if the trial design for the original hypothesis H_{0A} is not amended in any way, it is difficult to understand why adaptive analysis methods for this hypothesis would be necessary or beneficial. This is in contrast to Wassmer’s recommendations (Section 5.1.3) that p-value combination methods are likely to be required if the decision to add the current treatment was made based on information from the current trial. This point is perhaps a conceptual one, rather than necessarily statistical in terms of being able to calculate the effect of this knowledge on the error

rates, and would benefit from discussion with experts in the field and regulators to obtain a consensus of opinion.

In a long trial, or a trial with many hypotheses, in which arms are dropped and added on a rolling basis, multiplicity adjustment over the whole protocol could be a limitation that causes the design to be unfeasible. If different therapies are being assessed or the hypotheses contribute to different claims of effectiveness, it may be justifiable that adjustment is not required. However if, for example, a number of doses of the same treatment are added and dropped throughout the trial and adjustment is necessary, researchers would need to carefully consider the effect of adjustment on the power for the existing and new hypotheses, as well as the overall power to make a claim of effectiveness, when considering the feasibility of the amendment.

The recommendations here for applying multiplicity adjustment over the whole trial, rather than within stage, are limited to designs that are adapted by adding and stopping arms on a rolling basis. They cannot readily be extrapolated to other types of adaptive design; the Overall and Stagewise methods would need to be considered and compared to assess the FWER and power in these cases. The phenomenon in rolling designs in which Stagewise methods lead to low power in the case of one experimental treatment being effective and the other not is not seen in most other types of adaptive design, because there is no possibility for a stage to include only the poorer performing therapy, for example if a treatment is dropped for futility or selected for efficacy. The only other example of this phenomenon is if an arm is stopped at an interim analysis for efficacy, but other therapies continue, in which case an Overall adjustment method is also recommended. In designs where an adaptation is made at an interim analysis but all stages include all treatments, both methods do adjust over the whole protocol, and both strongly control the FWER although they work differently, so further consideration is needed.

If there is a planned interim analysis for either hypothesis, the consequences of this with relation to adding an arm need to be considered. Should the interim analysis for the original hypothesis be planned after the amendment, part of the design stage for the amendment should include consideration of whether any information arising from the interim is likely to affect the continuation of the new hypothesis. For example if there was a large efficacy benefit for the original experimental arm this could change practice, therefore altering the control therapy for the new hypothesis, and it should be

considered how this would be handled or whether any provisions need to be put in place to plan for this possibility. The inclusion of interim analyses introduce additional complexity when adding an arm, and the timing along with the intention of any analyses and their potential effect on the other hypotheses need to be considered on a trial by trial basis.

The recommendations in this section are primarily based on an example with time to event outcome data, but it has also been shown that there is no difference in recommendations for the general analysis methods assessed regardless of the outcome type. Multivariable analyses can adjust for stage using regression or logistic regression models as appropriate for continuous or binary outcomes; and p-value combination methods are unaffected by how the p-values are calculated, so the results here easily extrapolate. Time to event outcomes are the most complex of these because the weightings in the weighted combination methods need to be based on estimated numbers of events rather than actual numbers of patients.

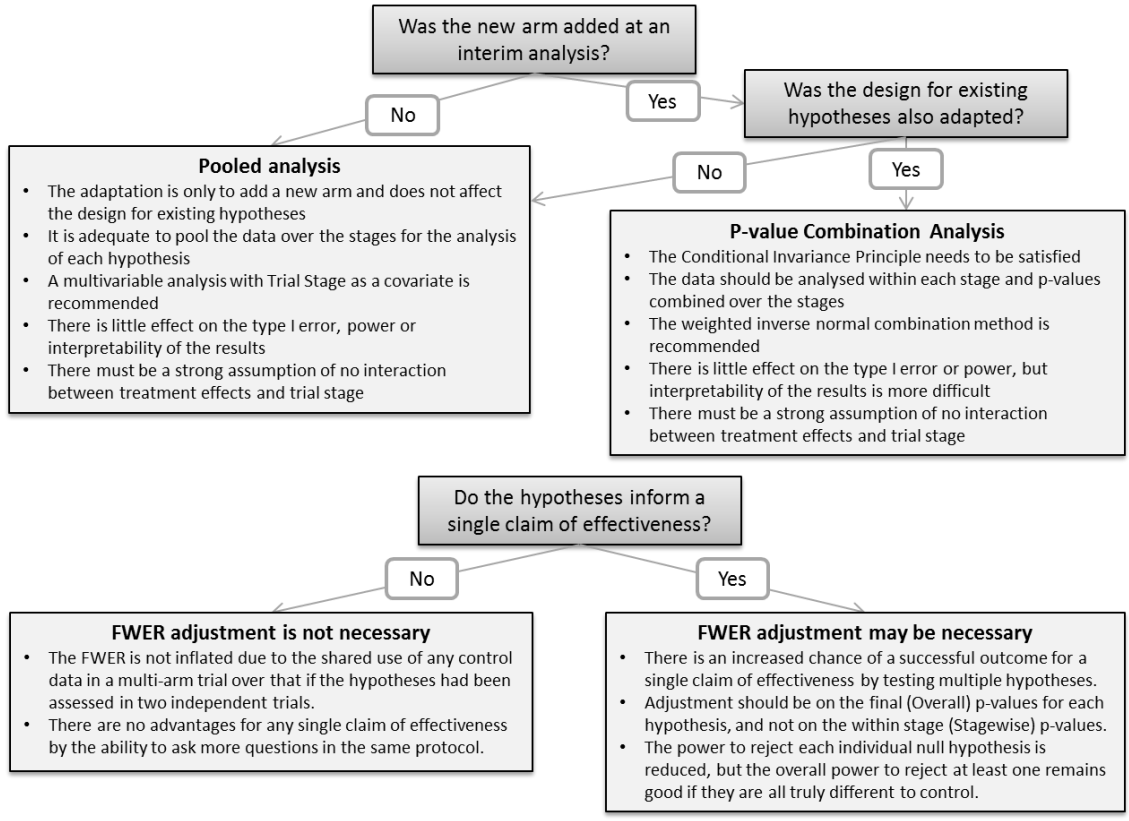
If there is a possibility of a treatment*stage interaction due to adding an arm, the final results could be uninterpretable. If it is considered possible that the change to the design could affect the outcomes differently within each trial arm, adding an arm is not advised. Note that this is also the case when an arm is dropped, as both situations could cause treatment*stage interactions. There is likely to be very low power to assess this statistically, so strong clinical justification of no interaction is needed at the design stage for the amendment.

In the survival case, if the assumption of an exponential survival pattern does not hold, adaptive analysis methods may not be appropriate. This is the case for all types of adaptive designs, and should be considered when deciding whether it is appropriate to adapt a trial. For example, in breast cancer ER+ and ER- patients have different relapse profiles, and so later stages with shorter follow-up would be weighted towards ER- patients, potentially biasing the combined results similarly to if there is a treatment*stage interaction. This would need further investigation on a trial by trial basis, with consideration of a biomarker stratified design.

5.6.2 Summary

Figure 5-3 below summarises the overall findings from this chapter.

Figure 5-3 Summary diagram of the methods of analysis and multiplicity requirements when a new treatment arm is added to an ongoing trial



Adding an arm to an ongoing trial does not always need to make the analysis of the trial complicated, and it could be extremely advantageous, with few statistical penalties. In particular, if the only change to the trial is that the arm is added to the protocol and randomisation, without adapting the design for the existing hypotheses, there is minimal impact on the analysis. If the new treatment arm is added at an interim analysis in which the design for the existing hypotheses is also adapted, p-value combination methods are required to satisfy the Conditional Invariance Principle for the existing hypotheses. In this case, if multiplicity adjustment is necessary, it should be performed on the final p-values calculated across the trial as a whole, rather than within the stages prior to their combination. The trial results could be compromised if there is an interaction between the treatment effects for different therapies over the trial stages, so care must be taken that this assumption is addressed and justified. In general, adding a new hypothesis to an ongoing trial could be relatively straightforward and widely applicable in practice to greatly improve the efficiency of clinical trials. The next chapter describes details of the FLAIR trial that has been amended to add a new

experimental therapy, including the advantages and efficiencies, statistical considerations and steps to prevent bias, and trial management considerations.

Chapter 6
Adding a new experimental research arm to an ongoing trial in practice: The FLAIR amendment in Chronic Lymphocytic Leukaemia

6.1 Background

6.1.1 Aims

FLAIR (Front-Line therapy in CLL: Assessment of Ibrutinib-containing Regimes) is a randomised, controlled phase III trial in patients with previously untreated Chronic Lymphocytic Leukaemia (CLL) sponsored and managed by the University of Leeds. The primary aim of the trial when it was originally designed was to assess current standard therapy with fludarabine, cyclophosphamide and rituximab (FCR) against ibrutinib with rituximab (IR) in terms of progression-free survival (PFS). At the outset of the trial, 754 patients were planned to be randomised in 4 years, with primary outcomes being available after a further 4 years of follow-up. The protocol describing the trial as originally designed has been published, for reference[102]. The trial opened to recruitment in September 2014, with 70 UK centres planned, and has recruited consistently ahead of target. During recruitment, early evidence emerged of another very promising treatment combination in this population, ibrutinib with venetoclax (I+V). In order to be able to assess I+V in a phase III trial in the same population in the United Kingdom in a timely manner, it was added into the existing FLAIR trial framework after 2 years 10 months of the planned recruitment period.

The aim of this chapter is to summarise the methodological and practical issues involved in successfully amending the FLAIR trial to include this promising experimental therapy so that its assessment could be expedited into a phase III setting. A summary is given of how the strategy used can improve the efficiency and relevance of phase III trials, reducing the time taken to answer new and important clinical questions without compromising the original design and with statistical validity. In this way, this type of amendment is not only acceptable to, but actively benefits patients, researchers, funders, regulators and the wider research community.

6.1.2 Introduction to Chronic Lymphocytic Leukaemia and its treatment at the time of designing FLAIR

Section 6.1.2 is amended from the FLAIR protocol[102], and was primarily written by the Chief Investigator, Professor Peter Hillmen. CLL is the most common adult leukaemia, affecting 6.6 per 100,000 population[103]. The incidence of CLL increases with age and almost twice as many men are affected as women. CLL results from the clonal proliferation of B-cells and is diagnosed by the pattern of expression of various cell surface antigens on the CLL cells. Patients most commonly present with lymphocytosis, lymphadenopathy, splenomegaly and systemic symptoms, such as fatigue, weight loss and malaise. The clinical course of CLL is highly variable with a median survival from diagnosis in the region of 7 years.

Combination chemotherapy with fludarabine, cyclophosphamide and rituximab (FCR) is the standard therapy in patients who are fit and young enough to tolerate it, although even in those patients only 75% tolerate a full 6 cycles. However FCR is associated with significant short and long term toxicity, such as myelodysplasia (MDS) and acute myeloid leukaemia (AML), with virtually all patients relapsing and eventually becoming refractory to therapy before dying either as a complication of the disease or its therapy, usually due to infection. Hence more effective, targeted therapies that improve remission rates and reduce relapses with fewer side effects are required.

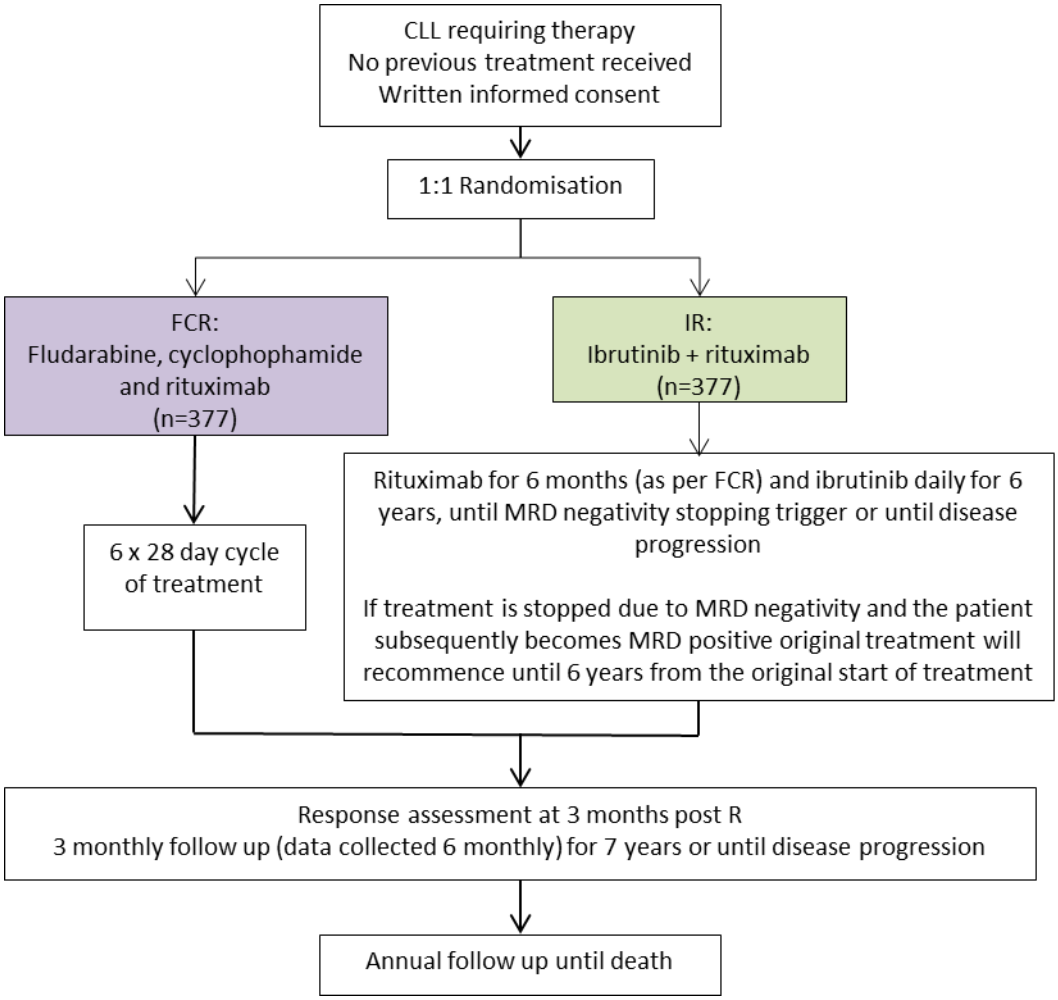
CLL cell proliferation is dependent on stimulation through the B-cell receptor (BCR) and since this pathway is specific for B-cells, including CLL cells, then this is a target in CLL. At the time of designing FLAIR there were several treatments targeting molecules on the BCR pathway including Syk, PI3K delta and Bruton's Tyrosine Kinase (Btk), and all demonstrated activity. Of these agents it appeared that ibrutinib was the most potent agent with relatively minor side effects in early Phase II trials. In addition, ibrutinib is not genotoxic and therefore would not be expected to lead to the late effects, such as MDS and AML, which are seen with FCR. One of the features of ibrutinib and other BCR pathway antagonists is that they have a characteristic pattern of response with an immediate improvement in symptoms and bulky lymphadenopathy but with a transient increase in circulating CLL cells which can take many months to resolve. The addition of rituximab to ibrutinib was hypothesised to prevent, or at least attenuate, this lymphocytosis so the combination of ibrutinib and rituximab (IR) was selected as the experimental therapy to be compared to FCR.

6.1.3 The original FLAIR trial design

The original design of FLAIR was a phase III, multicentre, randomised, controlled, open, parallel group trial comparing IR against the current standard FCR in patients with previously untreated CLL. Eligible patients have previously untreated CLL requiring therapy, with no more than 20% 17p deletion. A total of 754 participants were to be randomised on a 1:1 basis to receive therapy with FCR or IR. Participants randomised to FCR receive a maximum of 6 cycles with each cycle being repeated every 28 days. Participants randomised to receive IR receive 6 cycles of rituximab with each cycle being repeated every 28 days. Ibrutinib is taken daily for 6 years, until minimal residual disease (MRD) negativity stopping rules (Section 6.1.4) are reached or until disease progression.

The trial aims were to provide evidence for the future first-line treatment of CLL patients by assessing whether IR is superior to FCR in terms of PFS and whether IR toxicity rates are favourable. Other key endpoints to be assessed included: overall survival; attainment of undetectable MRD; response to therapy; health related quality of life and cost-effectiveness; as well as an evaluation of discontinuation and re-continuation of ibrutinib therapy if indicated based on levels of residual disease. Randomisation used minimisation with a random element to ensure treatment arms were well-balanced for the following participant characteristics: Binet stage (A progressive or B, C), age group (≤65 years, >65 years), gender (male, female) and centre (all participating centres). Figure 6-1 illustrates the original participant pathway.

Figure 6-1 Participant pathway into FLAIR prior to the amendment



The experimental arm is shown in green, and the control arm in purple.

The primary objective is to assess IR vs FCR in terms of PFS.

The sample size was based on testing the null hypothesis of no difference in PFS between the treatment arms. Based on results from the German CLL8 trial[104], the median survival in the FCR arm was assumed to be 4.5 years. To test a superiority hazard ratio of 0.75, which equates to an increase in median PFS to 6 years in the IR arm, with an overall two-sided 5% significance level and 80% power, assuming 4 years recruitment and 4 years follow-up, allowing for 5% drop-out, and inflating for a planned interim analysis, 754 participants were required to observe 379 events. A formal interim analysis on PFS was planned when half the numbers of events (191 progressions and/or deaths) were observed, in order to allow large differences between the treatment arms to be reported early to the Data Monitoring and Ethics Committee (DMEC). The O'Brien and Fleming alpha-spending function was used to account for testing at multiple time-points to conserve the overall type I error.

6.1.4 Minimal Residual Disease (MRD) negativity

MRD negativity is defined as the presence of <0.01% CLL cells in the peripheral blood or bone marrow. The detection of MRD above this level after therapy is an independent predictor of outcome[105], where detectable disease is prognostic of progression. It is hypothesised that once a patient's disease falls below a certain level it may reach a point at which the CLL cells cannot grow back to being detectable or progressing to a level which requires therapy. In these patients, continuing treatment may be unnecessary. The FLAIR trial therefore includes an MRD negativity stopping rule, in which participants receiving ibrutinib who become MRD negative stop therapy after a certain period of time, determined by the time it took to reach MRD negativity. If participants stop treatment due to MRD negativity and then relapse at the MRD level before the end of the trial treatment period, ibrutinib treatment is restarted to assess whether MRD eradication is re-achieved and to protect the primary endpoint of PFS. This is not considered a progression event.

6.1.5 Funding and approvals

FLAIR is partially funded by Cancer Research UK following review and approval by their Clinical Trials Advisory and Award Committee (CTAAC) in November 2012. Janssen Pharmaceuticals provide ibrutinib free of charge for use in the trial and provide funding via an educational grant. The trial received ethical approval from the NRES Committee Yorkshire and The Humber and regulatory approval from the Medicines and Healthcare Products Regulatory Authority (MHRA) in June 2014. The trial was registered on the ISRCTN registry (ISRCTN01844152) ahead of the first participant being recruited. An independent DMEC and Trial Steering Committee (TSC) were established during trial set-up and approved the original protocol and trial design. The DMEC and TSC both meet at least annually and the DMEC review safety reports on a 3-monthly basis.

6.1.6 Accrual

As with all clinical trials, recruitment is monitored closely by the Trial Management Group (TMG). The TMG decided that it would only be appropriate to add arms during recruitment if the trial recruited at least as well as anticipated and the addition of arms would not significantly delay the reporting timelines of the original design. By the end of 2015 it was clear that recruitment was going to continue at a rate that was 15-20%

ahead of target and the TMG agreed that this was sufficient for an amendment to add new arms.

6.2 Incorporating emerging evidence into FLAIR

6.2.1 Designing the original FLAIR trial to enable amendments

Due to treatment advances, PFS times are increasing, and whilst clearly beneficial to patients, this presents challenges for research in ensuring that trials are feasible and the outcomes remain relevant in the face of a long term endpoint. In addition, the drug development environment in CLL is rapidly changing. At the time of designing the original FLAIR trial, there was a series of phase II trials planned as part of the Bloodwise Trials Acceleration Programme (TAP)[106] run through the University of Birmingham assessing new treatment combinations with targeted therapies, some of which included ibrutinib. The new combinations were hypothesised to give deeper responses than IR, but there was very little evidence of activity or safety in patients with CLL. Our options at the time were to either: wait for the phase II outcomes in case they were positive, delaying the phase III assessment of IR; start the trial as planned, which would saturate the UK population for the coming years and deny the investigation of a new promising combination in a phase III trial; or start the trial but plan to be able to amend it to include new treatment arms if appropriate once early phase data were available. It was clear that in order to speed up the investigation of promising new therapies and improve the efficiency of the phase III trials process in CLL to mirror that in phase II, the latter option was necessary. For this reason, the FLAIR trial had a simple design so it could be more easily amended.

This thesis confirms that new treatment arms have rarely been added to ongoing confirmatory trials in practice, although phase III trials are the longest and most expensive part of the drug development process and doing so would greatly improve efficiency. The work here identifying the methodological considerations and subsequent recommendations has been applied to the FLAIR trial in order to incorporate a new experimental research arm during recruitment. The methodological and practical issues are presented, describing how they were addressed to ensure the FLAIR amendment was a success.

6.2.2 The emerging combination: Ibrutinib + Venetoclax

By the end of 2015, early stage data showed impressive response rates for venetoclax (V) (ABT-199) in combination with rituximab (V+R) in patients with relapsed/refractory CLL, and eradication of detectable MRD in 53% of patients, which had not previously been seen with any other targeted treatments[107]. Based on pre-clinical data it was anticipated that the combination of V plus ibrutinib (I+V) would be highly synergistic given the complimentary modes of action of the two agents, as the ibrutinib arrests CLL cell proliferation and venetoclax is pro-apoptotic leading to early cell death[108, 109]. As FLAIR incorporates MRD negativity stopping rules designed to reduce long-term toxicities and treatment costs it was important to identify a treatment combination with the greatest chance of inducing MRD negativity. It was hypothesised that the addition of venetoclax to ibrutinib would reduce MRD levels faster and more effectively than those expected with I alone or IR, and therefore allow the duration of therapy based on level of disease to be reduced, leading to a reduction in long-term resistance and toxicities and an overall cost saving. I+V was therefore chosen to be assessed in the phase II TAP trial 'CLARITY' (ISRCTN: 13751862) designed to assess I+V in 50 patients with relapsed CLL in a non-randomised setting.

Preliminary results from CLARITY were expected to be available during the first half of 2017, by which time FLAIR would have recruited approximately two-thirds of the planned sample size. The TMG agreed that these timelines were feasible to allow I+V to be added to FLAIR as a new arm but only if work began on designing the amendment and applying for approvals prior to the availability of the phase II safety or preliminary efficacy results from CLARITY. The approval applications were made with the caveat that they would be withdrawn if emerging data indicated. This strategy is discussed further in Section 6.6.1.2.

6.3 Design of the FLAIR amendment

6.3.1 Inclusion of an ibrutinib monotherapy control arm

It was decided that I+V would be added into the FLAIR trial as a new experimental arm, but in order to protect the trial from changes in practice in future, an ibrutinib monotherapy arm was also added as an additional control therapy.

At the time of designing the amendment to add I+V, FCR was still the standard of care in front line CLL, and thus was the required comparator for all experimental therapies. However, in 2016 the ibrutinib licence was extended to include use as a single agent for patients with previously untreated CLL. The TMG therefore felt that an ibrutinib containing therapy could become standard of care in the FLAIR population before the trial was fully reported. As IR was hypothesised to reach deeper responses than I alone, and was being assessed in clinical trials other than FLAIR, it was unclear whether IR or I alone was more likely to become the standard of care long-term in the UK. It was therefore proposed to include an I alone comparator arm at the time of the amendment alongside adding I+V.

In order to ensure the timely reporting of trial outcomes it was not feasible to include both the IR and I alone arms, which would have led to three comparator arms for I+V. It was originally proposed that IR would be closed at the end of the planned recruitment period to the IR vs FCR comparison however following feedback from CRUK it was agreed that a decision would be taken at that time to drop either IR or I alone. The decision on which arm to choose would primarily be made based on anticipated emerging MRD data from other trials that were due to report ahead of the decision point. In this way, the trial was protected in case FCR was superseded as standard of care by either I alone or IR. This mitigated the risk that the outcomes of the trial would be hugely devalued if it were to show that I+V was better than FCR, but FCR was no longer the standard of care. With the proposed design the amended FLAIR trial was future-proofed so that if FCR was no longer the standard when the trial reported, other comparator arms were also included and powered to be able to show a clinically relevant improvement.

The amendment therefore included the addition of two new arms, one experimental and one control. In addition, there were two new primary hypotheses, one comparing I+V to FCR, and the other comparing I+V to I or IR. The additional statistical implications of adding a control as well as an experimental arm are discussed in Section 6.4.

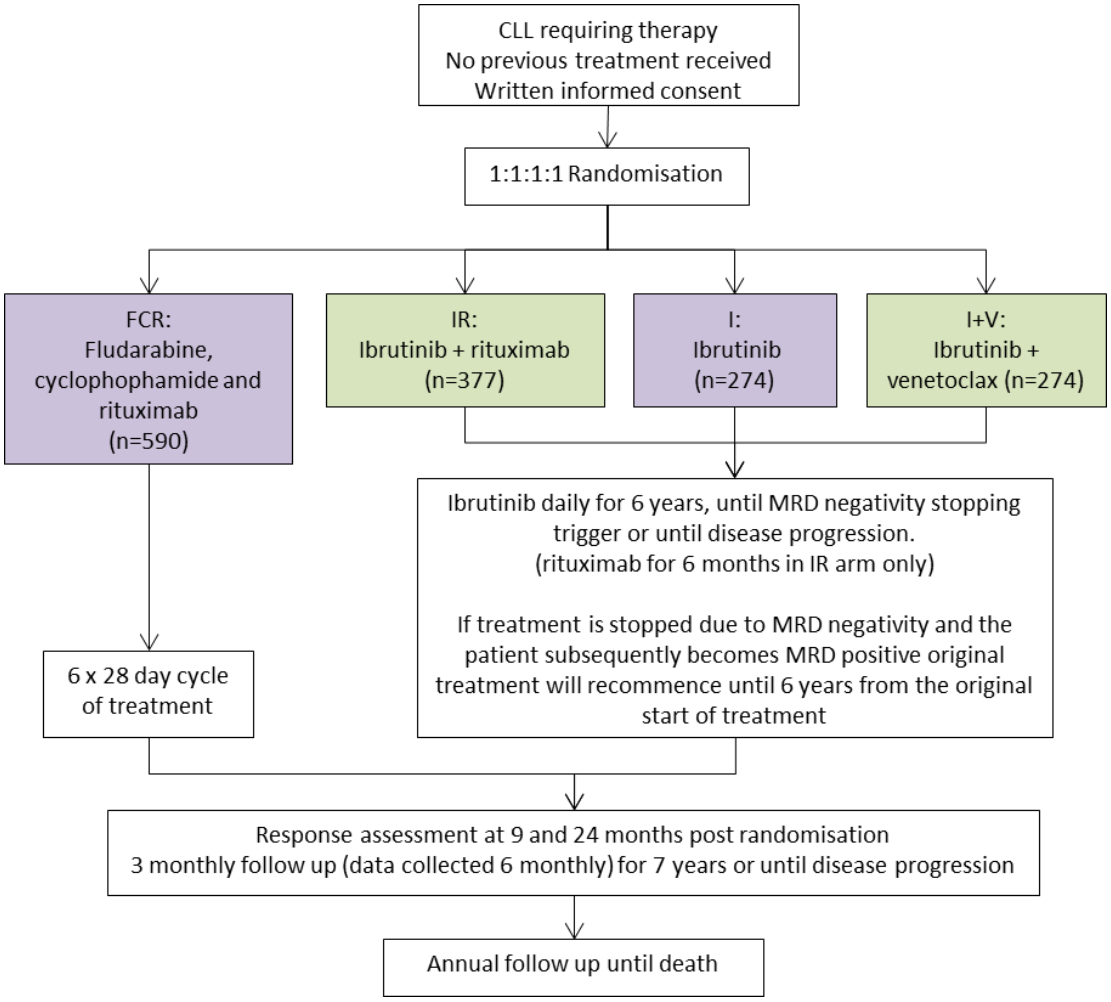
6.3.2 Amended Trial Design

The amended FLAIR design was a phase III, multi-centre, multi-arm, randomised, controlled, open, parallel group trial in patients with previously untreated CLL. Participants were randomised to receive FCR, IR, I monotherapy or I+V on a 1:1:1:1

basis. The eligibility criteria were unchanged from the original design. The treatment schedules for FCR and IR were unchanged. Participants randomised to I+V receive ibrutinib for 8 weeks before venetoclax is added over a four week dose escalation phase. In the I monotherapy, IR and I+V arms, ibrutinib and venetoclax (as relevant) are administered for 6 years, until the MRD negative stopping rules are triggered or until disease progression. If treatment is stopped and restarted due to MRD levels, as described in Section 6.1.4, participants randomised to I monotherapy or IR receive further I monotherapy, and participants randomised to I+V receive further I+V.

The amended trial aims are to provide evidence for the future first-line treatment of CLL patients by assessing whether IR is superior to FCR in terms of PFS, whether I+V is superior to FCR in terms of PFS, whether I+V is superior to I or IR (as appropriate) in terms of MRD negativity, and whether IR and I+V toxicity rates are favourable. The other key endpoints to be assessed remain unchanged from the previous design, but now also compare I+V with FCR and I+V with I or IR. Figure 6-2 illustrates the amended participant pathway.

Figure 6-2 Participant pathway into the amended FLAIR trial



The experimental arms are shown in green, and the control arms in purple.

The primary objectives are:

- *To assess IR vs FCR in terms of PFS*
- *To assess I+V vs FCR in terms of PFS*
- *To assess I+V vs I in terms of MRD negativity rate*

6.3.3 Sample Size

The sample size for I+V vs FCR was based on testing the primary null hypothesis of no difference in PFS between the treatment arms. However, the assumptions for the clinically relevant effect size differed from those for IR vs FCR due to evidence published in 2015 comparing I with chlorambucil[110], demonstrating that ibrutinib monotherapy leads to a better PFS in this population than thought at the time of the original design. To assess a superiority hazard ratio of 0.69 (for a median PFS increase of 4.5 to 6.5 years) with an overall 5% significance and 80% power, assuming a 2.5 year recruitment period and 3.5 years of follow-up, and allowing for a 5% dropout rate, 274 participants were required to be concurrently randomised to each of FCR and I+V

in order to observe 232 events. A total of 822 participants were therefore required to be concurrently randomised to FCR, I or IR and I+V. A formal interim analysis on PFS was planned when half the numbers of events (116 progressions and/or deaths) were observed in FCR and I+V, in order to allow large differences between the treatment arms to be reported early to the DMEC. The O'Brien and Fleming alpha-spending function was used to account for testing at multiple time-points to conserve the overall type I error.

In order to protect the primary outcome in the event that FCR was superseded as the standard treatment during the life of the trial, it was ensured that there was power to compare I+V against I or IR. The rationale for the addition of venetoclax was to reduce the treatment duration needed based on the MRD stopping rule, and therefore the appropriate endpoint for this comparison is the MRD negativity rate. PFS is also included as a key secondary endpoint, but it is confounded by the MRD stopping rule potentially affecting the duration of therapy differently in each arm. The analysis of MRD negativity will be carried out 2 years after the close of recruitment. At the time of designing the amendment the MRD negativity rates in the ibrutinib containing arms were not known so a range of power calculations were carried out. With 260 evaluable patients in each of the arms and a 5% two-sided significance level, there is 90% power to detect an improvement from, say, 10% to 20%. If there are a larger proportion that become MRD negative with IR, say 20%, there is 90% power to detect an improvement to 32.5%, and since a large increase in MRD negativity would be required in order to justify the addition of V, the planned number of patients is more than adequate.

6.3.4 Dropping I or IR

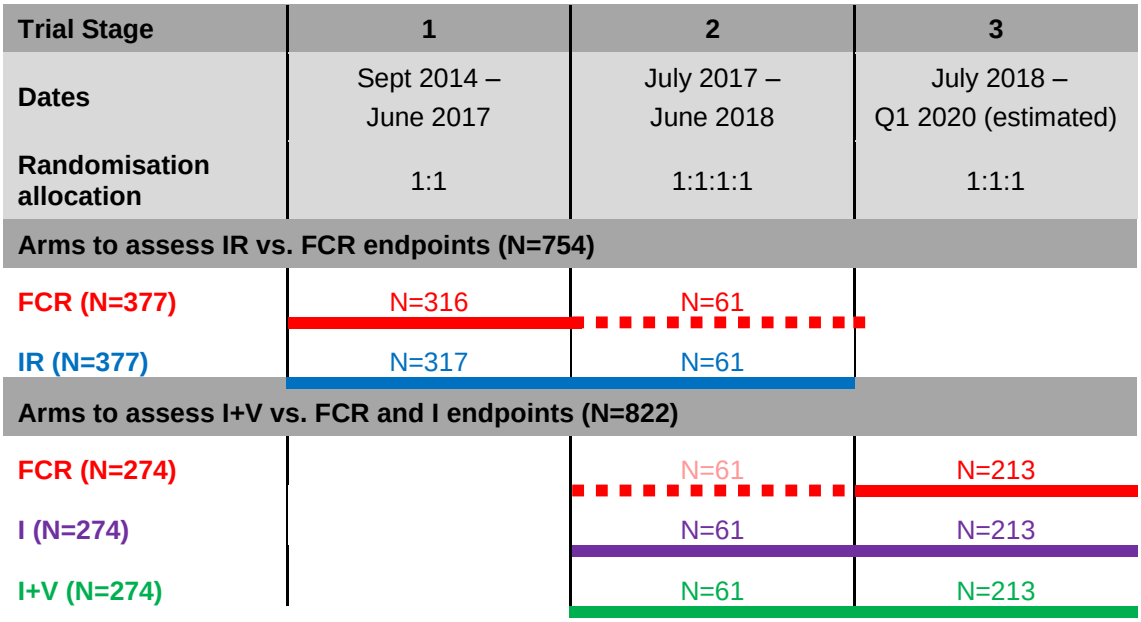
Recall that at the end of the recruitment period to the original FLAIR randomisation (FCR vs IR), it was planned to select the most appropriate comparator for I+V to be either I or IR, in addition to FCR, and to drop the other. The decision was discussed with the DMEC and TSC during February 2018, in order to make the amendment in July 2018, once 754 participants had been randomised to FCR and IR. The emerging evidence from external trials suggested that IR was no better than I in terms of PFS[111], and also that IR did not lead to good enough rates of MRD negativity. In addition, MRD negativity results from IR participants in Stage 1 of FLAIR were summarised for the DMEC, and these strengthened the external evidence. It was therefore agreed that the IR arm would be dropped, and the trial would continue to randomise on a 1:1:1 basis to

FCR, I monotherapy and I+V. There were no changes to the existing treatment schedules or eligibility criteria.

6.3.5 Overview of the trial stages

In total, 1516 participants will be randomised to the trial. A total of 754 participants are required to be randomised concurrently to FCR and IR (stages 1 and 2), and 822 participants to FCR, I and I+V (stages 2 and 3). In addition 61 FCR patients in stage 2 are included in both randomisations, and therefore the total sample size is less than it would have been in independent trials. Figure 6-3 outlines the treatment arms that are included over each stage in the trial. A dotted line indicates that the participants recruited to those arms will be used for more than one comparison.

Figure 6-3 Overview of trial stages



Note that the original FLAIR trial was planned to recruit in 4 years. Even with the addition of the extra arms, recruitment completed a couple of months ahead of schedule. The amendment included additional funding to open more centres, so over 100 were opened rather than the 70 originally planned. In this way, the delivery of the original trial was not compromised by the amendment. The statistical considerations to ensure that the outcomes for either the original or new hypotheses were not biased by the design amendment are discussed below.

6.4 Statistical considerations

6.4.1 Concurrent comparisons

For the analysis of the trial, all primary and secondary endpoint comparisons will only include patients randomised contemporaneously. That is because if there is a shift in the patient population due to the design change or changes in practice over time, it may shift the median survival and could bias the results[53]. Sixty-one FCR patients who were included in the original FLAIR design can also be used as comparators for the I+V arm, therefore reducing the numbers needed compared to a new trial. It will be possible to use data from non-concurrent patients across the whole protocol to carry out exploratory investigations. There will be more similarities between these patients than those from separate trials, and having such a wealth of data on this population could allow subgroups of patients, for example those with certain genetic markers, to be investigated to generate hypotheses that could inform future research. This trial was not designed to report comparisons between the non-concurrent trial arms I monotherapy and IR, or IR and I+V. This is discussed further in Section 6.6.1.3.

6.4.2 Type I error control

As discussed in Section 5.1.4 (Table 5-1), there are a number of ways the type I error could be inflated or bias introduced in a multi-arm adaptive trial design. These are separated out and addressed individually in the context of the FLAIR trial, as follows:

6.4.2.1 Multiple primary outcomes for I+V

The I+V arm is being assessed in two primary outcomes: against FCR for PFS; and against I for MRD negativity. In order for I+V to be acceptable and to be deemed a ‘success’, it needs to be significantly better than both of its control groups. As discussed in Section 4.6 (Figure 4-8), where both hypotheses are required to be superior there is no inflation of the type I error rate and therefore no adjustment is required. The type I error would only be inflated if either one of the hypotheses being positive could lead to I+V being taken forward.

6.4.2.2 Multiple hypothesis testing in the same protocol

This protocol allows the opportunity for both IR and I+V to be declared superior to the current standard within a primary analysis, therefore increasing the chance of a type I error for an ibrutinib containing combination. Whilst both give the opportunity for a

therapy containing ibrutinib to be declared superior, the aim of giving the additional treatments in combination (rituximab and venetoclax) is to be able to stop ibrutinib earlier. In fact, I+V is being compared directly against I monotherapy in terms of its ability to reduce the duration of treatment. Therefore, as discussed in Section 4.6, since a type I error for these comparisons does not directly benefit the same claim of effectiveness for an experimental therapy, FWER control is not necessary for this reason. If the two primary hypotheses had been assessed in separate protocols no adjustment would be required, and in this case it is feasible to assume that the questions would have otherwise been assessed in different trials. Since there is an overlap in recruitment, some of the control data is shared between the IR and I+V vs FCR hypotheses. In Section 4.3 it is also shown that the resulting correlation between the hypotheses reduces the overall type-I error over that if they had been assessed independently, and therefore FWER adjustment is also not necessary due to sharing control data. In summary, adjusting for multiple testing due to assessing multiple experimental arms would be an “unnecessary penalty for efficiency”[49] in this case, so is not planned.

6.4.2.3 Multiple analysis time-points

In order to account for the formal interim analyses allowing for early rejection of the null hypothesis for IR or I+V based on early evidence of efficacy, the O’Brien and Fleming alpha-spending function[112] adjusts for multiple testing in order to conserve the overall type I error. The method recommends that the interim results are compared to a p-value of 0.005, and the final results are then compared to a p-value of 0.048. This is applied to each of the hypotheses separately.

6.4.2.4 Analysis methods following adaptation of design features with combination of information across trial stages

It can be seen from Figure 6-3 that the trial consists of three stages, each with different randomisation options. At the end of stage 1, the design was amended to add two new treatment arms, and at the end of stage 2 the design was amended again to stop recruitment to the arm that had completed. The decision to add the new arms was made without reference to any internal trial data. At the end of stage 2, after the IR vs FCR randomisation had reached its target recruitment, IR was dropped from the trial. As discussed in Section 6.3.4, the decision to drop IR rather than I was made primarily based on data external to the trial, but also on a summary of MRD results from participants randomised to IR in stage 1 only. This has no implications for the analysis

of the IR vs FCR comparison across stages 1 and 2 because the planned recruitment had completed at the time that the amendment to drop IR was made. It also has no implications for the analysis of the I+V vs I comparison across stages 2 and 3 because the data summarised was for the IR arm from stage 1 participants only, and these are external to the concurrent randomisation across stages 2 and 3. In addition, summarising a subset of MRD results for IR patients for the DMEC does not affect the type I error for the final analysis of PFS for IR vs FCR because no randomised comparison was carried out.

In Section 5.6 (Figure 5-3), it was recommended that since the decision to add (and drop) arms was not informed by any analysis of data internal to the existing hypotheses at the time of the amendment, each hypothesis is analysed by pooling the data over the relevant stages, rather than needing to use adaptive analysis methods. A multivariable cox regression is therefore planned to analyse the PFS primary endpoints, and a multivariable logistic regression is planned for the binary primary endpoint of achievement of MRD negativity. These models will account for the trial stage as well as the stratification factors: disease stage, age group and gender.

Whilst the key eligibility criteria did not vary across the stages, it is possible that the different treatment options attracted slightly different patients into the trial. In the first stage, there was a 50% chance of receiving ibrutinib (IR), in the second stage this increased to a 75% chance (IR, I or I+V), and in the third stage 67% (I or I+V). In addition, the number of centres increased leading up to the second stage. In case of any stage effects caused by the changing treatment options or centres, the planned multivariable regressions for all analyses of primacy include Trial Stage as a covariate. However, measures were also put in place to try to prevent changes to the population across the stages due to the likelihood of receiving FCR or an ibrutinib containing therapy. Participants must be eligible and willing to receive any of the treatments, and randomising clinicians were required to ensure that this was the case before randomisation. Early withdrawals were closely monitored, and centres in which patients withdrew for reasons relating to not wanting their randomised treatment, particularly where they were randomised to FCR, were contacted for justification and to ensure their consent processes were appropriate. Therefore, the required assumption of no interaction between treatment effect and trial stage (Section 5.4.3) was felt to be realistic.

6.4.3 Other statistical details

In Section 2.3 the statistical considerations when adding an arm to an ongoing trial are summarised. In addition to the key points detailed above, the following were also considered to ensure the ability of the trial to answer all the primary hypotheses was protected.

6.4.3.1 Power

The new hypotheses comparing I+V concurrently against FCR and I were both formally powered, as described in Section 6.3.3. The design for the original hypothesis was unchanged by the addition of the new arms, so the power calculation remained appropriate, therefore ensuring that there was adequate power to assess each primary hypothesis in the protocol. Since FWER control was not determined to be necessary, no inflation was required to account for this.

6.4.3.2 Randomisation and allocation

Randomisation was by minimisation with stratification and a random element, and the stratification factors were unaltered for the duration of the trial. At each stage, the minimisation algorithm was reset. This was felt appropriate because there were enough patients in each stage that the arms were generally well balanced. Continuing the minimisation algorithm would not be appropriate when adding an arm as all totals would be zero for the new arm at first, distorting the algorithm. There is no reason why resetting the minimisation algorithm would add any bias.

It was decided to maintain an even allocation ratio to all arms in all stages, regardless of the number of experimental treatments. There are views in the literature[6] (Section 2.3.5) that randomising a higher proportion to control might be more efficient in terms of total patient numbers needed when there is more than one experimental arm, or that all arms should complete recruitment at the same time to avoid a third stage. However, in a trial with different treatments in different stages, varying the allocation ratio is not straightforward and having a different ratio in different stages for a single hypothesis would complicate the power calculation and affect the analysis[53]. Analysis methods to minimise bias where different stages include different treatment arms were researched as described in Chapter 5, so there was no need to avoid having a third stage.

6.4.3.3 Challenges due to staggered hypotheses

By the time there was sufficient early evidence of activity and safety for I+V and the amendment had been implemented with the stage 2 randomisation open to recruitment, 633 of the planned 754 participants (84%) had already been randomised into the trial. This reduced some of the advantages of sharing control data in terms of the total patient numbers needed, although there are still many other advantages of this strategy as discussed in Section 6.6.2. Having only a small overlap, however, extends the overall trial and recruitment length, and increases the time between analyses for the hypotheses, which could cause problems. The interim analysis for IR vs FCR was planned based on number of events, and this is expected to occur around mid-2019, which is whilst recruitment to Stage 3 is still ongoing. This interim analysis is initially only reported to the DMEC, but if the results are very positive it might be agreed to release these results more widely. This would have implications for the ongoing trial, which would be discussed in detail with the DMEC and TSC. Some of the FCR patients in this analysis are also in the FCR vs I+V analysis, and it would be considered how to minimise bias by releasing these results so that the integrity of the ongoing trial is not compromised. It may be that it is no longer ethical to recruit to FCR, and therefore that this arm should be dropped, with I+V vs I remaining as the sole primary comparison. In this case releasing the results is unlikely to affect this comparison as the patients are independent, but a fourth trial stage would be created, and in this case a stage effect is likely which will be accounted for in the analysis. Including the I monotherapy arm as a second control was a measure to protect the trial in this eventuality.

After the interim analysis for the IR comparison, the next planned analysis is likely to be the interim analysis for I+V vs FCR which will be reported to the DMEC in 2021, followed by the analysis of MRD for I+V vs I and the final analysis for IR vs FCR in 2022, and lastly the final analysis for I+V vs FCR in 2023. When each of these analyses take place, recruitment to the trial and FCR therapy will have completed, although some patients will still be receiving Ibrutinib or I+V therapy. At the time of each analysis, it will be considered how the results affect the patients still receiving trial treatment, and whether their treatment is still in their best interest. It will also be considered how to interpret and manage the results alongside the previous and future planned analyses. This will be determined by the Trial Management Group in discussion with the DMEC, TSC and NCRI CLL Subgroup once the results are available. The analyses for each experimental treatment can be reported as soon as possible, without affecting the other. In this way the trial is not impeded by its design over if it had been two independent trials.

6.5 Implementation / Trial Management

6.5.1 Approvals and funding

6.5.1.1 Oversight committees

The concept of adding new arms to FLAIR was initially discussed with the DMEC in November 2015 and with the TSC in October 2015. Both groups gave their approval for the necessary funding applications to proceed and agreed that they would approve the new design once funding was in place. The TSC includes a Patient and Public Involvement (PPI) representative who was actively involved in the discussions and the decision to approve the amendment, and felt that the efficiency of the design would be beneficial to patients. The final amended trial design was reviewed and approved by the DMEC in January 2017 and the TSC in February 2017.

The amendment was designed collaboratively with the NCRI CLL Subgroup Committee, which a number of FLAIR Principal Investigators and two PPI representatives are members of, and was agreed in November 2015. The amendment was also presented to, and approved by, the NCRI Haemato-oncology Clinical Studies Group. Both groups were supportive of the amended design and recognised the efficiencies associated with adding new arms to an existing trial rather than designing a separate trial to start after FLAIR had finished recruiting.

6.5.1.2 Cancer Research UK

A no-cost amendment application was submitted to the CRUK Clinical Research Committee in November 2015 for review by the committee in May 2016. This process included an international peer review by four reviewers. One of the peer reviewers identified I+V as a combination with “game-changing potential” and another that “with the amended design of adding ibrutinib and ibrutinib + venetoclax arms, this trial has the potential to help define the standard for frontline CLL treatment worldwide”. Some of the reviewers supported the design methodology of adding new arms with one saying “As the availability for novel agents increases across all types of cancer, studies such as this can be looked at as a model for efficiently answering key questions in a field.” and another that “The planned amendment is essential for this trial to ensure that the conclusions remain relevant when it is due to report”. However others were concerned about the complexity of the amended trial design and if this would impact

deliverability, whether it was reasonable not to adjust for multiple testing given the shared control patients, whether the design would be supported by the relevant pharmaceutical companies, and whether changes in practice could affect the trial long term. All of these points were addressed to the satisfaction of the committee, and approval was granted.

6.5.1.3 Pharmaceutical companies

Due to the relatively short timelines between I+V emerging as an important treatment combination and the original FLAIR design meeting the recruitment target, discussions with pharmaceutical companies had to happen in parallel with the amendment application to CRUK. The amended design included the use of the new Investigational Medicinal Product (IMP), venetoclax, manufactured by Abbvie and a considerably higher number of patients receiving ibrutinib. In advance of the design being discussed with the DMEC, TSC and NCRI committees, initial discussions had been held with Abbvie to establish provisional support for the design. A formal funding application was submitted to them in November 2015 and in February 2016 Abbvie agreed to provide free venetoclax and an educational grant for the additional running costs associated with the new arms, subject to successful contract negotiation. In June 2016 Janssen, the manufacturer of ibrutinib, agreed to provide free ibrutinib for the additional participants in the new arms and associated IMP distribution costs.

To finalise this additional support a contract amendment was required with Janssen and a new contract was required for Abbvie. These contracts were both signed in May 2017. Contract negotiation is a common factor impacting trial set-up times and delaying trials opening to recruitment. These negotiations are made more complex by having multiple pharmaceutical funders and negotiating contracts that comply with charitable funders terms and conditions. It was arguably simpler adding an additional pharmaceutical partner after the trial had opened as the principles around data sharing and intellectual property had already been agreed with one company so there was an understanding that those terms would be equivalent for new funders.

6.5.1.4 Ethical and regulatory

Protocol development, including associated documentation such as the Participant Information Sheet, was finalised in February 2017 following a number of reviews both by the TMG and by pharmaceutical companies. The Participant Information Sheet was also reviewed by the PPI representatives on the NCRI CLL Subgroup Committee.

Substantial amendments were submitted to the MHRA and ethics committee in March 2017 and April 2017 respectively. Ethics approval was received promptly within two weeks of the submission but MHRA approval wasn't received until May 2017. This was delayed as the MHRA requested additional information about the safety of the I+V combination.

6.5.2 Data management considerations

The trial CRFs were updated in line with the trial protocol and were finalised in April 2017. It was decided to amend the existing trial database rather than having a separate database for the new comparisons. This added some limitations in terms of how data for the new arms were collected as it needed to work within the existing database structure, but did not compromise the quality of the data that was collected for the analyses. A new randomisation system was implemented for the four-arm design which meant all centres had to be re-activated on the system and the minimisation algorithm was re-set.

6.5.3 Implementation at centres

A key consideration when the amendment was designed was that the addition of new arms should not significantly delay the reporting of the FCR vs IR comparison beyond the original planned timelines. As the trial was recruiting ahead of target, and the number of recruiting centres was planned to be increased from 70 to 110, the impact on the original analysis timelines was likely to be minimal. Set-up of the additional centres started ahead of the amendment opening to further increase the recruitment rate. Five existing centres decided not to participate in the amended trial, four due to lack of capacity and one because they were unable to cover the cost of MRD testing which was allocated as a treatment cost.

The new randomisation system went live at the beginning of July 2017. Thirty-nine centres opened to the new design within the first week. It was agreed that the old randomisation system would be switched off at the end of August 2017 with all centres needing to have approvals for the amendment in place before then or they would have been suspended to recruitment. Sixty eight centres opened before the original randomisation system was closed, rising to over 100 centres in the following months.

6.6 Discussion

In summary, the strategy of incorporating a new experimental treatment into the FLAIR framework was successful and hugely advantageous, without compromising either the original or new research goals. In this way, we were able to incorporate emerging evidence to test two experimental therapies instead of one, keeping the trial outputs timely and relevant, and minimising resources. There are many advantages to this strategy, and although there were also challenges, these were not unsurmountable.

6.6.1 Challenges

6.6.1.1 Perceived risk

Adapting a trial in any way introduces complexities, both real and perceived. A comment from a CRUK peer reviewer was “Trial design is now more complex, so additional risk that not all components will be completed as planned”. This general feeling that the more complex the design, the more risk is involved was echoed in discussions with clinical and patient representative members of the NCRI CLL Subgroup. Whilst a larger and longer trial with more components will naturally carry more risk, the trials team were careful to consider any potential sources of bias or disadvantages and address them, as discussed throughout this chapter. The original trial question of IR vs FCR is largely unaffected by the addition of the new arms. The number of planned centres were increased from 70 to 110 to ensure that recruitment to the original arms was not negatively impacted by the addition of the new arms, and in fact this comparison recruited ahead of target even with the amendment. The analysis is planned when the data in these arms are mature and without reference to the new arms, so the trial outcomes are not delayed by the amendment. The analysis includes trial stage as a covariate to account for any potential changes to the population caused by adding the arms and centres, although this effect is similar to dropping arms, which is now a commonly accepted strategy. The design for the original hypothesis was not affected by the addition of the new arms, which minimises the complexity as the trial is not truly adaptive. Each primary hypothesis is fully powered; is assessed based on concurrently recruited patients only, which protects against changes in the trial population over time; and statistical aspects relating to error rates due to sharing a protocol and control data have also been considered in detail. Therefore, although the trial is more complex, potential risks have been identified, managed and minimised.

6.6.1.2 Timelines of implementation prior to safety and activity data

In order to make the confirmatory assessment of the emerging therapy as seamless as possible following the phase II Bloodwise TAP CLARITY Trial, the amendment was planned and funding applications submitted prior to the availability of the CLARITY trial outcomes. At the time there was evidence of activity and safety of I+V in mantle cell lymphoma, but the combination had not yet been assessed in CLL. Due to the length of time it takes to obtain funding and approvals, the process was set in motion with the caveat that the applications would be withdrawn and the amendment dropped if the phase II data was not acceptable. Any changes to treatment schedule or safety monitoring that were required for CLARITY would have also been implemented into the FLAIR amendment. Had the emerging results been unacceptable and the amendment dropped, there would have been an amount of work done that had taken place unnecessarily, but this is similar to the risks associated with any grant application. Work on protocol development and other amendment processes was started before contracts with the pharmaceutical funders were signed which also presented a financial risk, however this was felt to be acceptable based on the preliminary approvals from both companies.

6.6.1.3 Protection against changes in practice

One of the concerns with long trials, particularly those with different hypotheses being assessed at different times such as in platform designs, is that practice will change and the outcomes become less relevant or the standard control therapy will be superseded. In order to pre-empt and protect against this, an ibrutinib monotherapy control arm was added concurrently to the I+V arm, so two different control groups were included, as discussed in Section 6.3.1. It is unusual for a confirmatory trial to include two control groups, and clearly the numbers of patients needed is increased compared to a standard randomised controlled trial. However, there was good evidence that the standard therapy could change over the life of the trial so this measure was felt to be necessary. Since the original FLAIR trial was designed, the advantage of ibrutinib monotherapy had been demonstrated in relapsed CLL with the Resonate Trial[113] leading to the marketing approval. In addition results were emerging showing ibrutinib monotherapy to be superior to conventional therapy in previously untreated elderly patients who are unfit for fludarabine-based therapy[110]. A NICE appraisal for ibrutinib use, which covers relapsed CLL and previously untreated CLL for patients with 17p deletion or P53 mutation who are unfit for fludarabine-based chemotherapy, was about to commence[114], making it very likely that ibrutinib monotherapy would become the standard in the UK and beyond for both of these patient populations. An application

had been submitted to extend the ibrutinib license to previously untreated elderly patients, and so a further NICE appraisal to extend front-line use seemed likely. However, as this had not yet happened, FCR was the current standard therapy and thus the necessary control therapy. The strategy of having two controls allowed I+V to be assessed against the most relevant therapies without delaying the research. During the course of the trial, if FCR was superseded it could be dropped from the randomisation without compromising the trial outcomes.

The other concern around changes in practice is that the trial population and therefore patient outcomes could shift over time. For this reason, as previously discussed, this trial is not able to report confirmatory comparisons between the non-concurrent arms: I monotherapy and IR; or IR and I+V. This was a key point raised within reviewers' comments. At the time of the amendment, the hypothesis of IR vs I was being assessed by the US Intergroup Alliance 041202 Trial (NCT01886872), which had randomised 350 newly diagnosed elderly patients to I and IR, and therefore this direct comparison was not novel. Until the results from the IR comparisons were known, it would not have be appropriate to include IR as a control against other experimental therapies, and this would not have been required in a new trial. Whilst early evidence suggested that IR is not better than I, if further evidence emerges that shows otherwise, it is possible to directly compare endpoint data between the 122 contemporaneous patients for each comparison as an exploratory investigation to inform future trial designs. In addition we would have data on MRD negativity, treatment duration and safety rates as well as health economic evaluations for I monotherapy, IR and I+V from within the FLAIR trial to input into this assessment. The outcomes from non-contemporaneous trial arms will be treated as they would had they arisen from different trials.

6.6.2 Conclusions

In this chapter it has been described how the FLAIR trial was able to successfully provide a platform for an emerging new therapy to be assessed within an existing confirmatory trial framework. It is demonstrated that despite challenges and some initial resistance, this type of adaptation is feasible and acceptable. The statistical considerations were addressed to the satisfaction of a number of reviewers, who were convinced that the trial outcomes will be appropriately powered, unbiased and statistically valid for both experimental therapies. In addition, any logistical challenges were not insurmountable. This strategy offered substantial gains in efficiency for

assessing the emerging therapy. For the original trial it took over 2.5 years from submission of the outline funding application to the first centre opening, and 3.5 years for all centres to be open, which is usual for a confirmatory trial. By amending FLAIR rather than planning a new trial the new hypotheses were incorporated almost seamlessly following on from the external phase II assessment, completely eliminating the time period between confirmatory trials. Due to opening additional centres even before the amendment was implemented, the original hypothesis is not delayed in recruitment or reporting. The primary assessment for the new hypothesis is planned just one year later than for the original, which is a saving of many years over planning and running a new trial. In addition, these hypotheses are able to be assessed in the same population at the same time without competing with one another.

The ability to amend FLAIR by adding new treatment arms has greatly benefitted patients because they have access to the latest therapies sooner. This was discussed in a recent article in the BBC news[115] which describes the extremely promising results from the I+V investigation in the phase II CLARITY trial, and how the use of an adaptive trial design enabled this treatment to be quickly incorporated into a randomised confirmatory trial. In addition, fewer patients overall are required to receive the control therapy because of the overlap in Stage 2, and since the trial has been amended patients have a higher chance of being randomised to a targeted therapy.

The FLAIR amendment has demonstrated that adapting a trial by adding experimental arms is feasible in practice without compromising the statistical validity or logistical integrity of the trial. This has opened up the potential for further therapies to be added within this framework following similar methodology should they emerge prior to the close of recruitment. In addition, a similar confirmatory platform including poorer risk and relapsed CLL patients as part of a master protocol would greatly benefit that population and is under discussion. There is no reason why the methodology in this thesis does not readily extend to more complex scenarios, and this is an area for future development.

Chapter 7
Discussion and Guidance

7.1 Summary of the research

Recent initiatives in clinical trials are aimed at speeding up research by making better use of scarce resources. For example, the FDA's Critical Path Initiative "to drive innovation in the scientific processes through which medical products are developed, evaluated, and manufactured" included the production of guidance on adaptive designs to increase the efficiency of studies. In the UK, the National Institute for Health Research Health Technology Assessment (NIHR HTA) programme released a themed call for 'Efficient Study Designs', with a focus on research that "will demonstrate particular design features to allow either more rapid conduct, or lower costs". It is becoming increasingly important to improve the efficiency of clinical trials in order to speed up the overall process of getting the best therapies to patients whilst minimising resources. If a promising treatment emerges whilst a trial in a similar population is ongoing, there would be many advantages to modifying the existing trial by adding the new arm, as long as the statistical considerations are addressed appropriately. Although there is a wealth of literature on adaptive and flexible designs, adding a new therapy into an ongoing trial is rarely discussed, and only a small number of trials were identified to have made this type of adaptation in practice. In this research the considerations necessary to ensure robust statistical validity were identified and addressed, in order to provide guidance so that researchers feel confident in amending ongoing trials to incorporate a new treatment arm.

This research topic is strongly supported by patient advocates and clinicians. It was discussed from the outset at the NCRI CLL Subgroup, in which a patient representative from the NCRI Consumer Liaison Group (CLG) was present, along with research active clinicians. They fully support the research concept as they feel it is in the best interests of patients by enabling promising emerging treatments to reach patients in trials faster than would otherwise be the case. The research was also discussed with Eric Low, Chief Executive of Myeloma UK, a group that informs and supports people affected by myeloma <http://www.myeloma.org.uk/>. Eric is very positive about the research, stating "I believe that these new statistical approaches that allow emerging novel

treatments to be incorporated into existing trials are vital to speed up the evaluation of available treatments, with the aim of enabling patients to receive the best treatments as soon as possible."

7.1.1 Framing the research

The aim of Chapter 2 was to frame the research by assessing the extent of current literature on adding new treatment arms, investigating how well this amendment has been made in practice; and identifying gaps where further research and guidance is needed. This review led to the generation of eight key statistical areas that researchers need to consider when implementing this type of amendment. In some cases it was clear from existing literature how to achieve statistical validity. Other areas concerned the efficiency of the trial design rather than necessarily statistical validity, and whilst these are important to consider, they should be addressed by the trial team at the time of the design of the amendment. In two areas, however, the literature was contradictory or scarce, and these formed the basis for the remainder of this research. Even though full recommendations were not available immediately after this review, a summary of the literature and statistical considerations identified was published as a review article in the Trials journal[6]. It has been cited four times to date in peer reviewed journal articles, and has been quoted in the draft CONSORT Statement extension to multi-arm parallel randomised clinical trials. Following the completion of this research, updated guidance comprehensively detailing the necessary considerations and methodology is now able to be provided, and is summarised in Section 7.2.

7.1.2 Multiple testing adjustment for multiple hypotheses

The first area identified to require further research was that of multiple testing adjustment due to multiple hypotheses being assessed within the same protocol and sharing some control data. In Chapter 3, it is confirmed that there is no consensus in the literature on whether FWER adjustment is necessary for multiple hypotheses. In order to be able to consider the impact of multiple hypotheses on the type I error rate, the need for adjustment was broken down into: the correlation between the test statistics due to shared control data; and the efficiency of sharing a protocol. In Chapter 4, the effect of the correlation on type I error rates was calculated exactly in the case of two and three hypotheses, and it was confirmed that the FWER is lower where the hypotheses share a control group than when they are independent. Whilst this is not a new finding, it is widely misunderstood. A referee's comment on the publication of this work[8] was that this is "an issue that I agree with the authors that, somewhat

surprisingly, has received little attention in the literature". By illustrating the type I error regions for two test statistics based on the bivariate normal distribution with correlation, and using this to quantify the probabilities in each of the error regions, the distribution of the various type I error rates defined in Section 4.2 was able to be visualised. The reason the FWER is reduced where there is shared control data is because the probability of observing more than one error across the hypotheses (FMER) is increased. However, this has rarely been addressed in the literature when considering the need for multiple testing adjustment. It is not clear why this is the case, perhaps because adjusting to control for the inflated probability that multiple experimental treatments will be falsely declared to be superior requires too stringent an adjustment. Multiple errors are a particular concern if more than one superiority claim from within the same protocol could be used to jointly inform a claim of effectiveness, as discussed in Section 4.5, but perhaps this is felt to be unlikely to be adequate to regulators, so protection of the MSFP rate would be unnecessary.

In the case of confirmatory trials, there is general agreement in the literature that where the hypotheses do not share control data, so are independent, no multiplicity adjustment is required. Recent discussions by representatives from the MHRA[116] and EMA[117] on master protocols portray the Biostatistics Working Party working hypothesis that "No multiplicity adjustment is required if the sub-trials are essentially independent trials, each testing a different, independent hypothesis". However, a violation of independence includes that there is an overlap of treatment, for example a common control arm. They say that this is an area of ongoing discussion, but the general consensus is that adjustment is more likely to be necessary in this case. It seems to be a contradiction that FWER control is required because of the shared control data and associated lack of independence, when the shared control data reduces the FWER over what it would have been in independent hypotheses. In discussion on this topic with Michael Proschan (personal communication), his response was that:

> "If you do not adjust for multiplicity, you are likely to have at least one false significant finding (if there are many comparisons). At that point, someone will point out that we cannot trust the other findings because there could be a bad control arm that caused the problem. I agree with you that the same thing could happen whether or not you adjusted for multiplicity, but if you did adjust for multiplicity, you are unlikely to discover any false positives, given that the FWER is controlled at level 0.05. On the other hand, if you didn't adjust for multiplicity, you ARE likely to discover at least one false positive. It only takes one false positive (not 2 or more, as the FMER considers) to create a problem

> of lost confidence. If there is any false positive, people will lose confidence in the findings. For that reason, it is important to prevent even one false positive. The way to do that is to control the FWER at a low rate such as 0.05. Therefore, I would not feel comfortable, in the setting of a single control arm, with doing comparisons each at level 0.05."

This is a very interesting argument, and illustrates that a concern over the FWER where there is shared control data is in fact related to the inflated chance of multiple errors. These points can still be addressed in terms of the FWER and FMER. If there is a concern over 'even one false positive', FWER adjustment is required regardless of whether there is shared control data or not. If the concern is over the conditional probability of a type I error given that another type I error has occurred in the case of a 'bad' control group, as Proschan suggests, this is directly related to the FMER (Section 4.2.3). The argument given is that FWER control is necessary to reduce the probability of multiple errors in the case of a bad control group. Referring back to Table 4-3 in which the error rates after multiplicity adjustment were assessed in the case of two hypotheses, it can be seen that FWER adjustment can reduce the probabilities of multiple errors to some extent. Therefore, adjustment may help to reduce the inflated FMER. However, even with adjustment, this probability is not reduced to that for independent trials, because the error being controlled is not the error that is inflated. In addition, some of the common adjustment methods are less conservative because they allow a higher probability of multiple errors in order to reduce the overall level of adjustment whilst maintaining control over the chance of at least one error. That is, they use the fact that the correlation increases the FMER, and therefore decreases the FWER, to reduce the amount of FWER adjustment. This is counterintuitive. If FWER adjustment is recommended where there is shared control data in an attempt to somewhat control the FMER, to then reduce the level of adjustment to allow more multiple errors is the opposite of what adjustment is trying to achieve.

In the case of two hypotheses this is easy to understand. If there is a 'bad' control group, there will clearly be an increased probability of an error in both hypotheses. If only one of the p-values is less than 5%, FWER control will reduce the probability of success for that hypothesis. However, in two independent trials there are two chances for one of the control groups to be bad, so the overall chance of at least one false-positive finding is higher, and adjustment is not required for this. In the case that both p-values are less than 5%, however, closed-testing adjustment methods have little or no effect on controlling the probability of a false-positive error given the other falsely significant finding. The only rationale for FWER control specifically due to the shared

control data is to somewhat reduce the inflated chance of two errors. For this reason, if FWER adjustment is required, closed testing methods should not be used. The more conservative Bonferroni adjustment performs best out of the available methods for this purpose.

In the case of three or more hypotheses, this becomes more difficult to understand. Table 7-1 summarises the effects of applying the Bonferroni, Hochberg and Dunnett's t adjustment methods on the various error rates, using the example of a four-arm trial with 1:1:1:1 allocation in which the three experimental arms are compared to a shared control group. The two-sided alpha is set at 0.05 for each unadjusted comparison, and therefore the FWER is required to be controlled at 0.05 when adjustment methods are used.

Table 7-1 FWER, FMER and MSFP comparisons for four-arm trials with three hypotheses (α=0.05 for each), a shared control group and even allocation ratio, after applying various multiple testing adjustments

	Independent case	Dependent case, 1:1:1:1 allocation			
		Un-adjusted	Bonferroni	Hochberg	Dunnett's t
Reject H_0 for each individual comparison	0.05	0.05	0.0166	0.0193	0.0188
FWER: Reject at least one H_0	0.1426	0.1254	0.0443	0.0453	0.0499
FMER2: Reject at least two H_0's (in any direction)	0.0073	0.0213	0.0049	0.0093	0.0058
FMER3: Reject all three H_0's (in any direction)	0.0001	0.0032	0.0005	0.0032	0.0007
MSFP2: Reject at least two H_0's in favour of *A*, *B* or *C*	0.0018	0.0107	0.0025	0.0047	0.0029
MSFP3: Reject all three H_0's in favour of A, B and C	0.00002	0.00160	0.00027	0.00160	0.00033

It can be seen that in the case of three hypotheses, the FWER is controlled at 0.05 using all methods, as required. The chance of at least two of the three hypotheses falsely reporting a superior false positive outcome (MSFP2) is 0.18% in the independent case and rises to 1.07% without adjustment. After Bonferroni and Dunnett's t adjustment the MSFP2 is reduced to 0.25% and 0.29% respectively, which are closer to the independent case. With the Hochberg adjustment the MSFP2 is somewhat reduced to 0.47%. The chance of three superior false positives is 0.16% in the unadjusted case and with the Hochberg adjustment, but 0.03% with the other

adjustment methods. Whilst this is higher than in the independent case, in which it is 0.002%, it is still very small. If the concern is that there are at least two superior false-positives out of the set of three that could inform a single claim of effectiveness, FWER adjustment using Bonferroni or Dunnett's t methods are reasonably effective, although it would be preferred to control this probability exactly if required. As the number of hypotheses increases, FWER adjustment becomes more stringent, and it is likely that the probabilities of at least two MSFPs from the set will be controlled, in fact they will probably become smaller than in the independent case. The probabilities of higher numbers of MSFPs will not be well controlled, but these probabilities will become extremely small. In trials with many comparisons, FWER adjustment does reduce the inflated chance of some multiple errors, although it does not do so in a considered way. It may be advantageous to investigate adjustment for the inflated probability of at least two errors declaring the experimental treatments superior (MSFP2), rather than at least one error (FWER) as is done currently, and apply this exactly. This is likely to become less stringent than FWER adjustment where there are more hypotheses. In summary, in the case of three or more hypotheses, FWER adjustment is a simple method to reduce the inflated probabilities of multiple errors to some extent, and therefore could be applied for this purpose, although it is recommended that closed testing methods are avoided.

Multiplicity adjustment for multiple hypotheses is a widely debated topic, but this way of breaking down the various probabilities of errors due to sharing control data has not been previously considered. The findings and associated recommendations challenge current literature and guidance documents that recommend FWER adjustment solely due to shared control data. Discussions are ongoing with the group at the EMA who are working on guidance on the necessity to control the FWER in the case of umbrella trials with shared controls, and with the team drafting a CONSORT Statement extension for multi-arm trials, with the aim to input into the discussions informing these important documents.

7.1.3 Analysis methods when adding an arm

The other area identified to require further research was the methods of analysis following amending an ongoing trial by adding a new treatment arm. In Chapter 5, analysis using adaptive p-value combination methodology was considered and compared to methods where the data are pooled over the stages before and after the amendment. It was determined that unless there is an interim analysis that also informs

a change in the design for the existing hypothesis in the second stage, adaptive analysis methods are not required. Instead, a multivariable analysis including Trial Stage as a covariate is recommended.

The concept of amending a trial by adding an arm with relation to the need for p-value combination methods was discussed with Michael Proschan, because of his co-authorship on a paper that informs the verification of the Conditional Invariance Principle[118]. He agrees with our view (personal communication) that if no adaptations are made other than adding an arm, “even if the decision to add B was based on interim data from A, the comparison of arm A to control restricted to stage 2 data is independent of the comparison of arm A to control in stage 1”. In that case, he believes that p-value combination methods would control the type I error rate, and using a multivariable model with stage as a covariate “should be asymptotically equivalent to the simpler combination of z-scores”. There was a question in the case of continuous endpoint data whether a change in variance in the second stage would not be accounted for in the multivariable model, whilst it would with p-value combination methods. However, he summarises “my opinion is that under reasonable assumptions, you should get essentially the same answer whether you combine independent z-scores or use a model that does the same thing. Including additional covariates to the model also seems asymptotically valid. Therefore, I believe that what you are suggesting will control the type I error rate pretty closely, and is quite reasonable”. Investigating the analysis methods where there is unequal variances across the trial stages is an area identified for further research.

This research has focused on the impact of a stage effect on the p-values and the probabilities of type I and II errors. Also of importance, in accordance with the ICH guideline E9 on statistical principles for clinical trials[51], is the ability to report the treatment effect estimate and associated confidence intervals. If a multivariable analysis is appropriate, these statistics are readily available, and are adjusted for stage and other covariates. However, if it is necessary to use adaptive analysis methods, estimates and confidence intervals must not ignore the adaptive nature of the trial. Methods have been proposed to derive appropriate point estimates and confidence bounds for p-value combination tests (see for example Wassmer and Brannath (2016)[36]), but these methods are not straightforward to apply. An important area for further consideration would be a comparison of the treatment effect estimate and confidence limits using the different analysis methods in the presence of a stage effect to further inform the recommendations. If there is a stage effect, although the

multivariable analysis accounts for this when calculating the p-value to assess the treatment effect, the treatment effect estimate and confidence interval should be interpreted with care. It would be advantageous to assess baseline characteristics of patients by stage, and if there are any apparent differences or if the stage effect is approaching significance within the model, subgroup or sensitivity analyses by stage could also be considered.

An interesting and novel finding from this work is that the order of applying multiplicity adjustment alongside p-value combination methods may be important. The majority of literature on adaptive designs applies multiplicity adjustment within stage prior to combining the p-values, which makes practical sense if there is an interim analysis and adjustment is required at each analysis point. However, in situations where treatment arms are added and stopped, such as platform or multi-arm multi-stage trials, this can have an undesirable effect on the power in the case that the experimental treatments perform differently to one another. Recommendations in these cases are therefore contrary to literature, in that if multiplicity adjustment is necessary for any sets of hypotheses, it should be made at the end of the trial on the final p-values rather than the stagewise p-values.

7.1.4 Practical application

In Chapter 6, the FLAIR trial in Chronic Lymphocytic Leukaemia was described in which a new experimental arm has been successfully added to the existing randomisation, showing that this type of amendment is feasible in practice both statistically and operationally. At the time of designing FLAIR, there was a promising treatment being investigated in a phase II trial, as described in Figure 1.1 in the Introduction, and this was part of the motivation for this research. However, the experimental treatment that was added to FLAIR was not the treatment that was originally planned because the phase II results were not as good as anticipated, and meanwhile venetoclax emerged from development with extremely promising early evidence of efficacy. An advantage of this type of flexible design is that since the amendment does not need to be specified from the outset of the trial, it can be planned as relevant at the time. The amendment is hugely beneficial in terms of resource savings, and allowed the new experimental therapy to be included in a large confirmatory trial in the UK years earlier than would have otherwise been possible. The FLAIR trial has now reached the recruitment target for the original hypothesis, and the IR experimental arm has been dropped from the randomisation. Recruitment remains

ahead of target, and the trial results for both the original and new hypotheses are eagerly awaited with the real potential to influence practice. This trial is an exemplar of how a new experimental arm can be added without compromising the validity of the research for either hypothesis.

7.2 Guidance and recommendations on adding an arm

In Chapter 2, the key statistical considerations when adding a treatment arm to an ongoing trial were identified. Throughout this research, uncertainties have been addressed so that it is now possible to provide guidance to researchers to help to ensure statistical validity and efficiency. A summary of the key considerations along with recommendations are provided in Table 7-2.

Table 7-2 Summary of recommendations on the key statistical considerations when amending an ongoing trial by adding a new treatment arm

Statistical Consideration	Recommendations
FWER control for multiple primary hypotheses	• Requirements for FWER adjustment are the same where an arm is added as they are for a standard multi-arm trial. • The FWER is not inflated due to the shared control group, so adjustment is not required for this purpose. • If the hypotheses contribute to the same claim of effectiveness, FWER adjustment is likely to be required due to the efficiency of asking more questions in the same protocol and having multiple chances for success. • The probability of multiple type I errors is inflated due to the shared control group, but FWER adjustment does not aim to control this. If more than one superior finding is a concern, for example where the hypotheses contribute to the same claim of effectiveness, the MSFP rate needs to be considered. In the case of two hypotheses, a more stringent p-value is necessary and has been proposed assuming concurrent randomisation. In the case of three or more hypotheses, FWER adjustment does reduce the inflated probability of some superior false-positive errors to some extent. However, if multiple errors are a concern, closed testing adjusting methods perform worse and should not be used. Bonferroni or Dunnett's t methods are recommended.
Methods of analysis over trial stages before and after the design amendment	• A multivariable analysis with Trial Stage as a covariate is recommended, to account for a potential stage effect caused by a shift in the population due to the amendment. • There must be a strong assumption of no interaction between the treatment effect and trial stage, which could be caused by a potential population shift. • P-value combination methods are only necessary if there is an interim analysis to adapt the design for existing hypotheses alongside adding the new arm. • If multiple testing adjustment is required, it should be made on the final p-values for each hypotheses, regardless of how much of the control data is overlapping. Adjustment within stage prior to combining p-values is not recommended.

Concurrent use of control data	• Control data for participants who were not randomised concurrently must not be used in primary analyses. • The power to test for heterogeneity across the trial stages is likely to be low, and therefore a finding of no significant heterogeneity is not meaningful.
Power for primary hypotheses	• Every primary hypothesis in the trial must be adequately powered. Patients randomised to the new arm must be in addition to the sample size for the original trial arms. Adding a new hypothesis will therefore increase the total size and associated resources of the trial.
Allocation ratio	• As for any randomised trial, the allocation ratio should be chosen to maximise efficiency whilst considering patient acceptability and other relevant trial aims as appropriate. • The allocation ratio does not need to be the same for all hypotheses, as long as each hypothesis has adequate power when compared to concurrent controls. • It is permissible to change the allocation ratio for an existing hypothesis at the time of the amendment, as long as the power remains adequate. This needs to be accounted for in the analysis[53].
End of recruitment for existing and new arms	• Depending on the allocation ratio and design assumptions, it is likely that the new hypothesis will need to recruit for longer than the existing hypotheses in order to have appropriate power. • When the original randomisation is complete, the experimental arms can be dropped. This would lead to another trial stage requiring a randomisation amendment, and should be planned for at the design of the amendment. The new trial stage will also need to be accounted for in the analysis, as described previously.
Change to the control group	• If the standard of care therapy is superseded during the length of the trial, the control therapy may become inferior to standard practice and unethical. At the design stage for the trial amendment, it should be considered whether this is likely, and if so this should be pre-empted as far as possible. For example, if an existing experimental trial therapy has the potential to become the standard of care, especially if the hypothesis is due to report before the end of recruitment to the amendment, the new treatment could be powered to be concurrently compared to both the existing and potential new control arms, if feasible. • If the new treatment is required to be superior to both the existing and new control therapies to be deemed a success, this does not raise multiplicity concerns for the FWER.

Amendment clearly mentioned in results publications	• Publications must transparently include the full trial design, detailing all treatments, comparisons and adaptations to the design, even if they are not the focus of that manuscript.

7.3 Limitations and extensions

This research primarily focused on multiplicity adjustment and analysis methods when adapting an ongoing trial by adding a new experimental arm, since it is essential that these are appropriate in order to ensure statistical validity. Other statistical considerations such as use of non-concurrent control data and optimal allocation ratios were touched upon, although were not investigated in detail. It may be possible that data from the control group collected prior to the addition of the new arm could be used appropriately to improve the power of the trial. The use of historical control data is of relevance more widely than just to the case of adding arms, and has previously been discussed with relation to frequentist confirmatory trials in the methodological literature. Whilst it could offer potentially large savings to a trial, there is also the risk of substantial bias[119], and it is less likely to be acceptable to regulators in a confirmatory setting. The allocation ratio could be calculated exactly to minimise patient numbers to the trial overall, but it has been shown that the savings are likely to be small[120], and the additional complexity of adjusting allocation ratios in the case of adding and dropping trial arms is unlikely to be worthwhile. Therefore it was decided not to focus on these issues.

It was determined from the outset that this research would only consider frequentist methodology, however Bayesian methodology could be advantageous in this setting. It may enable the incorporation of prior control data, and there is a different approach to decision rules that are not based on type I or II error rates. Bayesian inference in adaptive designs has been mainly limited to early-phase trials and may not be acceptable to regulators in a confirmatory trial[121], however it would be interesting to investigate in the case of adding an arm mid-trial.

As discussed in Chapter 5, in rolling designs where the hypotheses do not follow the same timelines, planned analyses for one hypothesis may impact on another. If an analysis for an original hypothesis is planned prior to the close of recruitment to the

new hypothesis, the potential implications of this need to be carefully considered. As well as potentially affecting the control treatment going forward, some data from the control arms for the new hypothesis could be reported. The implications of this have not been fully considered, and this is an area that would benefit from further investigation and discussion with experts.

In this thesis, the investigation into analysis methods has been based on simulations of an example where a new treatment arm is added to a trial with an existing control and experimental arm, in which both hypotheses assess superiority in terms of a survival endpoint. There is no reason to believe the findings would not extrapolate to trials with more hypotheses, in which more than one new arm is added either at the same time or at different times, and where there are different endpoints and objectives. However, in order to confirm the robustness and generalisability of the findings, it is planned to assess the analysis methods using simulations based on different scenarios. It is also planned to further investigate multiplicity adjustment alongside p-value combination in adaptive designs more generally following the interesting finding that the order in which they are applied is important.

Rather than assessing superiority of an new treatment, clinical trials can also assess non-inferiority, to test whether the experimental therapy is not unacceptably less efficacious than the current standard. This might be because it is thought that the treatment is as good in terms of efficacy, but has other benefits such as reduced toxicity. In this case, non-inferiority is not assessed using a p-value, but by comparing the lower limit of the confidence interval (CI) to a pre-determined non-inferiority margin. If the CI contains the non-inferiority margin, there is an unacceptable chance that the experimental treatment is inferior, and non-inferiority cannot be declared. If the entire CI lies above the non-inferiority margin, non-inferiority can be declared. In the case of multiple testing adjustment for multiple hypotheses, the width of the confidence interval can be adjusted in line with the chosen alpha adjustment method. For example, using a Bonferroni adjustment with two hypotheses and an overall significance level of 5%, the type I error rate is split so that significance level for each hypothesis is adjusted to 2.5%, and equivalent 97.5% CIs are calculated. This will widen the CIs, reducing the probability of falsely declaring non-inferiority. The principles that relate to the sharing of control data will apply to non-inferiority trials in the same way that they apply to superiority trials, but due to the differences in hypothesis testing methodology, this would benefit from further exploration. In terms of the methods of analysis over trial stages due to the amendment, the general recommendations from Figure 5-3 still hold

in the non-inferiority case. If the existing hypothesis is not adapted, a pooled multivariable analysis is appropriate, and CIs are easy to calculate. If adaptive analysis methods are necessary, however, it needs to be determined how to appropriately design and analyse an adaptive non-inferiority trial at the design stages for the amendment.

In order to influence practice, the recommendations from this research need to be discussed with regulators and the wider statistical community. The chapters on the review of statistical considerations when adding an arm and on multiplicity adjustment have been published in peer reviewed journals and presented at national and international conferences, seminars and workshops. In addition, the manuscript on multiplicity adjustment has been shared with the group updating the CONSORT Statement, the Biostatistics Working Party discussing guidance on multiplicity issues for master protocols, and the consultation on the EMA draft guideline on multiplicity issues in clinical trials. Initial feedback suggests that the research has been well received and is interesting and useful, and it is hoped that further discussions will be held leading to an impact on practice. The later work on analysis methods has yet to be shared, but it is planned to publish this work as part of an overall guidance document for researchers, funders or regulators with an interest in amending an ongoing trial by adding new experimental arms, as well as to submit abstracts to conferences and workshops.

7.4 Reflection

In this thesis, it has been described how an amendment to add an experimental arm into an ongoing trial can be made without compromising the statistical validity or integrity of the trial. The necessary statistical considerations have been identified, and the most contentious areas of multiple testing adjustment due to multiple hypotheses and analysis methods have been investigated. Some of the findings have been surprising, in particular the effect of multiplicity adjustment for multiple hypotheses with shared control data on the probabilities of type I errors; and the application of multiplicity adjustment alongside p-value combination analysis methods. Both of these areas are of relevance more widely than just to the case of adding arms, and this research has progressed and contradicted some of the existing guidance on these topics.

The recommendations to ensure statistical validity when adding an arm, particularly concerning multiplicity adjustment and analysis methods, are straightforward to implement, meaning that this type of amendment is relatively simple compared to adaptive deigns based on internal analyses. A new treatment arm can be added at any stage of the trial with efficiencies gained even if recruitment is close to completion, and it has been shown using a real exemplar that this can be hugely advantageous without compromising the ability to test the existing or new hypotheses. There is no reason why new therapies cannot be added to ongoing trials on a rolling basis using the same guidance and considerations. It is hoped that the findings from this research will encourage experimental arms to be added to confirmatory trials more regularly in order to improve the efficiency of the evaluation of promising emerging therapies.

List of References

1. Parmar MKB, Carpenter J and Sydes MR. More multiarm randomised trials of superiority are needed. *The Lancet*. 2014; 384: 283-4.

2. Quaresma M, Coleman MP and Rachet B. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. *The Lancet*. 2015; 385: 1206-18.

3. Cancer Research UK. Cancer Statistics for the UK. http://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk: CRUK, 2014.

4. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M and Pinheiro J. Adaptive Designs in Clinical Drug Development - An Executive Summary of the PhRMA Working Group. *Journal of Biopharmaceutical Statistics*. 2006; 16: 275-83.

5. Bauer P, Bretz F, Dragalin V, König F and Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine*. 2016; 35: 325-47.

6. Cohen DR, Todd S, Gregory WM and Brown JM. Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. *Trials*. 2015; 16: 179.

7. Connor JT, DeMichele A and Wittes J. University of Pennsylvania ninth annual conference on statistical issues in clinical trials: Where are we with adaptive clinical trial designs? (afternoon panel discussion). *Clinical Trials*. 2017; 14: 470-82.

8. Howard DR, Brown JM, Todd S and Gregory WM. Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. *Statistical Methods in Medical Research*. 2018; 27: 1513-30.

9. Cook JA, Julious SA, Sones W, et al. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ*. 2018; 363.

10. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B (Methodological)*. 1972; 34: 187-220.

11. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine*. 1982; 1: 121-9.

12. Woodcock J and LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. *New England Journal of Medicine*. 2017; 377: 62-70.

13. Saville BR and Berry SM. Efficiencies of platform clinical trials: A vision of the future. *Clinical Trials*. 2016; 13: 358-66.

14. Phillips AJ and Keene ON. Adaptive designs for pivotal trials: discussion points from the PSI Adaptive Design Expert Group. *Pharmaceutical Statistics*. 2006; 5: 61-6.

15. Hommel G. Adaptive Modifications of Hypotheses After an Interim Analysis. *Biometrical Journal*. 2001; 43: 581-9.

16. van Leth F, Phanuphak P, Ruxrungtham K, et al. Comparison of first-line antiretroviral therapy with regimens including nevirapine, efavirenz, or both drugs, plus stavudine and lamivudine: a randomised open-label trial, the 2NN Study. *Lancet*. 2004; 363: 1253-63.

17. Gatsonis C, Kass RE, Carlin B and Carriquiry A. *Case studies in Bayesian statistics*. New York: Springer-Verlag, 2001.

18. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C and Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine*. 2005; 24: 3697-714.

19. Bauer P. Adaptive designs: Looking for a needle in the haystack - A new challenge in medical research. *Statistics in Medicine*. 2008; 27: 1565-80.

20. Elm JJ, Palesch YY, Koch GG, Hinson V, Ravina B and Zhao W. Flexible analytical methods for adding a treatment arm mid-study to an ongoing clinical trial. *Journal of Biopharmaceutical Statistics*. 2012; 22: 758-72.

21. Sydes MR, Parmar MKB, Mason MD, et al. Flexible trial design in practice - stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomized controlled trial. *Trials*. 2012; 13: 168.

22. Parmar MK, Sydes MR, Cafferty FH, et al. Testing many treatments within a single protocol over 10 years at MRC Clinical Trials Unit at UCL: Multi-arm, multi-stage platform, umbrella and basket protocols. *Clinical Trials*. 2017; 14: 451-61.

23. James ND, Sydes MR, Clarke NW, et al. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *The Lancet*. 2016; 387: 1163-77.

24. Wason J, Magirr D, Law M and Jaki T. Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*. 2012.

25. Pong A and Chow S-C. *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*. New York: Chapman and Hall/CRC Press, Taylor & Francis Group, 2010.

26. Posch M, Bauer P and Brannath W. Flexible Designs. *Wiley Encyclopedia of Clinical Trials*. New York: John Wiley & Sons, Inc., 2007.

27. Hills RK and Burnett AK. Applicability of a “Pick a Winner” trial design to acute myeloid leukemia. *Blood*. 2011; 118: 2389-94.

28. Gaunt P, Mehanna H and Yap C. The design of a multi-arm multi-stage (MAMS) phase III randomised controlled trial comparing alternative regimens for escalating (COMPARE) treatment of intermediate and high-risk oropharyngeal cancer with reflections on the complications of introducing a new experimental ARM. *Trials*. 2015; 16: O16.

29. Burnett AK, Hills RK, Hunter AE, et al. The addition of gemtuzumab ozogamicin to low-dose Ara-C improves remission rate but does not significantly prolong survival in older patients with acute myeloid leukaemia: results from the LRF AML14 and NCRI AML16 pick-a-winner comparison. *Leukemia*. 2012; 27: 75.

30. Lieberman JA, Stroup TS, McEvoy JP, et al. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *New England Journal of Medicine*. 2005; 353: 1209-23.

31. Marson AG, Al-Kharusi AM, Alwaidh M, et al. The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. *The Lancet*. 2007; 369: 1000-15.

32. Goldberg RM, Sargent DJ, Morton RF, et al. A Randomized Controlled Trial of Fluorouracil Plus Leucovorin, Irinotecan, and Oxaliplatin Combinations in Patients With Previously Untreated Metastatic Colorectal Cancer. *Journal of Clinical Oncology*. 2004; 22: 23-30.

33. Goldberg RM, Sargent DJ, Morton RF, et al. NCCTG Study N9741: Leveraging Learning from an NCI Cooperative Group Phase III Trial. *The Oncologist*. 2009; 14: 970-8.

34. Alberts SR, Sargent DJ, Nair S, et al. Effect of Oxaliplatin, Fluorouracil, and Leucovorin With or Without Cetuximab on Survival Among Patients With Resected Stage III Colon Cancer: A Randomized Trial. *JAMA : the journal of the American Medical Association*. 2012; 307: 1383-93.

35. Jawahar MS, Banurekha VV, Paramasivan CN, et al. Randomized Clinical Trial of Thrice-Weekly 4-Month Moxifloxacin or Gatifloxacin Containing Regimens in the Treatment of New Sputum Positive Pulmonary Tuberculosis Patients. *PLOS ONE*. 2013; 8: e67030.

36. Wassmer G and Brannath W. *Group sequential and confirmatory adaptive designs in clinical trials*. Switzerland: Springer, 2016.

37. FDA. Draft Guidance on Adaptive Design Clinical Trials for Drugs and Biologics. Rockville, MD2010.

38. FDA. Adaptive Designs for Medical Device Clinical Studies. https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf: FDA, 2016.

39. CHMP (Committee for Medicinal Products for Human Use). Reflection Paper on Methodological Issues in Confirmatory Clinical trials with Flexible Design and Analysis Plan. London: EMEA (European Medicines Agency), 2006.

40. CHMP (Committee for Medicinal Products for Human Use). Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design. London: EMEA (European Medicines Agency), 2007.

41. Miller E, Gallo P, He W, et al. DIA's Adaptive Design Scientific Working Group (ADSWG):Best Practices Case Studies for "Less Well-understood" Adaptive Designs. *Therapeutic Innovation & Regulatory Science*. 2017; 51: 77-88.

42. Coffey CS, Levin B, Clark C, et al. Overview, hurdles, and future work in adaptive designs: perspectives from a National Institutes of Health-funded workshop. *Clinical Trials*. 2012; 9: 671-80.

43. Vandemeulebroecke M. Group sequential and adaptive designs - a review of basic concepts and points of discussion. *Biometrical Journal*. 2008; 50: 541-57.

44. Bauer P and Brannath W. The advantages and disadvantages of adaptive designs for clinical trials. *Drug Discovery Today*. 2004; 9: 351-7.

45. Chow SC and Chang M. Adaptive design methods in clinical trials - a review. *Orphanet Journal Of Rare Diseases*. 2008; 3: 11.

46. Wang S-J. Perspectives on the Use of Adaptive Designs in Clinical Trials. Part I. Statistical Considerations and Issues. *Journal of Biopharmaceutical Statistics*. 2010; 20: 1090-7.

47. Benda N, Brannath W, Bretz F, et al. Perspectives on the Use of Adaptive Designs in Clinical Trials. Part II. Panel Discussion. *Journal of Biopharmaceutical Statistics*. 2010; 20: 1098-112.

48. Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*. 2018; 16: 29.

49. Freidlin B, Korn EL, Gray R and Martin A. Multi-Arm Clinical Trials of New Agents: Some Design Considerations. *Clinical Cancer Research*. 2008; 14: 4368-71.

50. CPMP (Committee for Proprietary Medicinal Products). Point to Consider on multiplicity issues in clinical trials. EMEA (European Medicines Agency), 2002.

51. ICH (International Conference on Harmonisation). Statistical Principals for Clinical Trials E9. 1998.

52. Bauer P and Kohne K. Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics*. 1994; 50: 1029-41.

53. Altman DG. Avoiding bias in trials in which allocation ratio is varied. *Journal of the Royal Society of Medicine*. 2018; 111: 143-4.

54. Dunnett CW. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*. 1955; 50: 1096-121.

55. Schulz KF, Altman DG and Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010; 340.

56. EMA (European Medicines Agency). Guideline on multiplicity issues in clinical trials. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224998.pdf: European Medicines Agency, 2017.

57. Phillips A, Fletcher C, Atkinson G, et al. Multiplicity: discussion points from the Statisticians in the Pharmaceutical Industry multiplicity expert group. *Pharmaceutical Statistics*. 2013; 12: 255-9.

58. Dmitrienko A, Tamhane AC and Bretz F. *Multiple Testing Problems in Pharmaceutical Statistics*. CRC Press, 2009.

59. Wason JM, Stecher L and Mander A. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*. 2014; 15: 364.

60. Proschan MA, Follmann DA and Geller NL. Monitoring multi-armed trials. *Statistics in Medicine*. 1994; 13: 1441-52.

61. Proschan MA. A multiple comparison procedure for three- and four-armed controlled clinical trials. *Statistics in Medicine*. 1999; 18: 787-98.

62. Proschan MA and Waclawiw MA. Practical Guidelines for Multiplicity Adjustment in Clinical Trials. *Controlled Clinical Trials*. 2000; 21: 527-39.

63. Proschan M and Follmann D. Multiple comparisons with control in a single experiment ver. *The American Statistician*. 1995; 49: 144.

64. Bender R and Lange S. Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology*. 2001; 54: 343-9.

65. Westfall P and Bretz F. Multiplicity in Clinical Trials. *Encyclopedia of Biopharmaceutical Statistics, Third Edition*. Taylor & Francis, 2010, p. 889-96.

66. Cook RJ and Farewell VT. Multiplicity Considerations in the Design and Analysis of Clinical Trials. *Journal of the Royal Statistical Society Series A (Statistics in Society)*. 1996; 159: 93-110.

67. Rothman KJ. No Adjustments Are Needed for Multiple Comparisons. *Epidemiology*. 1990; 1: 43-6.

68. O'Brien PC. The Appropriateness of Analysis of Variance and Multiple-Comparison Procedures. *Biometrics*. 1983; 39: 787-8.

69. Hung HMJ and Wang S-J. Challenges to multiple testing in clinical trials. *Biometrical Journal*. 2010; 52: 747-56.

70. Marcus R, Peritz E and Gabriel KR. On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika*. 1976; 63: 655-60.

71. Dunnett CW and Tamhane AC. A Step-Up Multiple Test Procedure. *Journal of the American Statistical Association*. 1992; 87: 162-70.

72. Fernandes N and Stone A. Multiplicity adjustments in trials with two correlated comparisons of interest. *Statistical Methods in Medical Research*. 2011; 20: 579-94.

73. Kaplan R, Maughan T, Crook A, et al. Evaluating Many Treatments and Biomarkers in Oncology: A New Design. *Journal of Clinical Oncology*. 2013; 31: 4562-8.

74. Herbst RS, Gandara DR, Hirsch FR, et al. Lung Master Protocol (Lung-MAP)—A Biomarker-Driven Protocol for Accelerating Development of Therapies for Squamous Cell Lung Cancer: SWOG S1400. *American Association for Cancer Research*. 2015; 21: 1514-24.

75. Senn S and Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*. 2007; 6: 161-70.

76. Senn SS. *Statistical issues in drug development*. John Wiley & Sons, 1997.

77. Piccart-Gebhart M, Holmes E, Baselga J, et al. Adjuvant Lapatinib and Trastuzumab for Early Human Epidermal Growth Factor Receptor 2–Positive Breast Cancer: Results From the Randomized Phase III Adjuvant Lapatinib and/or Trastuzumab Treatment Optimization Trial. *Journal of Clinical Oncology*. 2016; 34: 1034-42.

78. Maughan TS, Adams RA, Smith CG, et al. Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *The Lancet*. 377: 2103-14.

79. Follmann DA, Proschan MA and Geller NL. Monitoring Pairwise Comparisons in Multi-Armed Clinical Trials. *Biometrics*. 1994; 50: 325-36.

80. Fisher LD. One Large, Well-Designed, Multicenter Study as an Alternative to the Usual FDA Paradigm. *Drug Information Journal*. 1999; 33: 265-71.

81. Shun Z, Chi E, Durrleman S and Fisher L. Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger vs two smaller pivotal studies Statistics in Medicine 2005; 24:1619–1637. *Statistics in Medicine*. 2005; 24: 1652-6.

82. Julious SA and Campbell MJ. Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. *Statistics in Medicine*. 2012; 31: 2904-36.

83. Abdi H. Bonferroni and Sidak corrections for multiple comparisons. In: Salkind NJ, (ed.). *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage, 2007, p. 103-7.

84. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*. 1979: 65-70.

85. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988; 75: 800-2.

86. FDA. Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products. Rockville, MD1998.

87. Westfall P, Tobias R, Rom D, Wolfinger R and Hochberg Y. *Multiple comparisons and multiple tests using the SAS system*. Cary, NC: SAS Institute Inc., 1999.

88. Jaki T and Parry A. Why are two mistakes not worse than one? A proposal for controlling the expected number of false claims. *Pharmaceutical Statistics*. 2016.

89. Maurer W, Branson M and Posch M. Adaptive Designs and Confirmatory Hypothesis Testing. In: Dmitrienko A, Tamhane AC and Bretz F, (eds.). *Multiple Testing Problems in Pharmaceutical Statistics*. Boca Raton: CRC Press, 2010, p. 193 - 237.

90. Brannath W, Posch M and Bauer P. Recursive Combination Tests. *Journal of the American Statistical Association*. 2002; 97: 236-44.

91. Fisher RA. *Statistical methods for research workers*. Oliver and Boyd, 1932.

92. Lehmacher W and Wassmer G. Adaptive Sample Size Calculations in Group Sequential Trials. *Biometrics*. 1999; 55: 1286-90.

93. Whitehead A. Dealing with Non-Standard Data Sets. *Meta-Analysis Of Controlled Clinical Trials*. John Wiley & Sons, Ltd, 2003, p. 215-40.

94. Becker BJ. Combining Significance Levels. In: Cooper H and Hedges LV, (eds.). *The Handbook of Research Synthesis* New York: Russell Sage Foundation, 1994.

95. Jenkins M, Stone A and Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints†. *Pharmaceutical Statistics*. 2011; 10: 347-56.

96. Gallo P and Chuang-Stein C. What should be the role of homogeneity testing in adaptive trials? *Pharmaceutical Statistics*. 2009; 8: 1-4.

97. Bauer P and Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat Med*. 1999; 18: 1833-48.

98. Bretz F, Schmidli H, König F, Racine A and Maurer W. Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: General Concepts. *Biometrical Journal*. 2006; 48: 623-34.

99. Bretz F, Koenig F, Brannath W, Glimm E and Posch M. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*. 2009; 28: 1181-217.

100. Friede T, Parsons N, Stallard N, et al. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine*. 2011; 30: 1528-40.

101. Kieser M, Bauer P and Lehmacher W. Inference on Multiple Endpoints in Clinical Trials with Adaptive Interim Analyses. *Biometrical Journal*. 1999; 41: 261-77.

102. Collett L, Howard DR, Munir T, et al. Assessment of ibrutinib plus rituximab in front-line CLL (FLAIR trial): study protocol for a phase III randomised controlled trial. *Trials*. 2017; 18: 387.

103. Haematological Malignancy Research Network (HMRN). Incidence and survival for haematological malignancies. https://www.hmrn.org/statistics: HMRN, 2018.

104. Hallek M, Fischer K, Fingerle-Rowson G, et al. Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *The Lancet*. 2010; 376: 1164-74.

105. Rawstron AC, Fazi C, Agathangelidis A, et al. A complementary role of multiparameter flow cytometry and high-throughput sequencing for minimal residual disease detection in chronic lymphocytic leukemia: an European Research Initiative on CLL study. *Leukemia*. 2015; 30: 929.

106. Bloodwise. Trials Acceleration Programme (TAP). https://bloodwise.org.uk/research/clinical-trials/tap: Bloodwise, 2018.

107. Ma S, Brander DM, Seymour JF, et al. Deep and Durable Responses Following Venetoclax (ABT-199 / GDC-0199) Combined with Rituximab in Patients with Relapsed/Refractory Chronic Lymphocytic Leukemia: Results from a Phase 1b Study. *Blood*. 2015; 126: 830.

108. Portell CA, Axelrod M, Brett LK, et al. Synergistic Cytotoxicity of Ibrutinib and the BCL2 Antagonist, ABT-199(GDC-0199) in Mantle Cell Lymphoma (MCL) and Chronic Lymphocytic Leukemia (CLL): Molecular Analysis Reveals Mechanisms of Target Interactions. *Blood*. 2014; 124: 509.

109. Deng J, Isik E, Fernandes SM, Brown JR, Letai A and Davids MS. Ibrutinib Therapy Increases BCL-2 Dependence and Enhances Sensitivity to Venetoclax in CLL. *Blood*. 2015; 126: 490.

110. Burger JA, Tedeschi A, Barr PM, et al. Ibrutinib as Initial Therapy for Patients with Chronic Lymphocytic Leukemia. *New England Journal of Medicine*. 2015; 373: 2425-37.

111. Burger JA, Sivina M, Ferrajoli A, et al. Randomized Trial of Ibrutinib Versus Ibrutinib Plus Rituximab (Ib+R) in Patients with Chronic Lymphocytic Leukemia (CLL). *Blood*. 2017; 130: 427.

112. O'Brien PC and Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35: 549-56.

113. Byrd JC, Brown JR, O'Brien S, et al. Ibrutinib versus Ofatumumab in Previously Treated Chronic Lymphoid Leukemia. *New England Journal of Medicine*. 2014; 371: 213-23.

114. NICE (National Institute for Health and Care Excellence). Ibrutinib for previously treated chronic lymphocytic leukaemia and untreated chronic lymphocytic leukaemia with 17p deletion or TP53 mutation. https://www.nice.org.uk/guidance/TA429: NICE, 2017.

115. BBC News. Have they found a cure for our cancer? http://www.bbc.co.uk/news/stories-42920045: BBC, 2018.

116. Saperia J. Basket, umbrella and platform trials: a regulatory perspective. *PSI Webinar: Basket, Umbrella and Platform Trials - Experiences and Practical Considerations*. https://www.psiweb.org/docs/default-source/default-document-library/julia-saperia-presentation-slides.pdf?sfvrsn=11a9dedb_0, 2018.

117. Tanniou J. Regulatory Hot Topics. *PSI Conference 2018*. Amsterdam, https://www.psiweb.org/docs/default-source/default-document-library/presentation-slides92e0bbff3ad665b3a176ff00001f6b97.pdf?sfvrsn=4936dedb_0, 2018.

118. Liu Q, Proschan MA and Pledger GW. A Unified Theory of Two-Stage Adaptive Designs. *Journal of the American Statistical Association*. 2002; 97: 1034-41.

119. Cuffe RL. The inclusion of historical control data may reduce the power of a confirmatory study. *Statistics in Medicine*. 2011; 30: 1329-38.

120. Wassmer G. On Sample Size Determination in Multi-Armed Confirmatory Adaptive Designs. *Journal of Biopharmaceutical Statistics*. 2011; 21: 802-17.

121. Chow S-C. Adaptive Clinical Trial Design. *Annual Review of Medicine*. 2014; 65: 405-15.

Appendices

Appendix A

Search strategy for the addition of a new treatment

The following search terms were used to identify relevant literature. The search strategy was piloted and agreed using Medline, and adapted as necessary for the other databases.

A.1 MEDLINE

1. Research Design/

2. exp Clinical Trials as Topic/

3. 1 or 2

4. ((adaptive adj3 design*) or (adaptive adj3 method*) or (adaptive adj3 trial*)).mp.

5. ((flexible adj3 design*) or (flexible adj3 method*) or (flexible adj3 trial*)).mp.

6. ((multi?stage adj3 design) or (multi?stage adj3 method*) or (multi?stage adj3 trial*)).mp.

7. ((platform adj3 design*) or (platform adj3 method*) or (platform adj3 trial*)).mp.

8. 4 or 5 or 6 or 7

9. ((adding or additional or incorporat* or extra) adj4 (arm* or treatment* or group* or therap* or randomi* or hypothes*)).mp.

10. 3 and 8 and 9

Key:

/ = subject heading (MeSH term)

exp = explode MeSH term topic to include sub-branches

.mp = (title, abstract, subject heading, heading word, drug trade name, original title, device manufacturer, drug manufacturer name)

* = truncation

? = single character or no character

Adjn = within n words either side

A.2 EMBASE

Includes conference proceedings as well as journals. Note that the MESH term ‘research design’ is mapped to ‘methodology’ on the EMBASE database. On MEDLINE ‘methods’ includes both ‘observation’ and ‘research design’.

1. Methodology/

2. exp "Clinical Trial (Topic)"/

3. 1 or 2

4. ((adaptive adj3 design*) or (adaptive adj3 method*) or (adaptive adj3 trial*)).mp.

5. ((flexible adj3 design*) or (flexible adj3 method*) or (flexible adj3 trial*)).mp.

6. ((multi?stage adj3 design*) or (multi?stage adj3 method*) or (multi?stage adj3 trial*)).mp.

7 ((platform adj3 design*) or (platform adj3 method*) or (platform adj3 trial*)).mp.

8 4 or 5 or 6 or 7

9 ((adding or additional or incorporat* or extra) adj4 (arm* or treatment* or group* or therap* or randomi* or hypothes*)).mp.

10 3 and 8 and 9

Key:

/ = MeSH term

exp = explode MeSH term topic to include sub-branches

.mp = (title, abstract, subject heading, heading word, drug trade name, original title, device manufacturer, drug manufacturer name)

* = truncation

? = single character or no character

Adjn = within n words either side

A.3 Science Citation Index (Web of Science)

No MESH terms available. Search restricted to the ‘statistics probability’ category, with an investigation carried out as to whether other relevant literature was likely to have been missed.

Topic=(adaptive near/3 design* or adaptive near/3 method* or adaptive near/3 trial* or flexible near/3 design* or flexible near/3 method* or flexible near/3 trial* or multi$stage near/3 design* or multi$stage near/3 method* or multi$stage near/3 trial* or platform near/3 design* or platform near/3 method* or platform near/3 trial*)

AND

Topic=((adding or additional or incorporat* or extra) near/4 (arm* or treatment* or group* or therap* or randomi* or hypothes*))

Key:

* = truncation

$ = single character or no character

Near/n = within n words either side

A.4 Cochrane Library

1. MeSH descriptor: [Research Design] explode all trees

2. MeSH descriptor: [Clinical Trials as Topic] explode all trees

3. #1 or #2

4. (adaptive near/3 design* or adaptive near/3 method* or adaptive near/3 trial* or flexible near/3 design* or flexible near/3 method* or flexible near/3 trial* or multi?stage near/3 design* or multi?stage near/3 method* or multi?stage near/3 trial*):ti,ab,kw

AND

(adding or additional or incorporat* or extra) near/4 (arm* or treatment* or group* or therap* or randomi* or hypothes*):ti,ab,kw (Word variations have been searched)

5. #3 and #4

Key:

* = 1 or more characters

? = single character

near /= within n words either side

A.5 ProQuest

Search in Books, Conference Papers & Proceedings, Dissertations & Theses and Scholarly Journals.

su("research methodology" OR "statistical methods" OR "clinical trials")

AND

noft(adaptive NEAR/3 design* OR adaptive NEAR/3 method* OR adaptive NEAR/3 trial* OR flexible NEAR/3 design* OR flexible NEAR/3 method* OR flexible NEAR/3 trial* OR multi*stage NEAR/3 design* OR multi*stage NEAR/3 method* OR multi*stage NEAR/3 trial* OR platform NEAR/3 design* OR platform NEAR/3 method* OR platform NEAR/3 trial*)

AND

noft((adding OR additional OR incorporat* OR extra) NEAR/4 (arm* OR treatment* OR group* OR therap* OR randomi* OR hypothes*))

Key:

su = subject headings

noft = ALL fields, no full text

* = any number of characters or no character

near /= within n words either side

Appendix B

Search Criteria for Guidance and Summary Documents on Adaptive or Flexible Designs

B.1 MEDLINE

1. Research Design/

2. exp Clinical Trials as Topic/

3. 1 or 2

4. ((adaptive adj3 design*) or (adaptive adj3 method*) or (adaptive adj3 trial*)).mp.

5. ((flexible adj3 design*) or (flexible adj3 method*) or (flexible adj3 trial*)).mp.

6. ((platform adj3 design*) or (platform adj3 method*) or (platform adj3 trial*)).mp.

7. 4 or 5 or 6

8. (guidance or discussion or reflect* or summary or (work* adj1 group)).mp.

9. 3 and 7 and 8

Appendix C

R code to calculate type I error rates and critical values assuming a multivariate normal distribution

C.1 R code to calculate the probabilities for the rejection regions based on two correlated test statistics, assuming a bivariate normal distribution.

```
#Install library first use
setInternet2(TRUE)
install.packages("mvtnorm")
library(mvtnorm)

#Bivariate normal case (2 experimental arms)

#set correlation
corr <- 0.5

#correlation matrix
corrmat <- matrix(c(1,corr,corr,1),ncol=2,byrow=TRUE)

#critical value
cval <- qnorm(0.975)

#Exactly 1 error (calculate probabilities for the edges
#excluding the corners of the square)

# Left hand side
leftside <- pmvnorm(lower=c(-Inf,-cval), upper=c(-cval,cval),
corr = corrmat )

# Right hand side
rightside <- pmvnorm(lower=c(cval,-cval), upper=c(Inf,cval),
corr = corrmat )

# Top edge
topside <- pmvnorm(lower=c(-cval,cval), upper=c(cval,Inf), corr
= corrmat )

# Bottom edge
bottomside <- pmvnorm(lower=c(-cval,-Inf), upper=c(cval,-cval),
corr = corrmat )
```

```
# Total chance of exactly 1 error
oneonly=leftside+rightside+topside+bottomside

#Exactly 2 errors (calculate probabilities in each of the 4
corners of the square)

# Lower left corner
lowleft <- pmvnorm(lower=c(-Inf,-Inf), upper=c(-cval,-cval),
corr = corrmat )

# Lower right corner
lowright <- pmvnorm(lower=c(cval,-Inf), upper=c(Inf,-cval), corr
= corrmat )

# Upper Left corner
upleft <- pmvnorm(lower=c(-Inf,cval), upper=c(-cval,Inf), corr =
corrmat )

# Upper Right corner
upright <- pmvnorm(lower=c(cval,cval), upper=c(Inf,Inf), corr =
corrmat )

# Total chance of exactly 2 errors
twoonly <- lowleft+lowright+upleft+upright

# FWER
FWER <- oneonly+twoonly

# Probability of any two errors (FMER)
FMER <- twoonly

# MSFP probability of two superior false positives
MSFP <- upright

#Output results
corrmat
FWER
FMER
MSFP
```

C.2 R code to calculate the probabilities for the rejection regions based on three correlated test statistics, assuming a trivariate normal distribution.

```
# Install library first use
setInternet2(TRUE)
install.packages("mvtnorm")
library(mvtnorm)

# Trivariate normal case (3 experimental arms)

#set correlation
corr <- 0.5

#correlation matrix
corrmat <-
matrix(c(1,corr,corr,corr,1,corr,corr,corr,1),ncol=3,byrow=TRUE)

#critical value
cval <- qnorm(0.975)

#Exactly 1 error (illustrated by the 6 side face of a cube minus
the upper and lower 5% around the edges)

Oneonly1 <- pmvnorm(lower=c(-cval,-cval,-Inf),
upper=c(cval,cval,-cval), corr = corrmat )
Oneonly2 <- pmvnorm(lower=c(-cval,-cval,cval),
upper=c(cval,cval,Inf), corr = corrmat )
Oneonly3 <- pmvnorm(lower=c(-Inf,-cval,-cval),
upper=c(-cval,cval, cval), corr = corrmat )
Oneonly4 <- pmvnorm(lower=c(cval,-cval,-cval),
upper=c(Inf,cval,cval), corr = corrmat )
Oneonly5 <- pmvnorm(lower=c(-cval,-Inf,-cval),
upper=c(cval,-cval, cval), corr = corrmat )
Oneonly6 <- pmvnorm(lower=c(-cval,cval,-cval),
upper=c(cval,Inf,cval), corr = corrmat )

# Total chance of exactly 1 error
Oneonly=Oneonly1+Oneonly2+Oneonly3+Oneonly4+Oneonly5+Oneonly6

#Exactly 2 errors (illustrated by the 12 edges of a cube minus
the #upper and lower 5% in the corners)

# the 3 edges that corner the triple rejection in favour of
control #(lower left front)
onlyllx <- pmvnorm(lower=c(-cval,-Inf,-Inf), upper=c(cval,-cval,
-cval), corr = corrmat )
onlylly <- pmvnorm(lower=c(-Inf,-cval,-Inf), upper=c(-cval,cval,
```

```
-cval), corr = corrmat )
onlyllz <- pmvnorm(lower=c(-Inf,-Inf,-cval), upper=c(-cval,
-cval,cval), corr = corrmat )

# the 3 edges that corner the triple rejection in favour of the
#experimental arms (upper right back)
onlyurx <- pmvnorm(lower=c(-cval,cval,cval),
upper=c(cval,Inf,Inf), corr = corrmat )
onlyury <- pmvnorm(lower=c(cval,-cval,cval),
upper=c(Inf,cval,Inf), corr = corrmat )
onlyurz <- pmvnorm(lower=c(cval,cval,-cval),
upper=c(Inf,Inf,cval), corr = corrmat )

#Off edges (l=lower u=upper f=front b=back l=left r=right):
onlylrz <- pmvnorm(lower=c(cval,-Inf,-cval),
upper=c(Inf,-cval,cval), corr = corrmat )
onlyfry <- pmvnorm(lower=c(cval,-cval,-Inf),
upper=c(Inf,cval,-cval), corr = corrmat )
onlylbx <- pmvnorm(lower=c(-cval,-Inf,cval),
upper=c(cval,-cval,Inf), corr = corrmat )
onlyufx <- pmvnorm(lower=c(-cval,cval,-Inf),
upper=c(cval,Inf,-cval), corr = corrmat )
onlyulz <- pmvnorm(lower=c(-Inf,cval,-cval),
upper=c(-cval,Inf,cval), corr = corrmat )
onlybly <- pmvnorm(lower=c(-Inf,-cval,cval),
upper=c(-cval,cval,Inf), corr = corrmat )

# Total chance of exactly 2 errors
Twoonly <-
onlyllx+onlylly+onlyllz+onlyurx+onlyury+onlyurz+onlylrz+onlyfry+
onlylbx+onlyufx+onlyulz+onlybly

# Exactly 3 errors (Calculate probabilities in each of the 8
corners of a cube)
x1y1z1 <- pmvnorm(lower=c(-Inf,-Inf,-Inf),
upper=c(-cval,-cval,-cval), corr = corrmat )

x2y1z1 <- pmvnorm(lower=c(cval,-Inf,-Inf),
upper=c(Inf,-cval,-cval), corr = corrmat )
x1y2z1 <- pmvnorm(lower=c(-Inf,cval,-Inf),
upper=c(-cval,Inf,-cval), corr = corrmat )
x1y1z2 <- pmvnorm(lower=c(-Inf,-Inf,cval),
upper=c(-cval,-cval,Inf), corr = corrmat )

x2y2z1 <- pmvnorm(lower=c(cval,cval,-Inf),
upper=c(Inf,Inf,-cval), corr = corrmat )
x2y1z2 <- pmvnorm(lower=c(cval,-Inf,cval),
upper=c(Inf,-cval,Inf), corr = corrmat )
x1y2z2 <- pmvnorm(lower=c(-Inf,cval,cval),
upper=c(-cval,Inf,Inf), corr = corrmat )
```

```
x2y2z2 <- pmvnorm(lower=c(cval,cval,cval),
upper=c(Inf,Inf,Inf), corr = corrmat )

# Total chance of exactly 3 errors
Threeonly <-
x1y1z1+x2y1z1+x1y2z1+x1y1z2+x2y2z1+x2y1z2+x1y2z2+x2y2z2


#FWER – the overall error region of the sides, edges and corners
FWER=Oneonly+Twoonly+Threeonly

# Probability of at least any two errors
twoerr <- Twoonly+Threeonly

# Probability of three errors - sum of the corner regions
threeerr <- Threeonly

# Two MSFP - probability of at least two superior false
positives
# Sum of the 3 edges meeting the upper right back
#(i.e. two false positives along the plane of the third
distribution) # and the upper right corner

TwoMSFP <- onlyurx+onlyury+onlyurz+x2y2z2

# Three MSFP - probability of three superior false positives
ThreeMSFP <- x2y2z2

#Output results
corrmat
FWER
twoerr
threeerr
TwoMSFP
ThreeMSFP
```

C.3 R code to calculate the critical value required to control the probability of two MSFP errors based on correlated test statistics to that for independent trials, assuming a bivariate normal distribution.

```
setInternet2(TRUE)
install.packages("mvtnorm")
library(mvtnorm)

#The MSFP is the upper right corner of the rejection regions
#based on the standard bivariate normal.
#The MSFP needs to be controlled at 0.000625 (0.025**2)

#set correlation
corr <- 0.5

#correlation matrix
corrmat <- matrix(c(1,corr,corr,1),ncol=2,byrow=TRUE)

#Solve the critical value for the upper right corner
#equalling 0.000625
upperx <- qmvnorm(p=0.000625,tail=c("upper.tail"),corr=corrmat)
uppertail <- upperx$quantile
adjcval <- 2*(1-pnorm(uppertail))
adjcval
```