The application of Adaptive Designs in Phase III Randomised Controlled trials in Cancer

By

Pankaj Mistry

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Health Sciences.

University of Warwick, Warwick Clinical Trials Unit

December 2020

1 Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to dedicate this thesis to Bhagwan Swaminarayan and my loving Gurus; Pramukh Swami Maharaj and Mahant Swami Maharaj. The completion of the work in this thesis has been purely to their blessings, compassion and grace.

I would like to take this opportunity to thank my amazing family. I am incredibly grateful to my mother and father for always having faith and trust in everything that I have done. I would like to thank my dear wife who has supported me and kept me focused throughout this journey. Also I would like to acknowledge the support I have received from my brothers, sisters, nephews and niece.

I would like to acknowledge my supervisors Dr Andrea Marshall, Professor Janet Dunn and Professor Nigel Stallard for their guidance, motivation and support throughout this PhD. I am extremely grateful for all the time, patience and invaluable advice they have given.

Finally, I am thankful to my friends at Warwick CTU, Fabio Rigat and the team at Janssen Pharmaceuticals who have supported me during my PhD.

Declaration and inclusion of published work

I declare this thesis is my own work except where it contains work based on collaborative research. The nature and extent of my contributions have been indicated where applicable. I declare this thesis has not been submitted for a degree at another university.

The work contained in Chapter three has been published in a peer reviewed journal (1).

1. Mistry, P., Dunn, J.A., Marshall, A., 2017. A literature review of applied adaptive design methodology within the field of Oncology in randomised controlled trials and a proposed extension to the CONSORT guidelines. *BMC medical research methodology*, *17*(1), p.108.

In addition to the above, I have also presented my research at the following conferences:

1. Mistry P, Dunn JA, Marshall, A. Application of Adaptive Design methods – A literature review. London: Young Statisticians meeting 2016.
2. Mistry P, Dunn JA, Marshall, A. Application of Adaptive Design methods – A literature review. Birmingham: International Society for Clinical Biostatistics (ISCB) 2016.
3. Mistry P, Dunn JA, Marshall A. A literature review of applied Adaptive Design Methodology within the field of Oncology in randomised controlled trials. Liverpool: International Clinical Trials Methodology Conference (ICTMC) 2017.
4. Mistry P, Dunn JA, Marshall A. The use of MAMS designs in phase III Lung Cancer trials. Melbourne: ISCB 2018.

5. Mistry P, Dunn JA, Marshall A. The use of MAMS designs in phase III Cancer trials. Melbourne: ISCB 2018.
6. Mistry P, Dunn JA, Marshall A. Investigating the application of a multi-arm, multi-stage (MAMS) design to compare optimal treatment duration of Herceptin in a non-inferior setting in treating early breast cancer patients. ICTMC 2019.

Abstract

The research conducted within this thesis investigated the application of adaptive designs within phase III cancer trials, in particular, the use of multi-arm multi-stage designs (MAMS) were investigated. The application of adaptive designs in phase III randomised controlled trials are becoming common. These designs have the ability to reduce the number of patients required, be efficient in reducing the time needed and effective in having the ability to answer many questions within one trial.

A literature review found that adaptive designs were regularly implemented but the reporting of these methods were poor. I proposed there should be an adaptive design extension to the CONSORT 2010 guidelines and subsequently this has been published.

A scoping exercise of cancer trials conducted within the Warwick Clinical Trials Unit found that adaptive methods were regularly implemented without classifying them as adaptive designs. This scoping exercise was extended to identify exemplar trials that have implemented adaptive design methodology and to understand the design, conduct and reporting of these trials.

MAMS designs within three common cancer sites (breast, colon and lung) were investigated using a superiority hypothesis, primarily focusing on the longer term outcome of overall survival. Simulations were used to demonstrate how MAMS designs can be extended for use with a non-inferiority (NI) hypothesis. A MAMS design was then applied to a trial investigating the optimal duration of Herceptin therapy for patients diagnosed with early breast cancer. A MAMS design with a NI hypothesis appeared feasible from a statistical viewpoint however the operational aspects must be considered to ensure the trials' success.

My research showed that implementing a MAMS design within cancer trials can be more efficient and effective. It is anticipated that the use of adaptive designs within phase III randomised trials will continue to increase and are now required to be fully reported.

List of abbreviations

ACE – Adaptive Designs CONSORT Extension

ADT – Androgen Deprivation Therapy

ALCHEMIST – Adjuvant Lung Cancer Enrichment Marker Identification and Sequencing

ALK - Anaplastic Lymphoma Kinase

APPROVe - Adenomatous Polyp Prevention on Vioxx

ARTemis - Avastin Randomised Trial with neo-adjuvant chemotherapy for patients with early breast cancer

AVAST-M - Adjuvant Avastin trial in high risk melanoma

BATTLE – Biomarker-integrated approaches of Targeted Therapy for Lung Cancer Elimination

CI – Confidence interval

CMF – Cyclophosphamide, methotrexate and fluorouracil

CompARE - Comparing Alternative Regimens for escalating treatment of intermediate and high-risk oropharyngeal cancer

CONSORT – Consolidated Standards of Reporting Trials

COVID-19 - Coronavirus Disease 2019

CRF – Case report form

CTU – Clinical Trials Unit

De-ESCALaTE - Determination of Epidermal growth factor receptor-inhibitor versus Standard Chemotherapy early And Late Toxicity Events in Human Papillomavirus-positive oropharyngeal squamous cell carcinoma

DEVA – Docetaxel Epirubicin Adjuvant

DFS – Disease free survival

DMC – Data and Safety Monitoring committee

EGFR – Epidermal growth factor receptor

EMA – European Medicines Agency

ER – Estrogen receptor

FA – Fulvestrant plus anastrozole

FDA – Food and Drug Administration

FEC – Fluorouracil, epirubicin and cyclophosphamide

FFS – Failure free survival

Focus - Fluoxetine or control under supervision

FP – Fulvestrant plus placebo

FWER – Familywise type I error rate

gBRCA – Germline breast cancer gene

HER2 – Human epidermal growth factor receptor 2

HERA – Herceptin Adjuvant

HoR – Hormone receptors

HORG - Hellenic Oncology Research Group

HR – Hazard ratio

IDFS - Invasive Disease free survival

IO – Intermediate Outcome

IRPAN - Irinotecan plus panitumumab

IV – Intravenous

LIHNCS - Lugol's Iodine in Head and Neck Cancer Surgery

LMWH – Low molecular weight heparin

MAMS – Multi-arm multi-stage

MCID – Minimally clinical important difference

MFS – Metastasis-free survival

NCI-MATCH – MCI Molecular Analysis for Therapeutic Choice

NEAT – National Epirubicin Adjuvant Trial

NHS – National Health Service

NI – Non-inferiority

NIHR – National Institute for Health Research

NSCLC – Nonsquamous non-small-cell lung carcinoma

ONS – Office for National Statistics

OS – Overall survival

Optima - Optimal Personalised Treatment of early breast cancer using Multi-parameter Analysis

PARTNER - Platinum and PARP inhibitor for Neoadjuvant treatment of Triple Negative and/or BRCA positive breast cancer

PATCH - Prostate Adenocarcinoma: TransCutaneous Hormones

pCR – Pathologic complete response

PET-CT - Position emission tomography - Computed tomography

PHARE – Protocol of Herceptin Adjuvant with reduced exposure

PFS – Progression free survival

PICCOLO - Panitumumab, Irinotecan, and Ciclosporin in colorectal cancer

PO – Primary Outcome

PR – Progesterone receptor

PWER – Pairwise type I error rate

RCT – Randomised controlled trial

RFS –Relapse free survival

RecFS – Recurrence free survival

SoFEA – Study of Faslodex with or without concomitant Arimidex vs Exemestane following progression on non-steroidal Aromatase inhibitors

STAMPEDE - Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy

TDE – Time dependent effect

TEAMM - Tackling Early Morbidity and Mortality in Myeloma

TTE – Time to event

VTE - Venous Thromboembolism

WHO – World Health Organisation

1 Introduction

1.1 Background

Prior to introducing a new medical intervention, a sequence of processes must take place to ensure the safety and efficacy of these interventions. This is a vital process as these interventions can be beneficial as well as detrimental. A few examples of interventions could be new or existing drugs, surgical procedures, treatment of behaviour or preventive care (World Health Organisation 2019). The process of evaluating these new interventions is commonly through clinical trials. The World Health Organisation (WHO) has defined clinical trials to be, "Any research study that prospectively assigns human participants or groups of humans to one or more health related interventions to evaluate the effects on health outcomes." (World Health Organisation 2019). The most commonly accepted method of treatment evaluation that results in a change to clinical practice are known as a randomised controlled trial (RCT) and comes under the umbrella of phase III clinical trials (Akobeng 2005). The use of random allocation within phase III RCTs ensures that selection bias is reduced when assigning patients to treatment groups. Furthermore, the use of stratification ensures that treatment groups are equally divided between prognostic groups (Sacks et al. 1982).

The first RCT recorded by the Medical Research Council (MRC) in the UK was performed in 1950 comparing the use of streptomycin in treating tuberculosis (Long et al. 1950). The first recorded RCT in the field of Oncology began in 1955 in the US and compared two chemotherapy regimens in patients diagnosed with acute leukaemia (Frei et al. 1958). Since then, phase III RCTs within the field of Oncology have continued to develop and have proved effective in evaluating treatments (Gehan 1979, Booth et al. 2008). However, clinical research continues to explore and evaluate new treatments and methods that will help to prevent, diagnose or treat cancer (Comprehensive Cancer Centers 2017). The research conducted within this thesis will primarily focus on statistical methods and trials related to the therapeutic area of cancer.

1.2 Bayesian and frequentist philosophies

The application of statistics in clinical trials can be categorised under two principle philosophies; Bayesian or frequentist (Berry et al. 2010). Some of the key differences between these methods are:

1. ***The probability of parameters***: The Bayesian approach believes that all unknowns have some form of a probability distribution compared to the frequentist approach where probabilities are defined based on data. Bayesian methods use prior and posterior distributions. Information parameterised before the data is collected is the prior distribution and information parameterised once data is observed is the posterior distribution.
2. ***Available evidence***: Frequentist methods are explicit to particular experiments compared to Bayesian methods which utilise all available information related to the parameters of interest.
3. ***Conditioning on the observed***: Bayesian methods focus only on observed results compared to frequentist methods that use probabilities of data to predict possibilities that may have not been directly observed.
4. ***Strictness***: Frequentist approach have more stricter criteria to complete an experiment in comparison to Bayesian methods (Berry et al. 2010). For example, before starting a trial, the frequentist approach requires a sample size compared to a Bayesian approach where sample size may not be required at the beginning.
5. ***Randomisation***: The role of randomisation is of utmost importance in phase III trials. Randomisation reduces selection bias and allows the balancing of pre-specified covariates. The Bayesian approach does not depend on randomisation and can use subjective probability compared to the frequentist approach where randomisation offers inference on clinical trial results (Berchialla et al. 2019).

The key differences between these philosophies have been highlighted here as Bayesian methods are used in adaptive trials, particularly in early phase trials, but the focus of this thesis is on phase III clinical trials, where more frequentist

methods are applied. Therefore, the review of adaptive design methods in chapter two will focus on the frequentist approaches but will highlight some of the Bayesian adaptive methods for completeness.

1.3 Survival Analysis

The term 'Survival Analysis' can be defined as the analysis of data in the form of different timepoints where there is a defined starting point measured up until the incidence of a certain event (Collett 2015). The rate of survival at different timepoints can be shown either through life-tables or more graphically through a survival curve. The limitations of these methods are that the survival time cannot be measured until an event has occurred therefore the output can only be produced once all patients have had an event. Hence, the concept of censoring, allows the use of partial information so patients can be censored either during the trial because they may have withdrawn or were lost to follow up or at the end of the trial.

In the context of a RCT of a control versus a new treatment with a time-to-event (TTE) endpoint, Kaplan-Meier curves can be estimated and compared using a log-rank test to test whether there is a difference between two survival curves and are useful only when the predictor is categorical variable, i.e. treatment A vs treatment B. The Cox proportional hazards model can incorporate categorical as well as continuous variables and can include more than one predictor variable. The hazard ratio (HR) can be used to measure the frequency of an event in one treatment arm compared to another. A HR greater than (or less than) one would indicate that survival was better (or worse) in one of the treatment arms. A HR equal to one would indicate that there is no difference between the arms.

1.3.1 HR calculation

In survival analysis, the hazard rate is the probability of an event occurring at time t given that it survives until time t or later. Therefore, the HR is the ratio of the hazard rates for the experimental treatment against a control. The HR can be obtained using a TTE survival rate at time t for the control and the experimental TTE survival rate by the following formula:

$$\frac{\log\left(Experimental\ TTE\ survival\ rate\ at\ time\ t\right)}{\log\left(Control\ TTE\ survival\ rate\ at\ time\ t\right)} = Hazard\ ratio$$

For sample size calculations, the hypothesised HR can be calculated using the TTE survival rate at time t for the control and the experimental TTE survival rate determined based on the absolute difference wanting to detect. For example, to demonstrate an improvement in treatment (superiority) of at least 7.5% from the control arm with a five-year survival rate of 85% for breast cancer, the HR is calculated by:

$$\frac{\log\left(0.925\right)}{\log\left(0.850\right)} = 0.480$$

Similarly, the following formula can be used when only the median of the TTE outcomes are available:

$$\frac{Control\ median\ TTE\ outcome}{Experimental\ median\ TTE} = Hazard\ ratio$$

For example, to demonstrate that an absolute improvement in median OS for the experimental arm of at least 30% from a median OS on the control arm of 21.6 months for lung cancer, the HR is calculated by:

$$\frac{21.6}{21.6\ \times 1.3} = 0.769$$

Therefore, a HR of 0.769 is the same as saying that there will be approximately 23% relative reduction in events in the experimental arm.

1.3.2 Parametric models

Parametric models are often used to simulate and analyse survival data (Machin 2006). The more popular distributions that are used are the exponential, Weibull and gamma distributions.

The exponential distribution can be used when the hazard rate is constant in a treatment group and is of the form, $S(t) = exp^{-\lambda t}$, where $S(t)$ is the survival function at time t and λ is the constant hazard. The shape of this distribution for different λ values can be seen in Figure 1.1.

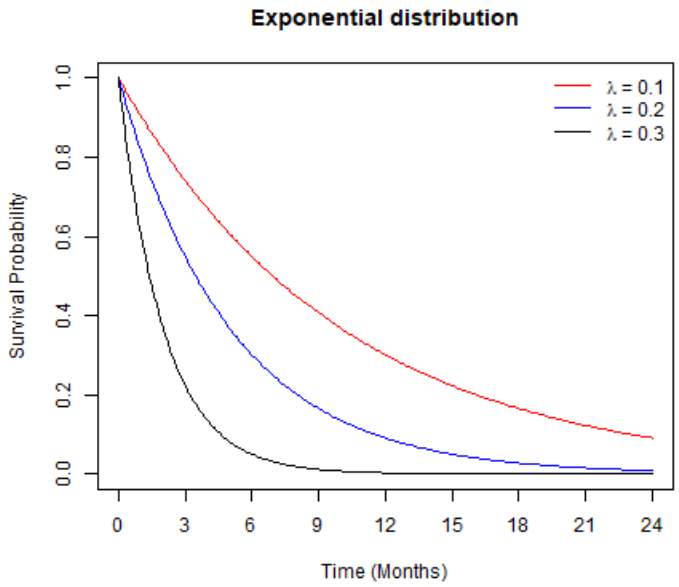

Figure 1.1: Exponential distribution when λ = 0.1, λ = 0.2 and λ = 0.3.

The Weibull distribution is used when either the hazard rate is constant or increasing. The Weibull distribution is a modification of the Exponential distribution and of the form $S(t) = \exp[-(\lambda t)^k]$, where the shape of this distribution is dependent on the constant *k*. An illustration of this distribution for λ = 0.1 and different values of *k* can be seen in Figure 1.2.

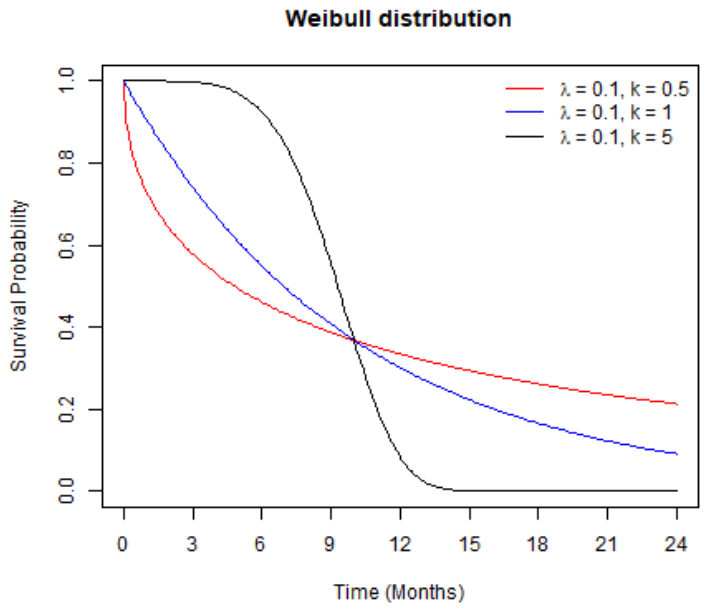

Figure 1.2: Weibull distributions when λ = 0.1 and k = 0.5, k = 1 and k = 5.

Similarly, the gamma distribution is a special case of the exponential distribution (Cox et al. 2007). The survival function is of the form $S(t) = 1 - I_k(\lambda t)$, where $I_k(\lambda t)$ is an incomplete gamma integral (Kiche et al. 2019). An example of the gamma distribution with different values for the constant *k* can be seen in Figure 1.3.

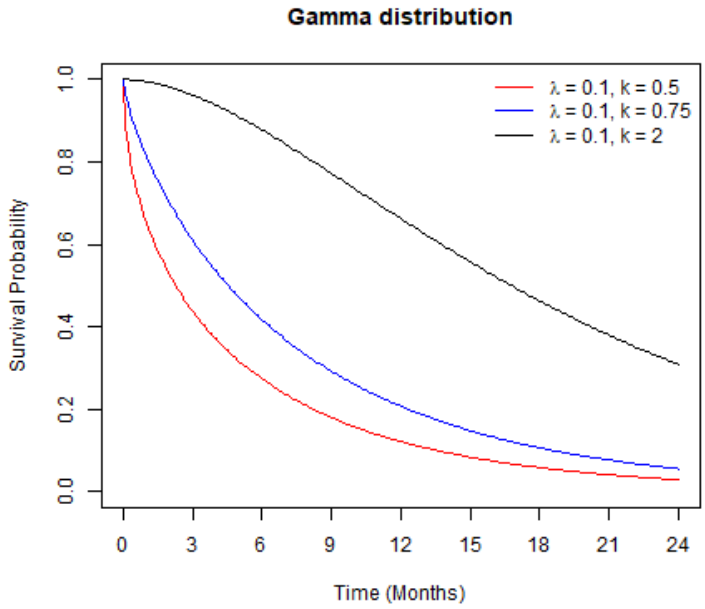

Figure 1.3: Gamma distribution when λ = 0.1 and k = 0.5, k = 0.75 and k = 2.

Further information on other popular distributions that could be used for analysing survival data such as the log-normal or log-logistic distributions can be found in the book 'Survival Analysis: A practical approach' by Machin et al (Machin 2006).

1.3.3 Time-to-event outcomes within cancer

Survival analysis methods are frequently used in the context of cancer trials due to many outcomes being the time to an event of interest. There are many types of 'time-to-event' outcomes within cancer which are of special interest and can be used to assess the impact of a treatment (US Food and Drug Administration 2007). Few of these common events of interest are defined and discussed below (Hudis et al. 2007):

- ***Overall survival (OS)*** - *d*efined as the time from randomisation till death due to any cause. Patients are censored at the date last known to be alive. Considered one of the easiest endpoints to assess however can extend the duration of a trial due to the potential length of follow-up required.
- **Disease-free survival (DFS)** – defined as the time from randomisation till disease recurrence or death due to any cause. DFS is considered a surrogate endpoint especially where death could be a prolonged and unrealistic endpoint. Furthermore, using DFS can accelerate the time it can take for a potential new drug or regimen to be approved. This is also known as relapse-free survival (RFS) or recurrence-free survival (RecFS).
- **Progression-free survival (PFS)** – defined as the time from randomisation till tumour progression, which can be clearly outlined in the trial protocol as this varies from trial to trial, or death. The definition of PFS has an outcome of death and hence can be considered a surrogate endpoint for OS.
- **Failure-free survival (FFS)** - defined as the time from when primary treatment for the cancer ceases till the time a patient experiences specified events, also known as event-free survival.
- **Metastasis-free survival (MFS)** – defined as the time from when primary treatment of cancer ceases till the cancer spreads to other parts of the body. Patients are censored at the date last known to be alive and free from metastasis.

1.3.4 Example – survival curves showing DFS

Figure 1.4 displays survival curves showing DFS (previously defined in section 1.3.3) for breast cancer patients that received Herceptin (trastuzumab) for one-year compared against observation only. Patients that received Herceptin for one year had a two-year DFS rate of 85.8%, patients on observation only had a two-year DFS rate of 77.4%. The HR for a risk of an event in the one-year Herceptin group compared to the observation group was 0.54 (95% CI: 0.43 to 0.67); this can be interpreted as just under half as many patients in the one-year Herceptin group would experience an event compared to the observation arm. A p-value < 0.0001 obtained from comparing the trial arms by the log-rank test shows there is a statistically significant difference between the two treatment arms.

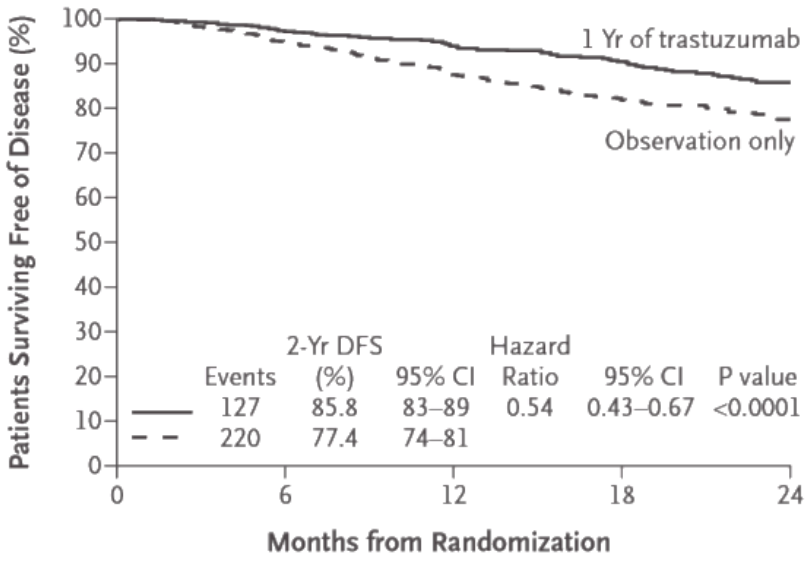

Figure 1.4: Survival curves showing DFS, breast cancer patients that received one year Herceptin compared against patients on observation only. Example taken from HERA trial (Piccart-Gebhart et al. 2005) . Hazard ratio with 95% CI and p-value shown. DFS – Disease free survival, CI – Confidence Interval.

1.4 Hypothesis testing options

The choice of hypothesis to implement within phase III trials is dependent on the objective of the study. It could be to demonstrate that a new treatment is better compared to the current standard treatment or a placebo, which is classified as a superiority hypothesis. Alternatively, the objective of a study could be to

determine if a new treatment is the same (equivalent) or no worse (non-inferiority) to a current standard treatment. Each of these hypotheses vary in terms of their design, analysis and interpretation; the hypotheses associated with each trial design are shown in Table 1.1.

Hypothesis testing options	Null Hypotheses	Alternative Hypotheses
Superiority	No difference between treatments	Difference between treatments
Equivalence	Treatments are not equivalent	Treatments are equivalent
Non-Inferiority	New treatment is inferior to current treatment	New treatment is not inferior to current treatment

Table 1.1: The null and alternative hypotheses for the different hypothesis testing options (Walker et al. 2011).

1.4.1 Superiority hypothesis

A two-sided superiority hypothesis to compare two treatments in a phase III cancer trial can be defined as follows:

$$H_0: \delta = 1$$

$$H_1: \delta \neq 1$$

where δ can be defined as the HR. The null hypothesis, H_0, specifies that there is no difference between treatments and the alternative hypothesis, H_1, specifies that there is a difference without indicating which direction i.e. the experimental arm could be better or worse than the control arm. The one-sided superiority alternative hypothesis specifies the direction for the alternative hypothesis (Bland et al. 1994), such that the experimental treatment is hypothesised to be better than the control by an amount δ_1. Hence a one-sided superiority hypothesis with a critical HR, δ_1, to compare two treatments in a phase III cancer trial can be defined as follows:

$$H_0: \delta \geq \delta_1$$

$$H_1: \delta < \delta_1$$

The null hypothesis indicates that the experimental arm is either the same or worse than the control arm. Comparatively, the alternative hypothesis indicates that the experimental arm is better than the control arm.

For this thesis, implementing a superiority hypothesis refers to a trial involving a one-sided significance level with a pre-specified absolute or relative margin, e.g. 3% absolute difference. The first step in analysing a trial with a superiority hypothesis is to perform a statistical significance test (p-value), to assess the assumption that there is no difference between treatments (null hypothesis). In most cases, a p-value < 0.05 means that the data suggests that there is a difference between the two treatments (alternative hypothesis) hence rejecting the null hypothesis (Table 1.1). If the data suggests that there is a difference between treatments then it is important to measure the size of the difference and if it is clinically relevant (CPMP 2001). Thereafter, a confidence interval (CI) is generated (usually at the 95% CI), which gives a range of values for which the true value will most probably be within. For example, using a time-to-event outcome within a cancer setting, a new treatment is considered superior to the control treatment if the HR and its respective CI is below the pre-specified margin as shown in Figure 1.5 (Dunn et al. 2018).

The potential results obtained from a one-sided superiority trial with a pre-specified margin can be seen in Figure 1.5 and are as follows:

1. HR favours the new treatment however the lower bound of the CI is greater than the critical HR and the upper bound of the CI is less than one suggesting that superiority of the experimental arm is demonstrated but it is not conclusive (**Superior, not clinically meaningful**);

2. HR favours the control treatment and the lower bound of the CI above one therefore suggesting that the new treatment is inferior to the control (**Inferior, clinically meaningful**);

3. HR favours the new treatment and the upper bound of the CI is below the critical HR therefore it is conclusive that superiority of the experimental

arm above the specified margin is demonstrated (**Superior, clinically meaningful**);

4. HR favours the new treatment however the lower bound of the CI is less than the critical HR and the upper bound of the CI is above one therefore neither superiority nor NI of the experimental arm are demonstrated (**Inconclusive**);

5. HR favours the new treatment however the upper bound of the CI is less than one and greater than the critical HR which indicates that superiority of the experimental arm is demonstrated but it is not conclusive (**Superior, not clinically meaningful**).

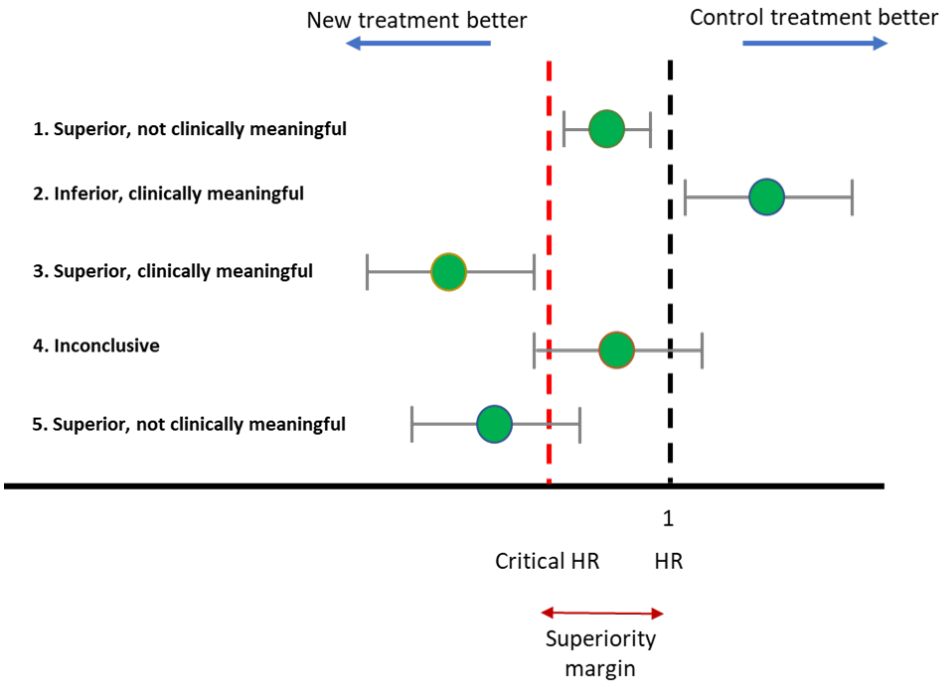

Figure 1.5: Potential results of using a superiority hypothesis comparing two drugs shown with HR and CI. HR = Hazard ratio; CI = Confidence interval.

1.4.2 Equivalence hypothesis

An equivalence hypothesis aims to determine that one treatment is 'similar' to another; this can be defined as follows:

$$H_0: \delta \geq \delta_U \; or \; \delta \leq \delta_L$$

$$H_1: \delta_L < \delta < \delta_U$$

where δ_U is defined as the upper critical HR and δ_L is defined as the lower critical HR.

For this hypothesis testing option, the null hypothesis would be that treatments are not equivalent, and the alternative hypothesis would be that the new treatment is equivalent to the current treatment (Table 1.1). Therefore, a p-value < 0.05 would suggest that treatments are equivalent. To assess the size of the difference between the two treatment within this setting, an equivalence margin is defined. The two treatments are classed as equivalent if the CI of the new treatment lies within the pre-specified margins. For example, using a time-to-event outcome within a cancer setting, a new treatment is considered equivalent to the control treatment if the HR and CI is between the two pre-specified margins (Figure 1.6).

The potential results obtained from using an equivalence hypothesis can be seen in Figure 1.6 and are as follows:

1. HR favours the control treatment however the lower bound of the CI is greater than one and the upper bound of the CI is less than upper critical HR indicating that the new treatment is inferior to the control but is not conclusive as the HR and 95% CI are below the upper critical HR (**Equivalent, Inferior, clinically meaningful**);

2. HR equals one indicating that there is an equal effect between the new and control treatment however the lower bound of the CI is below the lower critical HR and upper bound of the CI is above the upper critical HR resulting in an inconclusive trial (**Equivalent, not clinically meaningful**);

3. HR and the lower bound of the CI is above one and within the upper critical HR. However, the upper bound of the CI is greater than the upper critical HR therefore suggesting that the new treatment is inferior to the control but this is not conclusive as the HR and 95% CI are not greater than the upper critical HR (**Inferior, not clinically meaningful**);

4. HR favours the control treatment and the lower and upper bounds of the CI are between the lower and upper critical HR therefore equivalence is demonstrated (**Equivalent, clinically meaningful**);

5. HR favours the new treatment however the lower bound of the CI is less than the lower critical HR and the upper bound of the CI is less than one therefore superiority is demonstrated but it is not conclusive as the HR and upper 95% CI are not less than the lower critical HR (**Superior, not clinically meaningful**);

6. HR favours the new treatment and the upper bound of the CI is less than the lower critical HR therefore superiority is demonstrated (**Superior, clinically meaningful**).

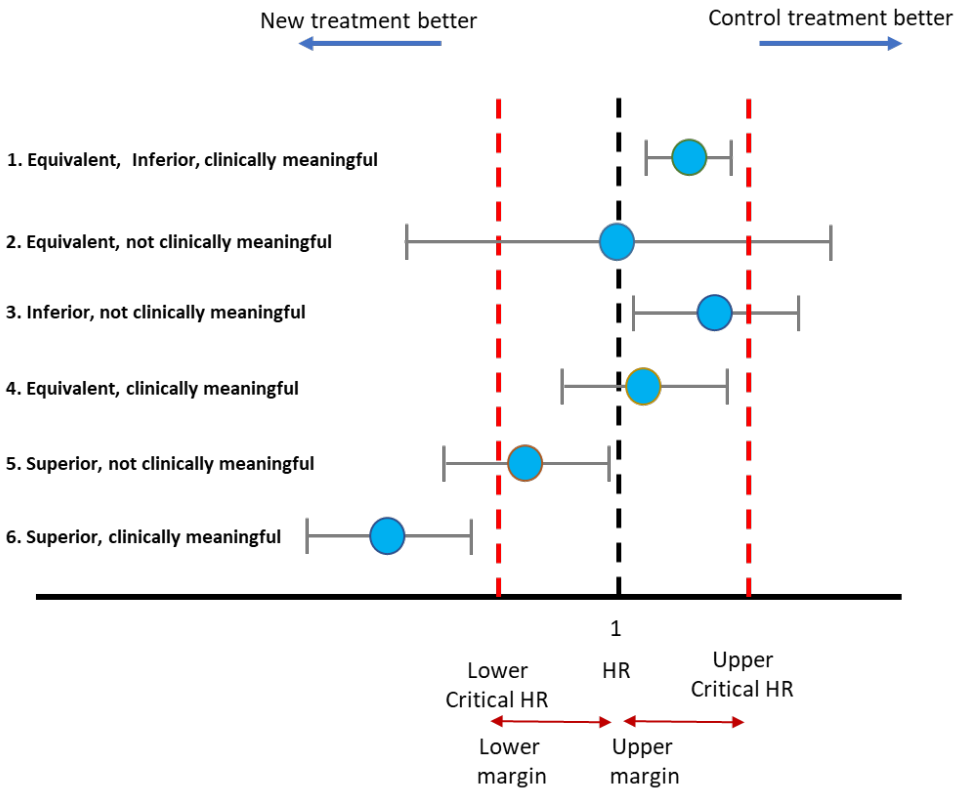

Figure 1.6: Potential results of using an equivalence hypothesis comparing two drugs shown with HR and CI. HR = Hazard ratio; CI = Confidence interval.

1.4.3 Non-inferiority hypothesis

A phase III trial may assess another treatment that could be similar to the current standard treatment but less toxic or cost-effective hence a non-inferiority (NI) hypothesis may be implemented (Weiskopf et al. 2016, Saad 2018). The null hypothesis would be that the new treatment is inferior to the current treatment and the alternative hypothesis would be that the new treatment is not inferior to the current treatment (Table 1.1). Therefore, a p-value < 0.05 would suggests that NI has been demonstrated. A NI trial aims to show that a new treatment is no worse than the current standard by a certain amount also known as the NI margin. Therefore, NI can be demonstrated if the upper CI is within the pre-specified NI margin. Using a time-to-event outcome within a cancer setting, a new treatment demonstrates NI if the HR and the upper CI is below the critical HR.

The use of NI hypothesis within phase III cancer trials has slowly increased over the years. The ClinicalTrials.gov website was used to search for all phase III cancer trials and from these trials how many implemented a NI hypothesis

(ClinicalTrials.gov 2020). The results found only 0.5% beyond 20 years ago (3/654), 1.2% between 10 and 20 years ago (33/2840), 2.7% between 5 to 10 years ago (50/1839) and 2.7% in the past five years (67/2476) have implemented a NI hypothesis.

The potential results obtained from a NI trial can be seen in Figure 1.7 and are as follows:

1. HR favours the control treatment however the upper bound of the CI is greater than the critical HR hence suggesting that NI is not demonstrated (**Not Inferior, not clinically meaningful**);
2. HR equals one indicating that there is an equal effect between the new and control treatment however the upper bound of the CI is greater than the critical HR suggesting that NI is not demonstrated (**Equivalent, not clinically meaningful**);
3. HR favours the new treatment however the upper bound of the CI is greater than the critical HR suggesting that NI is not demonstrated (**Superior, not clinically meaning**);
4. HR favours the control treatment and the upper bound of the CI is below the critical HR and the lower bound of the CI is greater one demonstrating NI but also suggesting that the new treatment is actual inferior to the control (**Not Inferior and inferior, clinically meaningful**);
5. HR equals one indicating equal effect and the upper bound of the CI is below the critical HR therefore NI is demonstrated (**Not inferior, clinically meaningful**);
6. HR favours the control treatment and the upper bound of the CI is below the critical HR but the lower bound of the CI is less than one suggesting

that the new treatment shows NI but is not conclusive (**Not-inferior, clinically meaningful**);

7. HR favours the new treatment and the upper bound of the CI is above one but within the critical HR therefore NI is demonstrated but superiority is not demonstrated (**Not inferior, clinically meaningful**);
8. HR favours the new treatment and the upper bound of the CI is less than one which indicates that both superiority and NI are demonstrated (**Superior and not inferior, clinically meaningful**);
9. HR favours the control treatment but the upper bound of the CI above the critical HR and the lower bound of the CI is greater one suggesting that the new treatment is actual inferior to the control but cannot declare NI (**Not inferior, not clinically meaningful**).

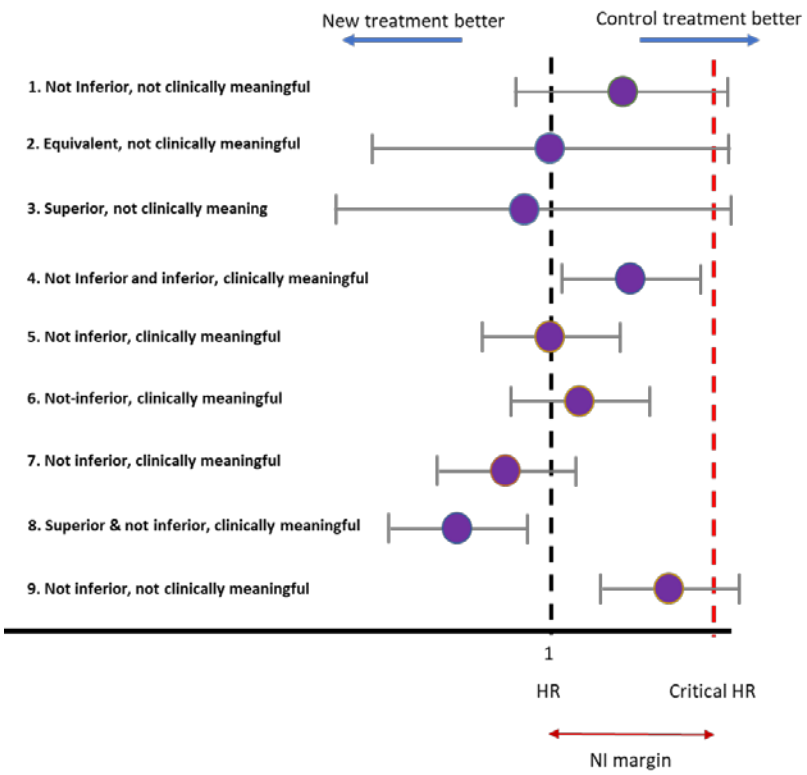

Figure 1.7: Potential results of a NI hypothesis comparing two drugs shown with HR and CI. HR = Hazard ratio; NI = Non-inferiority; CI = Confidence interval.

1.4.3.1 Why use a NI hypothesis?

Due to the many toxicities of cancer therapy, testing treatments in phase III cancer trials with a NI design can result in improving the patient care with more convenient and acceptable regimens (Riechelmann et al. 2013). Applying a NI hypothesis offer trial stakeholders the opportunity to compare alternate treatments with potentially less toxicity or reduced durations of treatments against the current standardised treatments which can also appeal to patients. For example, a recent trial implemented a NI hypothesis to compare the use of intravenous (IV) treatment using Daratumumab in combination with standard regimens (control) against administering Daratumumab subcutaneously for patients diagnosed with multiple myeloma (Luo et al. 2019). Administering treatments subcutaneously instead of having IV treatment would benefit both patients and trialist as patients would be able to administer treatment within a significantly reduced time in the comfort of their own home hence reducing time spent in hospital. Thus, implementing these designs could offer benefits such as reduced toxicity, improved cost effectiveness and may be more appealing to patients.

1.4.4 Selection of margin

The selection of the margin is a critical step when designing trials (Cook et al. 2018). The margin should be estimated based upon the performance of the control treatment in past trials, ideally within a placebo-controlled trial (US Food and Drug Administration 2016). It is then assumed that within the trial, the active control will have its expected effect; this is also known as assay sensitivity (US Food and Drug Administration 2016). Another way of interpreting assay sensitivity is that if the trial included a placebo then there would be a significant difference between the control treatment and placebo. Hence, a trial is said to be successful if there is an acceptable small difference between treatments however this may not have had assay sensitivity therefore not supporting the conclusion that the experimental treatment was effective.

The selection of a margin based on a placebo-controlled trial may show statistical significance (robustness of study results) but may not be clinically significant

(impact on clinical practice). In the case of a superiority or NI hypothesis, another margin less than the original margins used in previous placebo-controlled trials may be defined as the largest difference between two treatments to be clinically acceptable and hence demonstrate superiority or NI respectively (US Food and Drug Administration 2016). Furthermore, the ICH E9 guidelines suggests that the NI margin should be smaller than the differences observed in superiority trials of the control group (European Medicines Agency 1998).

In some studies, the obtained TTE outcome can differ to the original TTE event assumption made, therefore by using an absolute margin and adjusting the critical HR assumption at the time of analysis can preserve the power of the study.

Patients play a pivotal role in determining what the NI margin should be by determining the amount of time or percentage chance of dying they would be willing to accept in return for having a less toxic treatment and the symptoms that go with it. Furthermore, trial teams should consult the DELTA^2 guidance which provides recommendations in undertaking sample size calculations and choosing the target difference (Cook et al. 2018).

1.4.5 Relationship between superiority and NI hypotheses

For a one-sided trial with a superiority hypothesis comparing two treatment arms with a survival endpoint (Collett 2015), let δ_1 be defined as the critical HR where $\delta_1 < 1$ such that:

$$\delta_1 = \frac{log\ S_n(t)}{log\ \ S_c(t)}$$

where $S_n(t)$ and $S_c(t)$ are the estimated survival functions at time t on the new treatment and control treatment respectively.

The null hypothesis, H_0, and the alternative hypothesis, H_1, can be defined by:

$$H_0: \delta \geq \delta_1$$

$$H_1: \delta < \delta_1$$

where δ is the obtained HR. Therefore, the probability to conclude superiority is:

$$P(Conclude\ Superiority; \delta_{max} = 1) = \alpha$$

$$P(Conclude\ Superiority;\ \delta_{max} = \delta_1) = 1 - \beta = Power$$

where α and β are defined as the type I and type II error respectively (Collett 2015). Assuming that patients are equally allocated to the two treatment arms, the number of events, e, required to determine whether the experimental arm is superior above a specified margin or not can be obtained from the following equation:

$$e = \frac{4\left(z_{\alpha/2} + z_\beta\right)^2}{(log\ \delta_1)^2}$$

where $z_{\alpha/2}$ and z_β are the upper $\alpha/2$ and β points respectively of the standard normal distribution (Collett 2015). Therefore, the number of patients required, n, i.e. the sample size, can be calculated from the following:

$$n = \frac{e}{P(Event)}$$

where the probability of an event can be taken as:

$$P(Event) = 1 - 1/6\left\{\hat{S}(t_f) + 4\hat{S}\left(\frac{t_r}{2} + t_f\right) + \hat{S}(t_r + t_f)\right\}$$

where t_r and t_f are defined as the recruitment time and follow-up time respectively and the exponential survival function, $\hat{S}(t)$, can be calculated such that:

$$\hat{S}(t) = \frac{S_c(t) + S_n(t)}{2},$$

Hence, to implement this to a one-sided trial with a NI hypothesis comparing two treatment arms with a survival endpoint, let δ_2 be defined as the critical HR where $\delta_2 > 1$ such that:

$$\delta_2 = \frac{log\ S_n(t)}{log\ \ S_c(t)}$$

The null hypothesis, H_0, and the alternative hypothesis, H_1, can be defined by:

$$H_0: \delta \geq \delta_2$$

$$H_1: \delta < \delta_2$$

Therefore, the probability to conclude NI is:

$$P(Conclude\ NI; \delta_{max} = \delta_2) = \ \alpha$$
$$P(Conclude\ NI; \delta_{max} = 1) = \ 1 - \beta = Power$$

Therefore, assuming patients are equally allocated to the two treatment arms, the number of events can be calculated using the equation below, such that:

$$e = \frac{4\left(z_\beta + z\alpha_{/2}\right)^2}{(log\ \delta_2)^2}$$

Hence, the number of patients required can be calculated using the same equation above. This implies that using the same parameters in the NI design as used in the 1-sided superiority design will result in the same number of events required and therefore the same sample size.

1.4.5.1 Example – relationship between superiority and NI

For a superiority hypothesis, a new treatment for patients diagnosed with breast cancer is expected to increase the overall survival rate at five years from 75% in the control arm (standard treatment) to 80%. In contrast for a non-inferiority hypothesis, the overall survival rate at five years is 80% in the control arm and the new treatment is expected to be no worse than 75%. Therefore, the corresponding value of δ_1 and δ_2 for a superiority design and NI design respectively would be:

$$\delta_1 = \frac{log\ (0.80)}{log\ (0.75)} = 0.776 \qquad \delta_2 = \frac{log\ (0.75)}{log\ (0.80)} = 1.289$$

Let $\alpha = 0.05$ and $1 - \beta = 0.80$, hence the number of deaths required to compare the two treatments such that there is an 80% chance of detecting a HR of 0.776 (Superiority) or 1.289 (NI) to be statistically significant at the 5% significance level can be calculated by:

$$e = \frac{4 \times 7.85}{(log\ 0.776)^2} = 488 \qquad e = \frac{4 \times 7.85}{(log\ 1.289)^2} = 488$$

Therefore, it can be shown that 488 deaths are required to have 80% power to detect a HR of 0.776 or 1.289 for a one-sided superiority design or NI design respectively to conclude significance.

It is anticipated that patients will be recruited over a three-year period with a follow-up period of two-years. Therefore, the number of patients required can be calculated such that the probability of an event is:

$$P(Event) = 1 - {}^{1}/_{6}\left\{\hat{S}(2) + 4\hat{S}\left(\frac{3}{2} + 2\right) + \hat{S}(3 + 2)\right\}$$

$$= 1 - {}^{1}/_{6}\{\hat{S}(2) + 4\hat{S}(3.5) + \hat{S}(5)\}$$

Hence, the value of the different survival functions are:

$$\hat{S}(2) = \frac{S_c(2) + S_n(2)}{2} = \frac{0.91 + 0.89}{2} = 0.9$$

$$\hat{S}(3.5) = \frac{S_c(3.5) + S_n(3.5)}{2} = \frac{0.86 + 0.82}{2} = 0.84$$

$$\hat{S}(5) = \frac{S_c(5) + S_n(5)}{2} = \frac{0.8 + 0.75}{2} = 0.78$$

Using this information, the probability of death is:

$$P(Event) = 1 - {}^{1}/_{6}\{0.9 + (4 \times 0.84) + 0.78\} = 0.16$$

Therefore, the total number of patients required for the superiority design and NI design is:

$$n = \frac{e}{P(Event)} = \frac{488}{0.16} = 3050$$

Therefore 3050 patients (1525 in each arm) will be required over a period of three-years when implementing a superiority or NI hypothesis.

1.5 Adaptive designs

An alternative to the gold standard RCT parallel design, which has recently received much attention, are adaptive designs (Chow 2014, Bothwell et al. 2018). The definition of an adaptive design is any pre-planned modification made to the trial whilst preserving the validity and integrity of the trial (Gallo et al. 2006, US Food and Drug Administration 2018). These designs are attractive due to the efficiency and flexible nature in answering many questions within one trial (Berry 2012). They often involve re-estimating the sample size at a certain time after the trial has begun, considering more than one intervention at the same time rather than conducting separate trials or adding in treatment arms. Additionally, the long duration of Oncology trials, the lack of efficacy of phase III drug trials and the cost involved with clinical trials has led to the encouragement and adoption from regulators for more novel trial designs (US Food and Drug Administration 2010, Berry 2012).

Despite these advantages, adaptive designs in practice have lacked execution in comparison to the methodology that has been published. This may have been due to the Food and Drug Administration (FDA) classifying some of these designs as 'well-understood' and others as 'less well-understood' in their 2010 draft guidance (US Food and Drug Administration 2010). The lack of implementation for these novel trial designs may be as a result of the fear that if these designs aren't executed correctly, it could result in an unanswered trial (Chow 2014).

The main concerns of adaptive designs are the control of the type I and type II error rates for the duration of the trial; failure to control these rates could lead to possible bias assessed by the amount the treatment effect deviates from its true value. For example, Figure 1.8 shows the time at which different interim analysis will take place when the required number of events are reached for three event rates for a clinical trial. If the event rate was linear (event rate B), then the appropriate time for an interim analysis would occur at time 2 as this is when the

required number of events is reached however if the interim analysis took place earlier i.e. at time 1 or later at time 3, then it could lead to a possible bias of the true value of the treatment effect. For event rate A or event rate C the appropriate time for an interim analysis would be at time point 1 and 3 respectively.

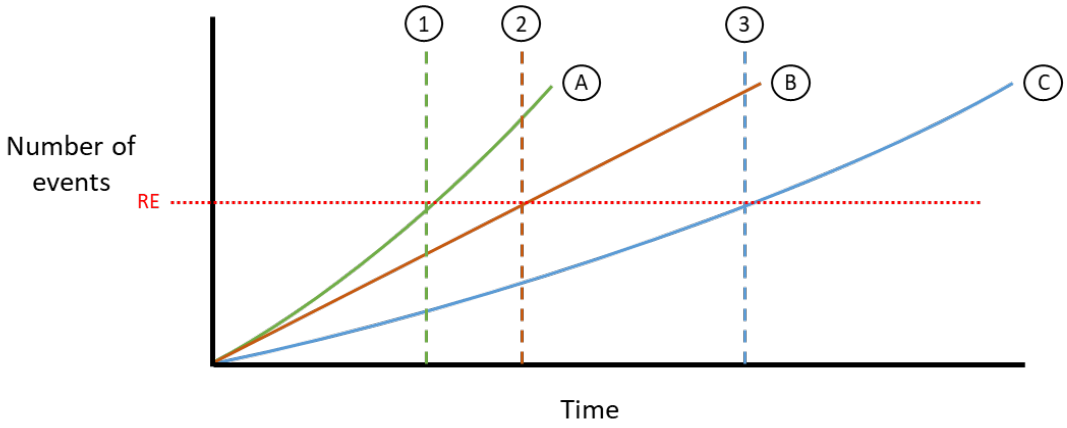

Figure 1.8: Different event rates over time and how the timing of the interim analysis changes accordingly. If interim analysis took place too early, it could lead to possible bias. 1,2,3 – represent the interim analyses. A, B, C – Different event rates. RE – Required number of events to trigger interim analyses.

1.6 Research aim and rationale

The research conducted in this thesis has been motivated by the rise in interest of adaptive design methods. Both the U.S FDA and the National Cancer Institute are investigating these novel designs due to the added pressures of speeding up both clinical trials and the evaluation of new drugs (Schmidt 2007).

Only 34% of phase III cancer trials achieve statistical significance whilst the time and the number of patients required to perform these trials continues to increase (Reitsma et al. 2015). Hence, the incentive of creating novel trial designs in diseases such as cancer will support the ideologies of the trialist and patients to ensure that clinical trials will continue to become more efficient and effective in the future.

The aim of this research is to examine the current types of adaptive designs within phase III cancer trials, within trials conducted at the Warwick Clinical Trials Unit

(CTU) and to explore innovative adaptive design methods that could be implemented.

1.7 Objectives

The objectives of this PhD thesis were to:

- Understand how adaptive designs are being applied and reported within phase III cancer clinical trials;
- Perform a scoping exercise of all cancer trials executed at Warwick CTU to assess the adaptive methods involved;
- Assess the application of multi-arm multi-stage (MAMS) designs to different cancer sites to examine what factors contribute to a 'successful' MAMS design
- Show through simulations how the MAMS designs can be used to implement a NI hypothesis;
- Implement a MAMS design with a NI hypothesis when investigating treatment duration and make recommendations on the operational aspects when implementing these designs.

1.8 Thesis Structure

This thesis begins by investigating some of the more common methods that are used and classed as 'adaptive design'. Details of the literature review that was undertaken to evaluate the use of adaptive design methods in RCTs within the field of Oncology are presented in chapter three. Guidelines for improving the reporting of trials using adaptive designs are also described in chapter three. The conclusions drawn from chapter three that the reporting of adaptive design methods were inadequate in clinical trials led to the scoping exercise performed at Warwick CTU in chapter four to understand what adaptive methods were being implemented. Hence, this chapter investigated the on-going/completed Oncology

trials within Warwick CTU and led to identifying any external RCT exemplars that had implemented adaptive design methods.

Chapter five investigates the use of hypothetical MAMS designs using the 'nstage' command in Stata (Blenkinsop et al. 2019) with a superiority hypothesis in three cancer sites: breast cancer, lung cancer and colon cancer. These specific cancer sites were selected due to their different survival rates and would provide a variety of results.

In chapter six, simulations were conducted to show that a MAMS design with a NI hypothesis could be implemented using the 'nstage' command. Thereafter, hypothetical MAMS designs using a NI hypothesis in breast and colon cancer were investigated.

Subsequently, in chapter seven, MAMS designs were applied to a proposed hypothetical trial investigating the optimum duration of Herceptin therapy in patients diagnosed with early breast cancer using a NI hypothesis. The idea behind creating this MAMS design was inspired by the Persephone trial (Earl et al. 2019) which was conducted at Warwick CTU and one of the numerous trials that followed the HERA trial in assessing various durations of Herceptin Therapy (Piccart-Gebhart et al. 2005).

Finally, chapter eight discusses all the issues raised throughout this research and summarises the findings from the thesis, addresses the limitations and the future work that could be undertaken.

2 Overview of Adaptive Design methodology

2.1 Introduction

The application of adaptive design methodology within a clinical trial allows for changes based on accrued information that was not available when the trial commenced. The FDA define an adaptive design as a study that includes prospectively planned adaptations (US Food and Drug Administration 2018). Various adaptive design methods have been developed and implemented such as: adaptive dose-finding, group sequential designs, sample size re-estimation, adaptive randomisation, seamless design etc.

This chapter describes some of the commonly considered adaptive design methods found in the literature. The categorisations of adaptive design methods are based upon categories specified within the book titled Cancer Clinical Trials: Current and Controversial Issues in Design and Analysis (George et al. 2016).

The focus of this PhD is on phase III cancer trials, thus adaptive dose-finding approaches/methods are only briefly discussed as these are more appropriate for early phase drug development.

2.2 Brief history of Adaptive designs

Adaptive design methods can be first traced back to the USA in 1933 whereby the methodology to modify the randomisation process to favour a more promising drug was developed (Thompson 1933). This method was implemented on ethical grounds; to reduce the exposure of patients to inferior treatments. Sequential sampling was first introduced by Stein (1945), where an initial sample size was calculated and the results of that sample size was used to calculated an additional sample size required to achieve the appropriate precision (Stein 1945). Thereafter, Wald (1947) implemented pre-defined stopping rules in this setting where if results were significant then sampling would stop (Wald 1947). The application of adaptive designs in the UK can be traced back to 1954 whereby a sequential trial was implemented to assess the use of calcium chloride against adrenaline as a bronchodilator agent (Kilpatrick et al. 1954). The results for each patient were assessed as soon as patients completed the intervention. The trial was stopped

after only four patients were assessed on the calcium chloride arm due to futility. Thereafter, Snell and Armitage (1957) and Smith (1958) provide early examples of the use of sequential trials (Snell et al. 1957, Smith 1958) and Todd (2007) and Bothwell et al (2018) provide a review of the literature and the developments of adaptive design methodology (Todd 2007, Bothwell et al. 2018).

In the past ten years, with the improvement in computer technology, adaptive design methodology and its application have increased, including the use of Bayesian techniques. The FDA initially released a draft guidance document for adaptive designs in 2010, where they categorised methods by 'Well-understood methods' and 'Less well-understood methods' (US Food and Drug Administration 2010). The latest draft guidance document has removed these categories and provided exemplars of clinical trials that illustrate the advantages of implementing adaptive designs which clearly demonstrate the development of these designs (US Food and Drug Administration 2018).

2.3 Adaptive dose-finding designs for identifying optimal dose

Phase I trials are conducted to identify a safe dose of a new drug to be given to humans. Adaptive dose-finding designs are usually applied within this phase whereby as the data is accumulated, the dose level is assessed against the toxicity level and decisions are made to either escalate, de-escalate or stay at the same dose (Zhang et al. 2006).

One particular adaptive dose-finding design is known as the continual reassessment method (O'Quigley et al. 1990, Garrett-Mayer 2006). The aim of this design is to find the maximum tolerated dose of a new drug and has shown to be more precise in comparison to other dose-finding methods (Wheeler et al. 2019). To successfully execute an adaptive dose-finding design using the continual reassessment method, it is essential that the following parameters are discussed and decided by all trial stakeholders:

- Which doses to study and how many doses are required;
- The target toxicity level;

- Clarify the dose-toxicity model; this model helps identify the probability of a patient being subject to any dose limiting toxicities;
- Clarity on how the trial data will be inferred for the dose-toxicity model;
- What decision rules to put in place to ensure that the maximum tolerated dose is reached quickly without overdosing patients;
- Planned sample size.

Once these parameters for the trial are determined, simulations are performed to assess the operating characteristics (Wheeler et al. 2019).

2.4 Group Sequential designs

The notion of sequential analyses has been defined as the repeated testing of hypothesis based on data that has accumulated over the course of a trial (Everitt et al. 2011). Initially, sequential analyses were performed after pairs of patients were allocated to each of the two treatment arms hence the number of analyses performed was subsequently high. Conversely O'Brien and Fleming, and Pocock developed designs whereby the number of analyses performed would be greatly reduced by performing analyses with groups of patients, hence these designs became popularly known as 'group sequential designs' (Pocock 1977, O'Brien et al. 1979, Jennison et al. 1999, Todd 2007). These group sequential designs have been developed to perform analyses at pre-specified time points also known as interims. At the interim, the accrued data can either indicate that there is no significant benefit of the new treatment (stopping for futility) or the new treatment is significantly beneficial (stopping for efficacy) (Pocock 1977, Todd 2007).

The application of group sequential methods are found in all four phases of clinical trials (Todd 2007, George et al. 2016). Initially within phase I trials, various doses are evaluated till the safe dose is found, this is subsequently evaluated within a phase II trial. Phase II trials are developed to assess whether a drug has an effect and to demonstrate thoroughly the efficacy of a drug. It is within phase II trials

that single-arm designs with multiple stages are implemented. The most common multiple stage designs are the minimax design, Simon's optimal two-stage design and flexible multi-stage designs (Simon 1989, Ensign et al. 1994, Chen 1997, Sargent et al. 2001). These single-arm phase II trials are followed by randomised phase II or phase III trials whereby the drugs are compared to the standard treatment or multiple treatments at multiple stages. It is within phase III trials that group sequential methods have been well developed and are well established (Todd 2007).

Although the focus of this PhD is phase III trials within the field of oncology, the VIETNARMS phase II/III trial for patients with hepatitis C has been included within this thesis as it demonstrates a group sequential type design, further details can be found in appendix 1. The VIETNARMS trial implements a sequential process whereby a Bayesian monitoring procedure has been used to allow the stopping of inferior arms early and hence successive patients are randomised to the remaining arms if there is 95% posterior probability of less than a 90% cure in any of the treatment arms (McCabe et al. 2020).

Group sequential methods can quicken the development of successful treatments and hence reduce the costs, resources and time taken to complete a trial. A retrospective analysis carried out for 72 cancer trials found that if group sequential methods were applied then approximately 80% of the trials would have stopped early (Rosner et al. 1989). Hence the FDA have stated that, "Early termination for efficacy should be generally reserved for circumstances where there is the combination of compelling ethical concern and robust statistical evidence" (US Food and Drug Administration 2018).

2.4.1 Stopping rules

In group sequential methods, the investigator can perform one or more interim analyses. Performing repeated interim analyses can inflate the false-positive-error rate or type I error rate, also known as the alpha value (α), due to simultaneously testing the same hypothesis multiple times, this is more commonly known as multiple hypothesis testing.

Therefore statistical stopping rules have been developed within group sequential designs to help control the type I error rate (Schulz et al. 2005). A graphical representation of a group sequential design can be seen in Figure 2.1, where over time, interim analyses are performed and if the test statistic is above or below a prespecified value then the trial can be stopped for either efficacy or futility.

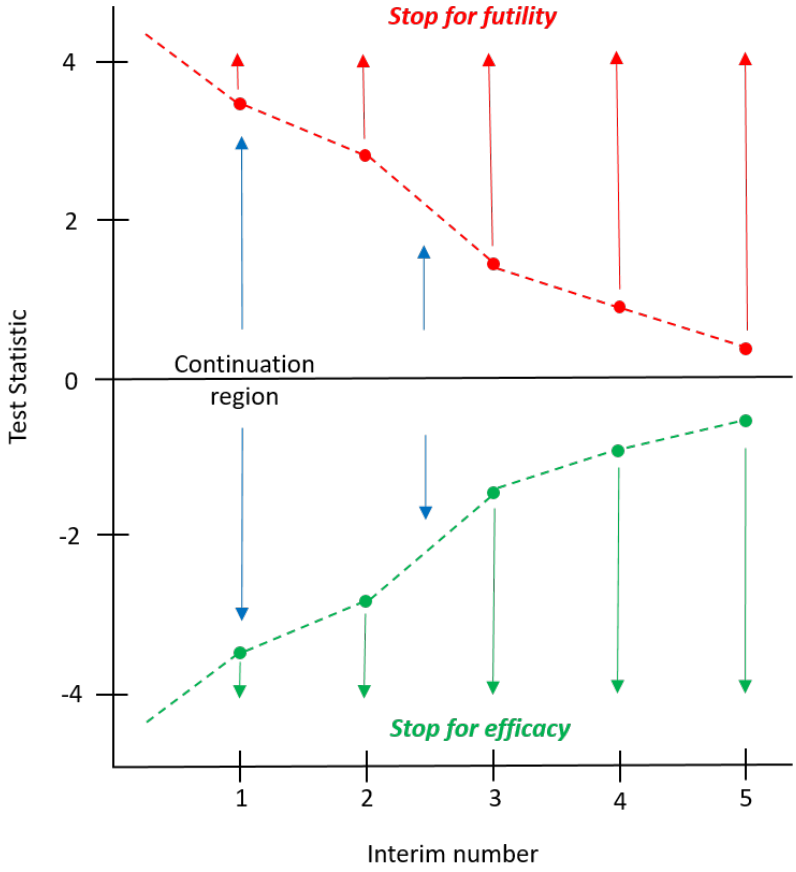

Figure 2.1: A graphical representation of a group sequential design.

One of these stopping rules incorporates the Haybittle-Peto boundaries (Haybittle 1971, Peto et al. 1977), whereby if at the interim analysis, the p-value is less than or equal to 0.001 then the trial is stopped early. This procedure continues at each interim stage up until the final stage, where the p-value is evaluated at the 5% significance level (Figure 2.2). Implementing a 5% significance level has made it easier for professionals from a non-statistical background to interpret the results, however it has been argued that this method is too stringent in deciding to stop a trial (Schulz et al. 2005).

Unlike the Haybittle and Peto stopping boundary, the O'Brien and Fleming stopping rule changes the significance level at each stage (O'Brien et al. 1979). Initially the stopping boundaries are conservative within the O'Brien and Fleming method, which is appealing to trial stakeholders, as at the initial stages of a trial the results are not stable but stabilise over time as more data is acquired (Figure 2.2). At each subsequent interim analysis, the stopping boundaries become narrower as more data is accumulated.

The Pocock approach (Figure 2.2) uses a fixed significance level and time for all interim and final stages calculated based on the number of stages altogether (Pocock 1992). This approach does not consider the potential instability of the data at the beginning of a trial by not having a conservative p-value in comparison to the aforementioned two methods. Furthermore, compared to a regular parallel trial with a final p-value of 0.05, the Pocock method would allow a less significant final p-value hence resulting in two different sets of results for two different trial designs.

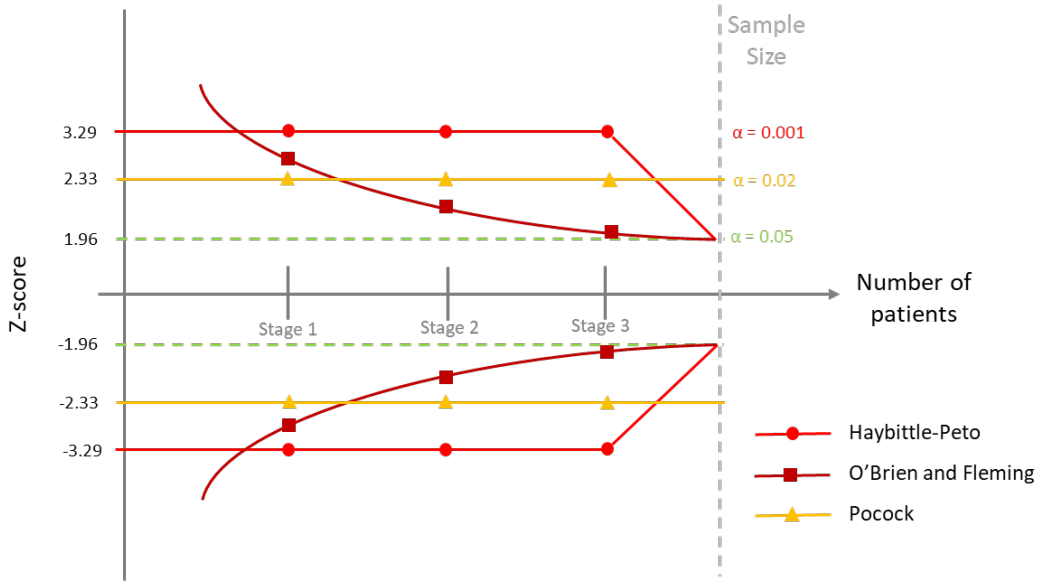

Figure 2.2: Graphical representation of different stopping rules that can be applied within group sequential designs.

Unlike the Pocock approach, where the interim stages are equally spaced, the Lan and DeMets approach allows for some flexibility for the frequency and timing of these interim stages (Gordon Lan et al. 1983). This approach has given additional

flexibility to group sequential designs by using stopping boundaries defined by an alpha spending function.

The triangular test proposed by Whitehead and Stratton (1983) not only applies a group sequential design but does not require pre-specification of when the interim analyses will be conducted and allows for continuous monitoring (Whitehead et al. 1983, Whitehead 1997). This approach has finite convergent boundaries giving an asymmetrical triangular continuation region (Figure 2.3). A trial is allowed to continue if it stays within the triangular region and the conclusion of the trial is dependent on which boundary is crossed, i.e. if the upper boundary is crossed then experimental arm is superior to control and if lower boundary crossed than the experimental arm is equal to or non-inferior to the control.

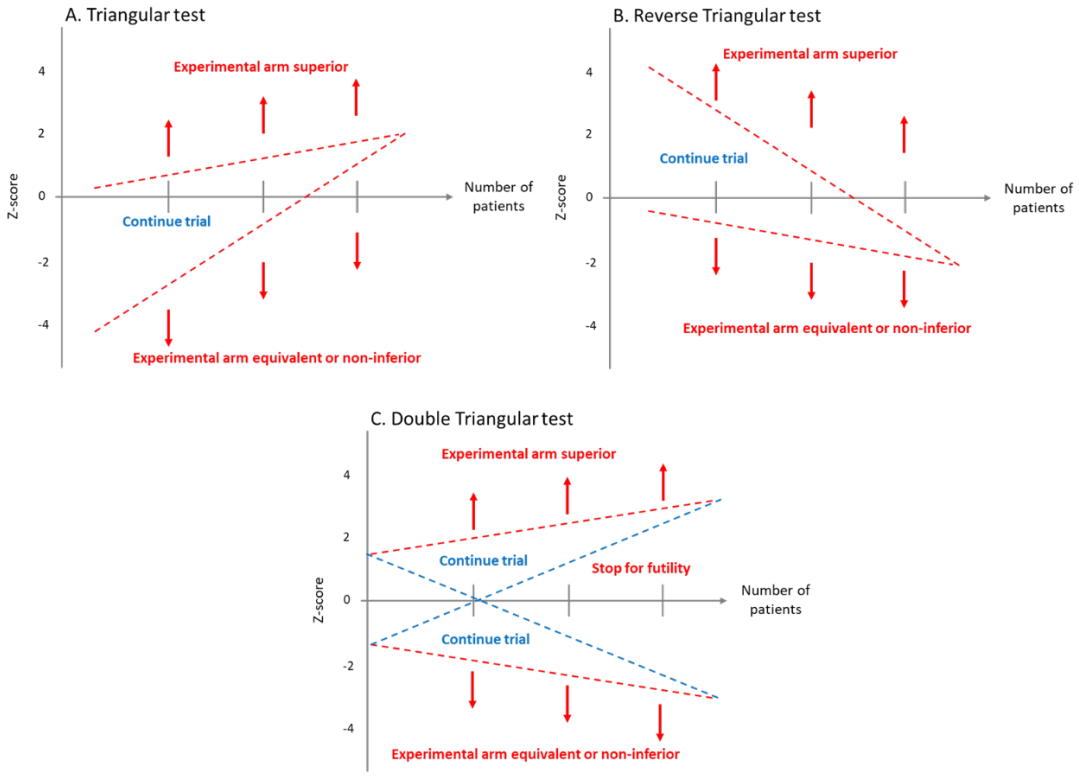

Figure 2.3: Graphical representation of the different triangular test proposed by Whitehead et al (1983)

The reverse triangular test (Figure 2.3) can be implemented if a NI hypothesis has been implemented. A double triangular test (Figure 2.3) is a combination of the triangular test and reverse triangular test which is used to detect either superiority or NI of an experimental arm. Unlike the triangular test, which stops the trial if

either the upper or lower boundary is crossed, the double triangular test continues the trial so definitive conclusion can be made of whether the experimental arm is equal, superior or non-inferior to the control arm.

2.5 Combination test and conditional error function approach

The combination test and the conditional error function are used when combining data from different stages and when there are one or more looks at interim data. These designs have been commonly known as a more flexible version of the group sequential design (Bauer et al. 2016, Pallmann et al. 2018).

If a trial has M pre-planned interim stages, each stage has a null hypothesis to test the treatment effect. The combined test is a combination of the null hypothesis at all stages against the alternative hypothesis. Changes in the primary endpoint or patient eligibility criteria can lead to different null hypothesis being tested at the various stages. Thus, one way to test the final null hypothesis taking the results of the interim null hypotheses into consideration is to combine the p-value (or Z-values) at stage m, denoted $P_m\ (Z_m)$, where $1 \leq m \leq M$ for which the two most popular methods are:

1. Inverse χ^2 (Fisher 1992), which rejects the null hypothesis if $-2\log(P_1 \times P_2 \times \ldots \times P_M) > \chi^2_{2M}\ (\alpha)$, where $\chi^2_{2M}\ (\alpha)$ is the upper tail point of the χ^2_{2M} distribution;
2. Weighted inverse normal (Mosteller et al. 1954, Becker 1994, Goutis et al. 1996) uses Z-values for which the null hypothesis is rejected if $w_1 Z_1, \ldots w_M Z_M > z(\alpha)$ where $w_1, \ldots w_M$ are pre-specified weights such that $\sum w_i^2 = 1$ and $z(\alpha)$ is the upper tail point of the standard normal distribution.

The conditional error approach allows for modifications to the design of the trial at any time during the trial if the new conditional error given the data observed does not exceed the original conditional error. Various efforts have been made to find the best conditional error approach that for a given power, minimises the sample size (Liu et al. 2001, Brannath et al. 2002).

The conditional error approach and the combination test are similar as they both allow flexibility with regards to the number of interim looks and changes in primary endpoint (Bauer et al. 2006).

2.6 Multi-arm multi-stage (MAMS) designs

In the clinical trial process, there can often be multiple promising treatments for a specific population; these treatments can be combinations of different drugs, different dosages of the same drug or different methods of administering a drug. Implementing MAMS designs can be an effective solution as these designs have the potential to evaluate several experimental arms against a single control arm and have the potential of accelerating treatment evaluation (Parmar et al. 2008, Royston et al. 2011, Wason 2015, Quartagno et al. 2018). Multiple stages can increase the efficiency in evaluating treatments as it allows for the elimination of ineffective treatments early (Barthel et al. 2009). The focus of this research is the MAMS design framework by Royston et al (2011). This type of MAMS design can utilise an intermediate outcome which can be a surrogate of the primary outcome and is used to assess the efficacy of the experimental arms at interim stages. At each stage, an interim analysis is performed whereby pairwise comparisons are made using an intermediate outcome for each of the experimental arms against the control arm. The results of these interim analyses will determine whether an experimental arm should be stopped for futility (the experimental arm has showed no benefit) or lack of benefit (showed benefit but not enough to progress to the next stage) or progress to the next stage. During an interim analysis, recruitment will continue to ensure that trials move seamlessly from one stage to the next. At the final stage, pairwise comparisons are made based on the primary outcome to decide upon the superiority of the remaining experimental arm(s) against the control.

A fundamental element to consider when implementing a MAMS design for this framework is the overall type I error rate (Bratton et al. 2016), of which there are two measures:

- Familywise type I error rate (FWER) is defined as the probability of rejecting at least one null hypothesis when these hypotheses are true for a given family of hypotheses for the primary outcome if the null hypothesis is true. The FWER is more necessary in multi-arm settings when controlling the type I error rate as a whole rather than at each pairwise comparison. One approach to calculate the FWER, discussed in section 2.6.7, is similar to the Dunnett's approach to compare multiple treatments against a control using a multiple comparison procedure (Dunnett 1955). Strong control of the FWER would apply when assessing the same drug for different durations.
- Pairwise type I error rate (PWER) is defined as the probability of recommending an ineffective experimental arm against the control arm for the primary outcome irrespective of all other experimental arms. The Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy (STAMPEDE) trial is a MAMS platform trial (see chapter four for more information) which emphasised control of the PWER because distinct hypotheses are being tested in each of the experimental arms.

MAMS designs have been applied in several phase III cancer trials (James et al. 2008, Shiu et al. 2013, Mehanna 2015). The subsequent sections will further explain the methodology implemented for this framework.

2.6.1 Nstage methodology

The MAMS methodology specified by Royston et al (2011) is available within Stata using the command called 'nstage' (Barthel et al. 2009, StataCorp. 2017, Blenkinsop et al. 2019). In the following sections, the theory behind this MAMS design is described.

2.6.2 Outcomes and hypothesis

For a MAMS design, let IO and PO be defined as the time-to event intermediate outcome and primary outcome respectively, where $IO = PO$ can also be a possibility.

Let $j \geq 1$ denote the number of experimental arms for which pairwise comparisons will be made against a common control arm, C. The true HR, Δ_{ij} is the effect of the experimental arm, j, on the control arm on the specified outcome at stage i, where i is the i^{th} stage for $i = 1, \ldots f$ where f is the final stage. The true HR is calculated using the IO from stage 1 to stage $f - 1$ and PO at stage f. It is assumed that the proportional hazards assumption holds for all treatment comparisons. The null and alternative hypotheses can be defined as:

$$H_{ij}^0: \Delta_{ij} = \Delta_{ij}^0,$$
$$H_{ij}^1: \Delta_{ij} = \Delta_{ij}^1, \qquad i = 1, \ldots, f, j \geq 1$$

A test of the accumulated data for each pairwise comparison is performed at stage i with nominal type I error rate α_i and power ρ_i. If this is statistically significant then the experimental arm continues onto the next stage up until stage $f - 1$, and at stage f superiority is concluded.

Under H_0 and H_1, the estimated log HR at stage i defined as $\ln \hat{\Delta}_i$ is distributed as:

$$H_0: \ln \hat{\Delta}_i \sim N(\ln \Delta_i^0, v_i^0)$$

$$H_1: \ln \hat{\Delta}_i \sim N(\ln \Delta_i^1, v_i^1)$$

where the estimated variances under H_0 and H_1 are v_i^0 and v_i^1 respectively. Therefore, the probability under the null or alternative hypothesis of reaching stage $i + 1$ or concluding superiority of either the control or experimental arm are based on normal approximations such that the appropriate critical HR and events are given by the multivariate tail areas.

The one-sided significance levels for each of the hypothesis tests can be thus related to the critical values for the test as follows:

$$\begin{aligned}\alpha_i = \quad & \Pr\left(\ln\hat{\Delta}_i < \ln\delta_i \middle| H_0\right) \\ & \Pr\left(\frac{\ln\hat{\Delta}_i - \ln\Delta_i^0}{\sigma_i^0} < \frac{\ln\delta_i - \ln\Delta_i^0}{\sigma_i^0} \middle| H_0\right) \\ & \Phi\left(\frac{\ln\delta_i - \ln\Delta_i^0}{\sigma_i^0}\right) \\ & \Phi\left(z_{\alpha_i}\right)\end{aligned}$$

where σ_i^0 is defined as square root of v_i^0, the critical HR at stage i is defined as δ_i and $\Phi(\,)$ is the standard normal distribution function.

Likewise, the power for each of the hypothesis tests is such that:

$$\begin{aligned}\rho_i = \quad & \Pr\left(\ln\hat{\Delta}_i < \ln\delta_i \middle| H_1\right) \\ & \Pr\left(\frac{\ln\hat{\Delta}_i - \ln\Delta_i^1}{\sigma_i^1} < \frac{\ln\delta_i - \ln\Delta_i^1}{\sigma_i^1} \middle| H_1\right) \\ & \Phi\left(\frac{\ln\delta_i - \ln\Delta_i^1}{\sigma_i^1}\right) \\ & \Phi\left(z_{\rho_i}\right)\end{aligned}$$

where σ_i^1 is defined as square root of v_i^1.

2.6.3 Critical HR

The critical HRs are calculated for each stage to determine whether the experimental arm continues to the next stage $i+1$ or if superiority is concluded at the final stage when $i = f$. The critical HR is denoted by δ_i where i is the i^{th} stage where $i = 1, \ldots, f$, is calculated by:

$$\ln\delta_i = \ln\Delta_i^0 + \sigma_i^0 z_{\alpha_i} = \ln\Delta_i^1 + \sigma_i^1 z_{\rho_i}$$

where z_{α_i} and z_{ρ_i} are the pre-specified normal equivalent deviate for a one-sided significance and power respectively. The σ_i^0 and σ_i^1 are calculated using the

variance which are approximately the same under the null and alternative hypothesis (Tsiatis 1981) as follows:

$$v_i^0 = v_i^1 = \frac{1}{e_i} + \frac{1}{\hat{e}_i}$$

where e_i and $\hat{e}_i$ are the number of control arm events and experimental arm events respectively for each arm at stage i.

2.6.4 Number of control arm events

The event rate in the control arm governs the time at which the interim and final analyses takes place. The number of control arm events at each stage i can be calculated using the following equation:

$$e_i = \frac{2\left(z_{\alpha_i} - z_{\rho_i}\right)^2}{\left(\ln\ \Delta_i^1 - \ln \Delta_i^0\right)^2}$$

Using the number of control arm events at each stage, a numerical search is performed to obtain the sample size that will give the required power at each stage (Royston et al. 2011), and hence the trial duration, assuming that the event times required for the IO and PO are exponentially distributed (see section 1.3).

The rate at which patients are recruited into the control arm of the trial per unit time during stage i, is denoted by a_i and it is assumed that recruitment occurs at a uniform rate in each given stage i.

The period of stage i, can be defined by l such that:

$$l_i = t_i - t_{i-1}$$

where $i = 1, \ldots, f$ and $t_0 = 0$. Therefore, the number of patients recruited during stage i into the control arm is defined by:

$$n_i = a_i l_i$$

The number of patients recruited into one experimental arm during stage i is also n_i. Therefore, on the condition that all experimental arms, j, where $j \geq 1$,

continue past $f-1$ stages, the total number of patients required is calculated by the following:

$$n = (1+j)\sum_{i=1}^{f} a_i l_i$$

In practice, trial stakeholders may wish to set a period for recruitment, $\hat{t}$ where $\hat{t} < t_f$ and follow-up patients until the required number of events are observed. The only practical option is for $\hat{t}$ to occur between stage $f-1$ and stage f because ending recruitment prior to an intermediate stage would eliminate any possibility of ending recruitment to experimental arms early based on the assumption that at least one experimental arm reaches the final stage.

However, there may be a situation where all experimental arms stop prior to stage $f-1$, where $f \geq 3$, due to futility or lack of benefit resulting in ending recruitment earlier than anticipated and reporting the outcome earlier than planned.

If $\hat{t}$ has been specified, then $\hat{l} = \hat{t} - t_{f-1}$. Hence, patients are followed up until the required number of control arm events are observed. Therefore, the number of patients required can be calculated by:

$$n = (1+j)\left[a_f \hat{l} + \sum_{i=1}^{f-1} a_i l_i\right]$$

2.6.5 Significance level and power at each stage

Within the MAMS framework, the value of α_i directly influences the number of control arm events required at each stage and hence the stage duration. The MAMS framework recommends using descending values of α_i and larger than the usual significance values at the interim stages to allow for decisions to be made early on in the trial as to whether to drop experimental arms due to futility or lack

of benefit or continue to the next stage. The effect of using large significance values has been compensated by having a high power at each stage. For example in the STAMPEDE trial, at stages one to four the significance levels are 0.5, 0.25, 0.1 and 0.025 respectively and the power value at each of these stages are 95% for stages one to three and 90% at stage four[1] (Sydes et al. 2012).

2.6.5.1 Stagewise significance level and power

An experimental arm reaches stage i with pairwise significance level, α_i and power, ρ_i on the condition that it has passed stage $i-1$ which implies that the experimental arm has passed previous stages $i-2,\dots,1$. Therefore, $\alpha_{i|i-1}$ is denoted as the probability at stage i under H_0 of rejecting H_0 given that the experimental arm has passed stage $i-1$. In the same way, the stagewise power is the probability at stage i under H_1 of rejecting H_0 on the condition that it has passed stage $i-1$.

Therefore, based on the rules of conditional probability,

$$\alpha_{i|i-1} = \frac{\Phi_i\left(z_{\alpha_1}, \dots z_{\alpha_i}; B^{(i)}\right)}{\Phi_{i-1}\left(z_{\alpha_1}, \dots z_{\alpha_{i-1}}; B^{(i-1)}\right)}$$

$$(\rho)_{i|i-1} = \frac{\Phi_i\left(z_{(\rho)_1}, \dots z_{(\rho)_i}; B^{(i)}\right)}{\Phi_{i-1}\left(z_{(\rho)_1}, \dots z_{(\rho)_{i-1}}; B^{(i-1)}\right)}$$

Where Φ_i is the i-dimensional standard multivariate normal distribution function with mean 0 and variance matrix $B^{(i)}$, where this is the matrix containing the first i rows and columns, for the standardised test statistics at the first i interim analyses, with appropriate correlations. The overall PWER is given by the equation above when $i = f$.

Using the Royston et al (2011) framework when $IO \neq PO$, the maximum FWER occurs when all experimental arms are effective at each interim stage for the IO but ineffective on the PO (Bratton et al. 2015). This effectively results in the MAMS

[1] http://www.stampedetrial.org

design reducing to a one-stage trial as the interim stages become redundant and therefore $PWER_{max} = \alpha_f$.

2.6.6 Overall significance level and power

Once the stagewise significance level and power have been defined, the overall significance level, α, and power level, ρ for when IO and PO are the same can be defined by:

$$\alpha = P\left(\widehat{\Delta}_1 < \delta_1, \dots., \widehat{\Delta}_f < \delta_f | H_0\right)$$

$$\alpha = \Phi_f\left(z_{\alpha_1}, \dots, z_{\alpha_f}; B\right)$$

$$\rho = P\left(\widehat{\Delta}_1 < \delta_1, \dots., \widehat{\Delta}_f < \delta_f | H_1\right)$$

$$\rho = \Phi_f\left(z_{\rho_1}, \dots, z_{\rho_f}; B\right)$$

where it is assumed that $\left(\ln \widehat{\Delta}_1, \dots \ln \widehat{\Delta}_f\right)$ is a multivariate normal distribution and $\Phi_f(; B)$ is the standard f-dimensional multivariate normal distribution. B is the correlation matrix where the i^{th} and k^{th} element B_{ik} of $B(i, k = 1, \dots, f)$ is the correlation between the log hazard ratios, $\ln \widehat{\Delta}_i$ and $\ln \widehat{\Delta}_k$, for the outcome at the end of the i^{th} and k^{th} stages respectively.

However, if IO and PO differ, the overall significance level and power for the combined $(f-1)$ stages is defined as:

$$\alpha_{IO} = \Phi_{f-1}\left(z_{\alpha_1}, \dots, z_{\alpha_{f-1}}; B^{f-1}\right)$$

$$\rho_{IO} = \Phi_{f-1}\left(z_{\rho_1}, \dots, z_{\rho_{f-1}}; B^{f-1}\right)$$

Hence for the value of B_{if}, the lower limit, l, and upper limit, u of the overall type I error and overall power can be calculated as follows:

$$\alpha_l = \alpha_{IO}\alpha_f, \alpha_u = \min\left(\alpha_{IO}\alpha_f\right)$$

$$\rho_l = \rho_{IO}\rho_f, \rho_u = \min\left(\rho_{IO}\rho_f\right)$$

When there is 100% correlation between $\ln \widehat{\Delta}_l$ and $\ln \widehat{\Delta}_f$ the minima occurs, i.e. $B_{if} = 1 \; \forall \; i$. When there is no correlation between $\ln \widehat{\delta_l}$ and $\ln \widehat{\delta_f}$ the maxima occurs, i.e. $B_{if} = 0 \; \forall \; i$.

It is important to note that within this MAMS design framework, the overall values of α and ρ are not required to make decisions at the interim stages or final stage to conclude superiority (Royston et al. 2011), however these overall values can lead one to change the selected values at α_i and ρ_i.

2.6.7 FWER

The probability of passing all i stages for one or more unsuccessful treatments is known as the familywise error rate (FWER). Within the multi-arm setting, it may be more necessary to assess the overall (FWER) type I error rate instead of the PWER . Furthermore, strong control of the FWER has been recommended more for confirmatory settings (phase III) in comparison to exploratory MAMS trials (phase II) (Wason et al. 2016).

The Dunnett's test is a procedure to compare many to one comparisons, hence using the Dunnett test (Dunnett 1955), $FWER_{max}$ can be calculated as follows:

$$FWER_{max} = 1 - \Phi_J\left(z_{1-\alpha_f}, \dots z_{1-\alpha_f}; R\right)$$

where R is the $J \times J$ correlation matrix between arms with off-diagonal entries equal to 0.5.

2.6.8 Example - applying the 'nstage' command

To demonstrate the application of the 'nstage' command by Barthel et al (2009), a sample size calculation was performed for a hypothetical four arm three stage MAMS trial for patients diagnosed with colon cancer. OS was used for both the IO and PO where patients had a five-year OS rate of 50.5% and the trial aimed to detect a 7% difference between the control arm and the experimental arm. It was hypothesized that one experimental arm would not significantly differ from the control arm at each interim stage resulting in an experimental arm being dropped at each stage. The stagewise significance values and stagewise power values were selected based on values suggested by Royston et al (2011) and initially implemented in the STAMPEDE trial (see Table 4.2). The parameters inputted into the 'nstage' command in Stata can be found in Table 2.1.

Nstage command	Parameter	Value
nstage	Total number of stages	3
arms	Number of arms in each stage	4, 3, 2
accrue	Rate per unit time that patients enter the trial during each stage	625, 625, 625
aratio	Allocation ratio	1
s	Survival Probability for OS	0.505
t	Survival Time (years)	5
hr0	HR under H0 specified for intermediate and final outcome	1, 1
hr1	HR under H1 specified for intermediate and final outcome	0.81, 0.81
alpha	One-sided alpha for each stage	0.5, 0.25, 0.025
omega	Power for each stage	0.95, 0.95, 0.9
tunit	Time units	1 (= one year)
tstop	Time of recruitment stopping (years)	6

Table 2.1: Parameters inputted for a hypothetical four arm three stage MAMS trial using the 'nstage' command in Stata, the intermediate and primary outcome was OS. OS – Overall survival, HR – Hazard ratio, H0 – Null hypothesis, H1 – Alternate hypothesis.

It can be seen from the sample size calculation in Figure 2.4 that 3750 patients would be required to be recruited over six years and followed up for approximately two years. The first interim analysis would take place when 134 control arm events are obtained at approximately 3.8 years. The second interim analysis would take place when 258 control arm events are obtained at approximately 5.4 years and the final analysis would take place when 489 control arm events are obtained at 7.8 years. The critical HR at stages one and two are 1 and 0.942 respectively; at these interim stages pairwise comparisons are made for each experimental arm against the control arm; if the obtained HR is less than the critical HR then that experimental arm would continue to the next stage. The critical HR at the final stage is 0.882, it was assumed that one experimental arm would reach this stage and if the obtained HR was less than the critical HR then superiority of that experimental arm over the control is declared.

```
. nstage, nstage(3) accrue(625 625 625) arms(4 3 2) alpha(0.5 0.25 0.025) hr0(1 1) hr1(0.81 0.81)
> omega(0.95 0.95 0.90) t(5 5) s(0.505 0.505) aratio(1) tunit(1) tstop(6)

n-stage trial design                    version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 5.1 time units

Operating characteristics
------------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
------------------------------------------------------------------------------
1           0.5000    0.950    1.000    0.810    1.000    3.853    3.853
2           0.2500    0.950    1.000    0.810    0.942    1.580    5.433
3           0.0250    0.900    1.000    0.810    0.882    2.380    7.814
------------------------------------------------------------------------------
Pairwise Error Rate          0.0218             Pairwise Power   0.8584
Familywise Error Rate (SE)   0.0555 (0.0005)
------------------------------------------------------------------------------
Note: patient accrual stopped at time  6.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ----------Stage 1----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      625      156      469
Patients*     2408      602     1806
Events**       470      134      336
           ----------Stage 2----------
           Overall  Control   Exper.
Arms             3        1        2
Acc. rate      625      208      417
Patients*     3396      931     2465
Events**       694      258      436
           ----------Stage 3----------
           Overall  Control   Exper.
Arms             2        1        1
Acc. rate      625      313      313
Patients*     3750     1108     2642
Events**       909      489      420
           ------------------------------
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE
```

Figure 2.4: Syntax and output obtained when implementing the 'nstage' command in Stata for a four arm three stage trial with a superiority hypothesis in patients diagnosed with colon cancer.

2.7 Sample size re-estimation

A suitable sample size calculation that has an appropriate power and significance level is essential to achieving the objectives of a trial. Adjusting the sample size as an adaptive design method allows for the re-calculation of the sample size based on observed data during an interim analysis. The European Medicines Agency (EMA) Committee for Medicinal Products for Human use states that "Blinded sample size reassessment that properly control the type I error should be used, especially if the sole aim of the interim analysis is the re-calculation of sample size" (CHMP 2012). However, depending on the reason, i.e. investigating a drug in a single arm trial in an unknown population, unblinded methods may need to be considered. For example, the HYPAZ trial aimed to explore the effect of a new cancer drug in an unknown population (Bond et al. 2011). Hence, an adaptive sample size re-estimation design was implemented due to this limited knowledge at the design stage.

The EMA imply that sample size re-estimation should be prospectively planned, yet unplanned methods have been created whereby the sample size can be adjusted even though these were not planned at the start of the trial (Bauer et al. 1994, Proschan et al. 1995, Fisher 1998). Unplanned sample size adjustments occur within clinical trials, with many of these adjustments due to changes to the on-going protocol. They could be based on characteristics of the accrued data at an interim stage (Chuang-Stein et al. 2006, Chow et al. 2011). For example, a randomised trial comparing adjuvant imatinib versus no further therapy for patients that have undergone surgery for localised, high or intermediate risk of gastrointestinal stromal tumour (Casali et al. 2015) involved an ad-hoc adaptation of the sample size. The sample size for the trial increased from 400 patients to 900 patients due to a larger sample of patients being recruited with low and intermediate risk tumours than expected, which resulted in a higher than expected survival rate in the control group. Furthermore, the VICTOR trial (see section 4.2) which assessed the use of rofecoxib in colorectal cancer patients, required 7000 patients to detect an increase in survival to detect a relative difference of 20% with 85% power and using a 5% two-sided significance interval

(Midgley et al. 2010). The recruitment of patients to the trial was stopped early as the drug being used in the trial was withdrawn completely due to safety concerns. This resulted in the VICTOR trial team re-evaluating the statistical analysis plan to follow-up the 2434 patients to detect an increased relative difference of 25% in survival with 87% power using a two-sided 5% significance level rather than detecting the original 20% difference with a much lower power. These unforeseen changes required the trial team to act rapidly to ensure that this drug was recalled and no longer being consumed. This example demonstrates that unforeseen circumstances can arise which can force investigators to report using a much reduced sample size. Pragmatically speaking this adaptation is classed as an adaptive design method, however it would not be classed as an adaptive design in accordance to the guidelines provided by the FDA (US Food and Drug Administration 2018) as it was undertaken retrospectively and not pre-planned.

Re-estimating the sample size during interim stages can assist a trial in various ways. For example, if there is limited information about the population in question, then an internal pilot study could be conducted whereby the sample size can be re-estimated based on the information accrued at the interim analysis. Birkett and Day (1994) have developed a method of using an internal pilot study to help calculate the final sample size. Usually sample size calculations are based on external data which may differ in terms of patient criteria or different treatment duration hence they may not truly represent the trial being designed. Birkett and Day‘s approach is useful as the final sample size is specified based on the results of the internal pilot study compared to other methods which calculate the pre-planned final sample size and only allow for an increase in sample size (Wittes et al. 1990, Birkett et al. 1994, Gould 1995).

Sample size re-estimation can be extended to allow implementation within fully sequential, group sequential, and dose response designs (Bhatt et al. 2016). Sample size re-estimation methods are implemented to increase the sample size (Jennison et al. 2015). However, decreasing the sample size can be accomplished if it is within a group sequential design setting with thorough planning to ensure the type I error rates are controlled.

The following sections will discuss some of the different methods to re-estimate sample size.

2.7.1 Internal Pilot studies

Internal pilot studies can be executed prior to a main trial to help calculate sample size if there are any unknown parameters. These unknown parameters can be initially estimated to calculate the sample size and then after the accrual of data, the sample size can be re-estimated at an interim analysis.

One simple approach is when there is a single analysis of a primary outcome measure. Given that the hypotheses, type I and type II error rates have all been specified then the sample size can be expressed as the function $f(\theta)$, where θ is an unknown parameter. The internal pilot study begins with an estimate for the unknown parameter, θ_1 and hence with an initial sample size of $f(\theta_1)$. After the accrual of some data, a pre-planned interim analysis is performed whereby a new estimate, θ_2 is found. This value is utilised to re-estimate the sample size and hence the trial continues with the aim to reach a sample size of $f(\theta_2)$ and it is with this final sample size that all data is analysed.

2.7.2 Proschan and Hunsberger method

The Proschan and Hunsberger method (Proschan et al. 1995, Posch et al. 1999) implements a two-stage design whereby decisions related to the sample size are made after the first stage based on conditional power. This method uses the significance of the treatment difference within the first stage to determine how many additional observations are required and the critical value to use at the end of the study.

This method was implemented in a RCT assessing the effectiveness of phenytoin in preventing seizures in patients with brain tumours (De Santis et al. 2002). This trial aimed to detect a decrease in the incidence of seizures from 15% to 5%. To achieve this, a target sample size of 356 patients was required with 80% power and 5% significance level. The trial performed a pre-planned interim analysis after 200 patients were recruited at which the incidence of seizures in the control arm was 11% and 13% in the treatment group. This resulted in a conditional power of approximately 27%, which meant that there would be less than a 27% probability

to detect the hypothesised difference between the two groups. Based on this interim result, recruitment was stopped.

In contrast to the Proschan and Hunsberger method, the MAMS designs uses large pre-planned type I error values with fixed treatment difference(s) which is compensated by having a high power at each stage to help make decisions with regards to stopping or continuing experimental arms.

2.7.3 Cui, Hung and Wang method

The Cui, Hung and Wang method of sample size re-estimation investigates the use of observed data within group sequential clinical trials. This testing procedure preserves the type 1 error rate and may provide a considerable increase in power together with an increase in sample size (Cui et al. 1999).

The VALOR trial was a phase III RCT that investigated vosaroxin plus cytarabine in patients diagnosed with leukaemia. This trial implemented the Cui, Hung and Wang sample size re-estimation method based on the number of events. A total of 450 patients with 375 deaths were initially required to detect an improvement in median survival from 5 months (placebo) to 7 months (vosaroxin plus cytarabine) with 90% power and 5% two-sided significance level. Pre-specified statistical guidelines were set out for the Data and Safety Monitoring committee (DMC) to follow, based on the Cui, Hung and Wang sample size re-estimation method. The pre-planned decision criteria to be made at the interim analysis were to either stop the trial or increase to 562 deaths (approximately 675 patients). The pre-planned interim analysis was performed when 173 deaths were reached. The DMC followed the pre-specified guidelines and made the decision to increase the sample size to 562 deaths. The independence of the recommendation from the DMC and the use of pre-specified criteria for the amendment to the sample size ensures the credibility of the study.

2.7.4 Brutti, De Santis, and Gubbiotti method (Bayesian)

The Bayesian predictive approach for sample size re-estimation by Brutti et al (2009) considers a mix of prior information for the quantity of interest (De Santis 2006, Brutti et al. 2009). This method is an extension to the proposed methods by Gajewski and Mayo, and Wang (Gajewski et al. 2006, Wang 2007).

This method uses the predictive approach to choose the data probability distributions for the posterior calculations. For this method two distinct prior distributions are calculated, a design prior and an analysis prior. The design prior models the uncertainty on the parameter values at the planning stage and the analysis prior combines the data obtained with information collected prior to the study. At the interim stages, the weights of the prior distributions are updated which result in the sample size being re-estimated using the method by De Santis (2006).

2.8 Adaptive randomisation

Randomisation is an essential procedure that is used within clinical trials to ensure that there is an equal balance of allocated treatments across patients. Furthermore, randomisation can ensure that the allocation of treatments can be balanced across specified baseline characteristics by employing stratification variables. There are two main types of adaptive randomisation methods; outcome-adaptive randomisation and covariate-adjusted randomisation, which can also be combined to form covariate-adjusted outcome-adaptive randomisation (George et al. 2016). Covariate-adjusted randomisation uses baseline demographics associated with prognostic features for each patient to assign them to relevant treatment groups. Many specific procedures have been developed to implement covariate adjustment of which two of the most popular methods are stratified block randomisation and minimisation (Kahan et al. 2012). Adaptive randomisation methods can be applied to outcomes that are binary, ordinal or continuous and can be applied to studies with multiple treatment arms.

2.8.1 Outcome-adaptive randomisation

The FDA define outcome-adaptive randomisation as “a form of treatment allocation in which the probability of patient assignment to any particular treatment group of the study is adjusted based on repeated comparative analyses of the accumulated outcome responses of patients previously enrolled” (US Food and Drug Administration 2010, US Food and Drug Administration 2018). This method can be used to allocate patients to the treatment that suggests better outcomes, alternatively it can reduce the number of patients allocated to a

treatment that suggests that there may be a higher probability of patients having an adverse event. Implementing this method in trials can motivate patients to enrol in latter stages as it may increase their chances of being allocated to a 'more successful treatment'. The practicality of implementing outcome-adaptive randomisation methods may not be suitable for trials that either have a long-term outcome or treatment duration, or trials with a large sample size (Chow et al. 2008). The allocation could be dependent upon the outcome of the previous participant; thus, the need to wait until treatment is complete or the outcome obtained before implementing adaptive randomisation would significantly increase the completion time of the study.

A frequentist approach to outcome-adaptive randomisation developed by Hu and Rosenberger demonstrate various procedures when comparing two or more treatment arms when the outcome is binary (Hu et al. 2003, Hu et al. 2006).

Hu and Rosenberger consider the case of a trial with two treatment arms denoted i with a binary outcome. The response rates for treatment one and two are denoted by p_i where $q_i = 1 - p_i$ for $i = 1, 2$. The optimal randomisation probability (ε_1) is calculated to capitalise on the power to compare the response rates by the equation:

$$\varepsilon_1 = \frac{\sqrt{p_1 q_1}}{\left(\sqrt{p_1 q_1} + \sqrt{p_2 q_2}\right)}$$

As the trial progresses, the optimal randomisation probability can be changed based on accrued data at interim analysis.

An example of an outcome-adaptive randomisation trial that adjusts for covariates can be found in the Stroke Hyperglycemia Insulin Network Effort (SHINE) trial which aimed to give patients different amounts of insulin in hyperglycaemic acute ischemic stroke patients (Bruno et al. 2014). This two-arm trial aimed to ensure there was a balance in important prognostic factors and encourage patients to be randomised to the better performing treatment. Therefore, initially a 1:1 allocation was employed but as the trial progressed the ratio changed to either

ensure baseline prognostic factors remained balanced or give more weight to the treatment group with the better outcome. The primary analysis for this trial assessed the proportion of patients in each treatment group with a favourable outcome after controlling for the key baseline prognostic factors.

2.8.2 Bayesian adaptive randomisation

There are various methods to perform adaptive randomisation using Bayesian methods (Hardwick et al. 1991, Thall et al. 2007, Lee et al. 2010). A common method applied is simple Bayesian adaptive randomisation (George et al. 2016), for example in a two arm study, patients are assigned to treatment one with probability ε_1 if the following equation is satisfied at time of randomisation:

$$\varepsilon_1 = Prob(\lambda_1 > \lambda_2),$$

where λ_i is the posterior probability for the treatment arms i where $i = 1, 2$.

One example of where Bayesian adaptive randomisation has been used is in the Biomarker-integrated approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE) trial. Patients with advanced nonsquamous non-small-cell lung carcinoma (NSCLC) that had been treated with chemotherapy and then experienced disease relapse were eligible for the trial (Liu et al. 2015). Patients were randomised to four different treatment arms based upon their biomarker profiles. A Bayesian adaptive design was incorporated to regularly monitor the data for futility. Furthermore, Bayesian adaptive randomisation was implemented whereby the performance of each treatment arm was evaluated and this updated information was used to guide the randomisation allocation. The trial allowed the assessment of four treatments in five biomarker groups within one trial rather than performing 20 separate phase II single arm studies. Having all these patients within one trial can make it easier to compare patients across the treatment arms (Liu et al. 2015).

2.9 Seamless designs

The phases (phases I - III) of a clinical trial are usually conducted separately and in sequence to ensure robust conclusions about a new drug. However, it can be inefficient and costly to create separate trials for each phase, hence seamless

designs transition the new drug sequentially from one phase to another (George et al. 2016). The overall aim of these designs is to reduce the time taken for a new drug to pass through the different phases of drug development as well as reduce patient exposure, which has been demonstrated by reductions in sample size by at least 30% (Inoue et al. 2002, Maca et al. 2006).

Seamless designs are implemented within the field of Oncology by firstly assessing outcomes that can be measured sooner also known as surrogate outcomes such as DFS compared to longer-term outcomes such as OS (see section 1.3). These surrogate outcomes are measured to see the effect of a new treatment on the size of a tumour within a phase II setting and then a longer term outcome within a phase III setting (Stallard 2010). However, the surrogate outcome should correlate with the longer-term outcome. The success of a seamless phase II/III trial is dependent on the choice of outcome.

Pick the winner designs can be used in seamless designs (Scher et al. 2002). A pick the winner design can have multiple arms with two stages; a selection stage (phase II) at which 'the winner' is chosen to be carried into the confirmation stage (phase III), where new patients are recruited and randomised to receive either the control or 'the winner'. This design has been compared to a modified MAMS design as it can evaluate multiple treatments across at least two stages however the difference is that the MAMS design allows the continuation of a treatment based on critical values whereas the pick the winner approach evaluates improvement in outcomes regardless of statistical testing (Hills et al. 2011).

An example of a phase II/III seamless study evaluated the safety and efficacy of pegteograstim on chemotherapy-induced neutropenia against pegfilgrastim in patients diagnosed with breast cancer (Lee et al. 2016). At phase II, patients were administered four to six cycles of chemotherapy and assigned to two different groups. It was at this phase where the optimal dose of pegteograstim was chosen to be seamlessly carried to a phase III trial.

2.10 Master Protocols, Umbrella trials, Basket Trials, Platform-based trials

The concept of precision medicine entails finding the biomarker of a patient's cancer which then leads to personalised therapy (West 2017). Redman and Allegra et al (2015) have defined master protocols to be a top-level protocol within which multiple biomarker-based sub-trials can be contained each with their own protocols (Redman et al. 2015, Hirakawa et al. 2018). The different types of trial designs within a Master Protocol are Umbrella trials, Basket trials and Platform-based trials. These trial designs all look at more than one treatment arm or population at a time, but they do not have their own exclusive adaptive methodology to undertake these trials. They implement methods that have been described in earlier sections to make decisions, for example the STAMPEDE trial is a platform trial but implements the MAMS methodology.

A recent systematic review of Master protocol trials (Park et al. 2019) identified 83 master protocols (49 basket trials, 18 umbrella trials and 16 platform trials) for which the majority of these were in oncology (76/83, 92%). These designs offer trialist improved efficiency by accelerating late stage development by assessing multiple drugs at once in comparison to two-arm trials resulting in the rapid rise of implementation of these designs (US Food and Drug Administration 2018, Park et al. 2019). Furthermore, these designs can be extremely useful in times of global health pandemics such as the coronavirus disease, to rapidly identify efficacious treatments and stop unnecessary research into non-effective treatments (Noor et al. 2020).

An example of a master protocol is the Adjuvant Lung Cancer Enrichment Marker Identification and Sequencing (ALCHEMIST) trial as it integrates three trials (ALCHEMIST-Screening, ALCHEMIST-EGFR and ALCHEMIST-ALK) for patients with early stage lung cancer and allocates drugs that target specific genomic changes occurring within patients (Gerber et al. 2015). The ALCHEMIST-Screening trial aims to screen between 6000 and 8000 stage 1-3 NSCLC patients and assess the tumours for Epidermal Growth Factor Receptor (EGFR) mutations and Anaplastic Lymphoma Kinase (ALK) rearrangements. Patients that have the aforementioned tumour types will be referred to either the ALCHEMIST-EGFR (sample size = 410

patients) or ALCHEMIST-ALK (sample size = 360 patients) trials respectively. The remaining patients will be followed up for relapse and survival within the screening cohort.

2.10.1 Umbrella trials

Umbrella trials involve patients of the same cancer type being recruited and a test performed to detect which biomarkers are present in order to allocate patients to the relevant drug (West 2017) (Figure 2.5). Umbrella trials are usually phase II or phase II/III trials. An advantage of umbrella designs is that they allow multiple drugs to be tested simultaneously as different arms may open and close over time. Furthermore, patients are treated for the same tumour type and stage of disease, therefore any benefits observed in patients on the experimental arm can be linked to the biomarker hence assisting with prognostic homogeneity (Renfro et al. 2017, West 2017). Implementing such a design could result in a potentially large trial with a long duration which comes about due to funnelling patients with one type of cancer into smaller sub-studies based on their biomarkers (Renfro et al. 2017). Trial stakeholders may not find the long duration of these designs appealing if in the situation where answers are swiftly required.

An example of where the Umbrella design has been implemented is within the LUNG-MAP trial (Ferrarotto et al. 2015, LUNG-MAP 2017). The LUNG-MAP trial assesses patients diagnosed with lung cancer and endeavours to compare targeted treatment based on a patient's genomic makeup versus current standard therapy. Hence there are many sub-studies within the LUNG-MAP trial based upon the treatment that is matches the patients' tumour profile.

Similarly, the FOCUS4 trial (FOCUS4 2019) implements an umbrella MAMS design with biomarker-stratified and non-stratified comparisons all with separate protocols within one master protocol. More details of this trial can be found in chapter four.

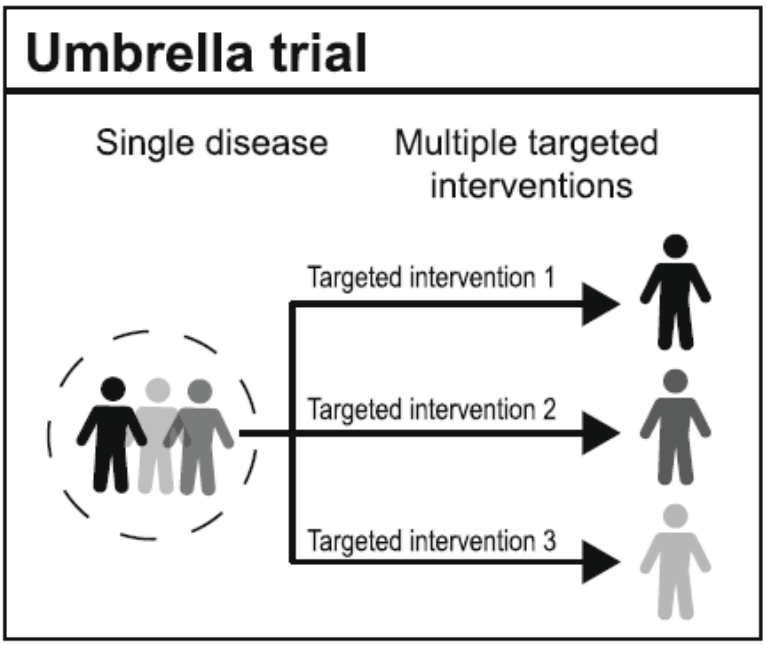

Figure 2.5: A graphical representation of an Umbrella trial design (Park et al. 2019).

2.10.2 Basket trials

A Basket trial enrols patients that have the same or similar biomarkers but could be diagnosed with different cancers and allocates them the same treatment (West 2017) (Figure 2.6). Normally Basket trials are phase II, single-arm small (30-40 patients) studies where within each cohort of patients there is a predefined target-response hypothesis. However, they can also be implemented within phase III settings like the ADD-ASPIRIN trial (Coyle et al. 2016). The ADD-ASPIRIN trial investigates the use of aspirin in avoiding recurrence and improving survival in patients diagnosed with early stage cancer in four solid tumours (breast, colorectal, gastro-oesophageal and prostate). Every patient has a run-in period whereby 100mg aspirin is taken daily for eight weeks, after which patients are randomised separately within each tumour-specific cohort to one of three arms: placebo, 100mg aspirin daily or 300mg daily. For the patients that are aged over 75 years, they are only randomised to either placebo or 100mg aspirin daily.

The advantages these trials offer are that a variety of treatments can be beneficial to a variety of disease types (Moore et al. 2016). However the disadvantages with these trials are the prognostic heterogeneity, whereby patients within the same

cohort may have completely different characteristics due to the different cancer types (Moore et al. 2016, Renfro et al. 2017).

An example where a Basket trial design has been utilised is within the NCI Molecular Analysis for Therapeutic Choice (NCI-MATCH) trial. This trial assigns patients to receive treatments according to the genetic changes occurring within their tumours (National Cancer Institute 2017). Currently the trial has 18 different treatment arms available to patients, with most of the treatment arms enrolling up to 35 patients. The trial aimed to recruit 25% of patients that have a rare cancer type but currently have 60% of patients with rare cancer types thus exceeding their expectations.

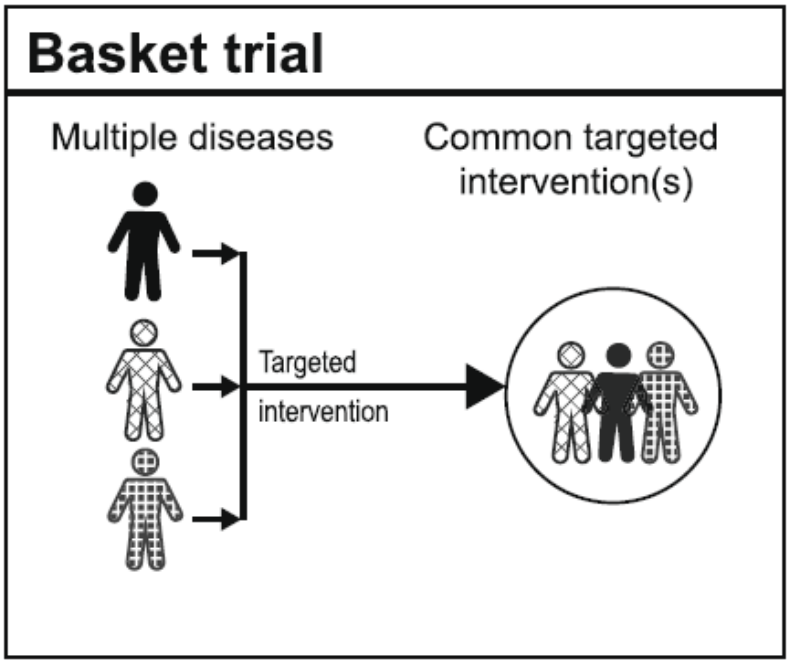

Figure 2.6: A graphical representation of a basket trial (Park et al. 2019).

2.10.3 Platform trials

A platform trial can be defined as a trial in which various sub-trials continue to enter and exit; sub-trials in a phase III setting may exit the platform due to futility or lack of benefit (Renfro et al. 2017) (Figure 2.7). Bayesian methods can be used within platform trials whereby Bayesian decision rules are utilised to determine whether treatments should continue or stop (Saville et al. 2016). The main advantage these types of designs offer are the flexibility of dropping ineffective

experimental arms at earlier stages and introducing new arms. Furthermore, trials can seamlessly progress from a phase II to a phase III trial. The potential size, duration and operational requirements of a platform trial can be a discouraging factor for stake-holders to implement this trial design. These designs can be useful in times of global health pandemics such as the coronavirus disease 2019 (COVID-19) as there would be an urgent need to rapidly identify efficacious treatments and reject treatments that are futile, lack benefit or too toxic (Noor et al. 2020)

An example of a platform trial is the STAMPEDE trial; which was initially classed as a MAMS design but evolved into a platform trial with the addition of treatments arms[1]. It is a phase II/III trial looking at patients with locally advanced prostate cancer who are commencing long-term Androgen Deprivation Therapy (ADT). The addition of experimental arms are treated like two-arm multi-stage trials (Sydes et al. 2011). More details of this trial can be found in chapter four.

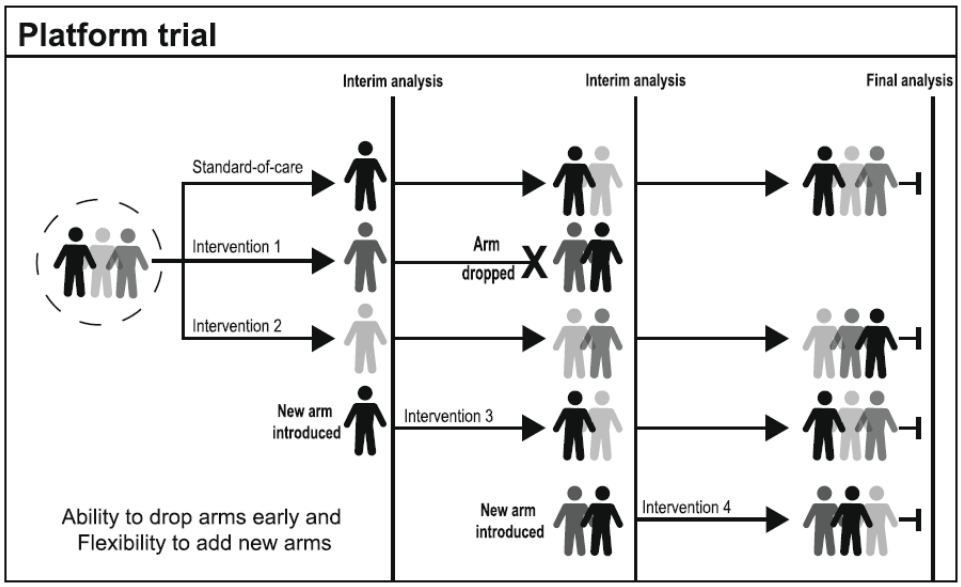

Figure 2.7: A graphical representation of a Platform trial design (Park et al. 2019).

2.11 Operational considerations

The appropriate choice of adaptive design method to adopt is critical but it is also of utmost importance to ensure trial teams are fully aware of the operational and

[1] http://www.stampedetrial.org

resource intensive demands required to execute adaptive trials. Underestimating this could result in the early closure of a trial due to lack of resource.

At the planning stages of a trial, it is important to outline all trial processes, evaluate the hurdles that may occur and ensure strategic planning to either avoid these pitfalls or increase the efficiency of these processes. For example, remote online data entry at sites instead of using paper case report forms (CRFs) may be a more efficient way to collect the data and avoid data input errors by the trial team. If an adaptive trial has interim analyses, then the time between each interim stage must be sufficient to cater for the operational demands required prior to any analysis, like sufficient time required for data entry, validation and querying to ensure that all data is clean and up to date, the logistical preparation required such as organising meetings and being ready to action the outcome of these meetings, i.e. if the decision at an interim analysis meeting is to discontinue an experimental arm then ensure the trial team is prepared.

Trials that employ adaptive designs should always have an independent Data Monitoring Committee (DMC) to ensure the credibility/integrity of the decisions made during the trial. Trial teams can face logistical difficulties in arranging DMC meetings which can lead to delays in making decisions. Therefore, to ensure quoracy is met, trial teams may want to employ a larger than normal DMC.

Non-trial specific processes must be taken into consideration, for example, it is inevitable that there will be some sort of staff turnover during the life of the trial and hence trial teams must ensure that processes are in place to ensure efficient training of new staff members.

The operational papers published by the STAMPEDE and FOCUS4 trial teams discuss at the data management challenges, operational aspects and personal experiences of the researchers involved on these teams and should be referred to prior implementing adaptive trials (Hague et al. 2019, Morrell et al. 2019, Schiavone et al. 2019).

A few of the many aspects to consider when implementing any sort of adaptive trial have been outlined above. Trial teams can also refer to the Practical Adaptive

and Novel Designs and Analysis (PANDA) project[1] which is currently under development, however will be a useful resource and aims to educate researchers with regards to the practical applications when applying adaptive designs in RCTs (ScHARR CTRU 2020). More considerations when designing an adaptive trial are discussed in chapter seven.

2.12 Software for Adaptive designs

The design and analysis of adaptive designs can be complex in comparison to standard trials and hence the use of customised software is required (Gallo et al. 2006). The development of computational software for adaptive design methodology is continually increasing and specialised software such as the FACTs (Fixed and Adaptive Clinical Trials Simulator), EAST and COMPASS software have been created to assist in designing adaptive trials (Corporation 2000, Berry et al. 2010, Berry Consultants 2012, Bolognese 2017).

The FACTs software created by Berry consultants[2] allows users to design, simulate and compare adaptive design based clinical trials. This software can simulate dose escalation trials, dose finding trials and enrichment trials. Furthermore, this software allows various key components to be modified during the course of the trial such as treatment arms, allocation ratio and sample size and is extremely useful for designing phase I studies. The EAST software can also assist in the creation and simulation of adaptive based trial design. However, this software has different options compared to the FACTs software to assist in implementing adaptive designs such as ensuring strong error control of the FWER with multi-arm trials, providing an array of accurate early stopping boundaries to consider different options such as sample size, cost effectiveness and differences in time for implementing a MAMS design with options for sample size re-estimation, dose selection and early stopping. The COMPASS software focuses on designing adaptive dose-finding trials, giving both Bayesian and Frequentist methods and furthermore allowing users to compare a variety of design options.

[1] https://www.sheffield.ac.uk/scharr/research/centres/ctru/panda
[2] https://www.berryconsultants.com/software/

Programs have also been created to be used within commonly used software such as SAS, R or Stata (Chang 2014, Grayling 2017, Grayling 2019). For group sequential designs the 'SEQDESIGN' package within SAS can be used and is aimed at creating interim analyses for clinical trials, the 'gsDesign' package within R focuses mainly on designs to implement alpha and beta spending functions however only the Wang-Tsiatis method, O'Brien-Fleming and Pocock methods are available (Pocock 1977, O'Brien et al. 1979, Wang et al. 1987).

Michael Grayling and his colleagues have created a group of packages to design group sequential clinical trials with normally distributed outcome variables within Stata (Grayling 2017) such as the 'HaybittlePeto' package that implements a group sequential design with Haybittle-Peto boundaries (Haybittle 1971, Peto et al. 1977) and the 'WangTsiatis' package that calculates the sample size and the boundaries required to implement a group sequential design with Wang-Tsiatis boundaries (Wang et al. 1987). The 'LanDeMets' package in Stata has been created by Ignacio Ulibarri and calculates the boundaries required for a group sequential design using the method of Lan and DeMets (Gordon Lan et al. 1983, Ullibarri 2013). This command further allows the user to select either the O'Brien-Fleming method or Pocock method to use as the alpha spending function.

As mentioned previously, an extension to the group sequential design is the MAMS design, the 'mams' package within R has been created by Thomas Jaki and his colleagues to assist in designing these trials (Jaki et al. 2019). Prior to the release of this command, Patrick Royston and his colleagues developed the following packages within Stata to assist with the design and implementation of the MAMS designs:

- The 'artsurv' command calculates both sample size and power for trials that have time-to-event outcomes (Royston et al. 2010).
- The 'artpep' command addresses the practical issue of staggered recruitment and accumulation of data in time-to-event outcome based trials (Royston et al. 2010). Sample size calculations are based on critical

assumptions and uses the number of events as an important indicator of trial progression. The 'artpep' command is used to project the power and events in trials.

- The 'nstage' command offers calculations of sample size, number of events and trial duration for trials with two or more treatment arms with a time-to-event outcome (Barthel et al. 2009, Blenkinsop et al. 2019). The command influences the decision to reject treatment arms that show lack of or no benefit against the control arm by comparing the HR obtained against a critical HR. Therefore, encouraging those treatment arms that indicate a difference to proceed based on the HR obtained. An intermediate outcome measure can be used instead of the primary outcome measure at the interim stages of testing. A pairwise comparison occurs between each experimental arm and control arm to determine which treatments continue to the next stage. Conclusively, treatments that have continued to the final stage are compared against the control arm on the primary outcome measure and superiority is concluded. This command can be used in conjunction with the 'artsurv' and 'artpep' to measure and calculate the timings of interim analyses based on real-time recruitment.

Three MAMS based cancer trials (FOCUS4, STAMPEDE and CompARE trials) are described in chapter five, all three of these trials have been designed using the Stata commands, 'artpep' and 'nstage'.

2.13 Conclusion

There are numerous ways in which adaptive designs can be conducted within Oncology, as discussed within this chapter. Adaptive designs are implemented more in oncology than other disease areas due to the often long duration of these

trials. Therefore, it would be advisable to consider more than one treatment at a time and adapt a trial sooner rather than later for efficiency where possible in these disease settings. Adaptive trial designs have the capacity to modify sample sizes in comparison to standard trial design to ensure the most efficient number of patients are treated with the best possible treatment. However, the execution of these trials from an operational perspective must not be undermined and diligence is required to ensure a 'successful' trial.

The next chapter will investigate whether these innovative adaptive trial design methodologies are being translated in practice and being implemented in cancer RCTs.

3 Literature review of Cancer RCTs

3.1 Introduction

The previous chapter discussed some of the common adaptive design methods and provided examples of how these methods have been applied. This led onto investigating the current literature to understand the use of adaptive design methods in RCTs and how these trials are reported.

Previous literature reviews have been conducted to investigate specific adaptive design methods or to identify all adaptive design RCTs using clinical trial registries and the National Institute for Health Research (NIHR) register (Bauer et al. 2006, Stevely et al. 2015, Hatfield et al. 2016).

One such review by Hatfield et al. (2016) investigated the condition and characteristics of registered adaptive design trials. This review used the ClinicalTrials.gov website and the NIHR register to collate all phase II, II/III and III adaptive design trials registered between 29 February 2000 and 1 June 2014. Hatfield et al. (2016) found that the use of adaptive designs have increased over time and out of the 143 trials found, the highest were within the field of Oncology (35/143, 24%). The limitations of the interface of the ClinicalTrials.gov website and NIHR register restricted the capture of all trials with adaptive designs. This led to the authors proposing that clinical trial registers should promote the use of the term 'Adaptive design' in either the title, summary or design sections of the register or allocate a section to adaptive design.

Stevely et al. (2015) investigated the reporting of phase III group sequential RCTs against the Consolidated Standards of Reporting Trials (CONSORT) 2010 checklist. Full text papers in the English language published from the 1 January 2001 to 23 September 2014 were reviewed. Papers were eligible for inclusion if they were parallel group RCTs with confirmatory objectives that applied group sequential methods using a Frequentist approach, papers that applied Bayesian methods were part of the exclusion criteria. Many eligible papers reviewed were from the field of Oncology (76%). The authors concluded that the reporting of group

sequential trials was not in line with the CONSORT 2010 guidelines and therefore these issues could be alleviated by creating an extension to the current CONSORT.

Bauer and Einfalt (2006) performed a review to assess applications of applied adaptive design methods based on the combination test or conditional error function approach (see section 2.5). A list of 60 papers that were published between 1989 and 2004 were selected. Many other papers were eligible for assessment however these 60 papers were selected by the authors as they thought that these papers would be the most cited when practically implementing adaptive design methods. The review concluded that the standard of reporting statistical methods needs improving and suggested a list of points that should be addressed when providing a description of the statistical methods used.

All the above-mentioned reviews have either investigated a specific adaptive design method or explored trial registries to find adaptive design trials prior to 2014. However, it is unclear what the state of applications of adaptive design methods in RCTs within the field of Oncology is and whether reporting has improved over time. The aim of this literature review was to capture full-text papers that incorporated adaptive design methods within phase II, II/III (also known as seamless trials – see section 2.9) or III RCTs in the field of Oncology published in 2015 as previous reviews have shown that the majority of adaptive trials have been applied in this field and the context of this PhD is focused on phase III cancer trials. The objectives of this review were:

- To understand the different applications of adaptive design methods;
- Whether these methods were being explicitly stated or if it had to be inferred that adaptive design methods were applied;
- Whether these methods were pre-determined or if circumstances influenced the need to implement any adaptive changes (ad-hoc).

3.1.1 Extension to the CONSORT guidelines

In 1996, the first CONSORT guidelines were published with the aim of improving the reporting of RCTs (Begg et al. 1996, CONSORT 2010). Since then, the CONSORT guidelines have evolved with updates being made in 2001 and the latest version in 2010 (Moher et al. 2001, CONSORT 2010). The CONSORT guidelines are based on a two-arm parallel design however various extensions have been added to the CONSORT to accommodate different design aspects, interventions and types of data such as cluster trials, pragmatic trials, acupuncture interventions (Campbell et al. 2004, Zwarenstein et al. 2008, MacPherson et al. 2010).

The contents within this chapter are based upon research that has been published during the course of this PhD (Mistry et al. 2017a) (appendix 2). At the time of this research, the CONSORT 2010 contained sections related to the use of interim analyses, however some authors had suggested for the CONSORT to be extended exclusively for trial designs implementing adaptive design methods (Stevely et al. 2015, Hatfield et al. 2016). Therefore, an extension to the CONSORT guidelines to incorporate adaptive designs was proposed (Mistry et al. 2017a). Furthermore the outcome of this review and extension to the CONSORT guidelines were presented at the International Society for Clinical Biostatistics conference, Birmingham 2017 and at the International Clinical Trials Methodology conference, Liverpool 2017 (Mistry et al. 2017b).

Since this publication, the Adaptive designs CONSORT Extension (ACE) group introduced an official extension to the guidelines and have published various papers (Dimairo et al. 2018, Dimairo et al. 2019, Dimairo et al. 2020). The ACE group was approached with the intention to collaborate with them; however, it was too late as most of the work by the ACE group was completed. Therefore, another part of this chapter compares the similarities and differences between the two CONSORT extensions (Table 3.6).

3.2 Methods

3.2.1 Literature search

The literature search was conducted using the Embase, Ovid and PubMed databases. The review was constrained to the following inclusion criteria:

- Application of adaptive design methods;
- Phase II, phase II/III, phase III RCTs;
- Within the field of Oncology;
- Disclosed results of the primary outcome;
- Published in 2015;
- Full obtainable text;
- Written in the English language.

Any duplicate records were removed based on the title, authors, abstract and year of publication. Any potential modifications made to the trial/statistical process that were either premeditated, implemented spontaneously or retrospective were used as the definition to recognise when adaptive design methods had been utilised (Chow et al. 2008, Bretz et al. 2009).

A free text search was conducted using the key search terms to maximise the capture of phase II, II/III or III RCTs in the field of Oncology, the list of these search terms can be found in appendix 3.

3.2.2 Data extraction

An Excel spreadsheet was used to record the following data:

- Standard demographics such as first author, title, name of the trial;
- The journal that the paper was published in;
- The sponsor or funder of the study;
- Trial phase;

- Cancer type being reported;
- Nature of the primary outcome;
- Number of trial arms;
- What sort of intervention was being implemented;
- How many planned/unplanned interim analyses;
- At what stage the trial is being reported, i.e. interim or final analysis;
- If there was any planned/unplanned stopping criteria and reason;
- If the trial was concluded early and the reason;
- Initial planned sample size and the reported sample size;
- Type of adaptive design method implemented;
- Number of adaptive design methods that were applied;
- Whether the adaptive design methods applied were pre-determined, ad-hoc or both;
- If the use of adaptive design methods were explicitly stated or if it had to be inferred;
- The trial identifier if registered on clinical trial registries.

The papers identified by the literature search were all reviewed, and the information was extracted and recorded in the Excel spreadsheet. Data that could not be found in the paper was researched by using the trial identifier or trial name to find out the relevant information. If no further information was available, then the data was classified as missing. The categorisation of the adaptive design methods were based upon those made by the FDA (US Food and Drug Administration 2018).

The eligibility of all the papers and reasons for full text exclusions were checked by myself and supervisors (Andrea Marshall and Janet A. Dunn). A 10% sample was

checked for accuracy of all information by the supervisors. Any discrepancies were discussed and a consensus agreed.

3.2.3 CONSORT extension

The proposed CONSORT extension was developed through a series of discussions and an iterative process with the supervisors (Janet Dunn and Andrea Marshall). Each item of the CONSORT statement was evaluated to determine whether clarification for adaptive designs was necessary based on prior knowledge and evidence from the literature review. Suggestions for how adaptations should be incorporated into the extension were discussed until a consensus was agreed. Furthermore, proposed extensions were submitted as a manuscript and extensively reviewed prior to publication by Joseph Koopmeiners, Abigail Shoben and Kristen May Cunanan all of whom have experience with the adaptive designs, for the suitability and applicability of the proposed CONSORT extension.

3.3 Results

Across the Ovid, Embase and PubMed databases, 8288 records were recognised as RCTs in the field of Oncology. There were 734 records which were related to adaptive design methods within phase II, phase II/III and phase III trial setting. The removal of duplicates reduced the number of papers to be screened to 464 records. Of the 464 records, 364 were deemed ineligible due to the following reasons:

- Only the abstract was published (n=263);
- Did not meet all inclusion criteria (n=68);
- Record was either a review or a methodology paper hence would not contain any results (n=33).

This left 100 full text records which were assessed in more detail for eligibility, of which a further 46 records were removed for the following reasons:

- Not a RCT (n=33);
- The trial was not cancer related (n=3);

- No analysis on the primary outcome measure (n=9);
- No relevant information provided (n=1).

A total of 54 records were included in the final review which has been presented below using the PRISMA flow diagram in Figure 3.1 (Liberati et al. 2009). A detailed list of all 54 papers used can be found in appendix 4.

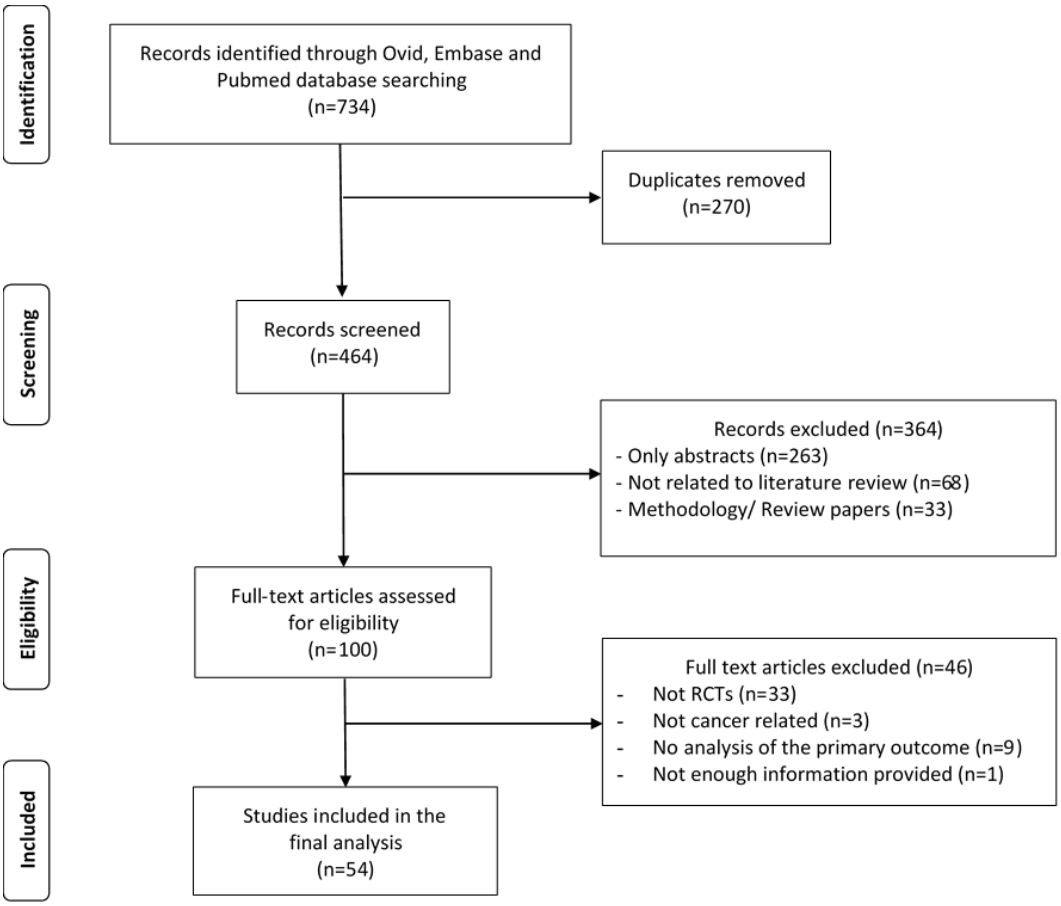

Figure 3.1: PRISMA flow diagram showing the process to identify the papers for final review.

The 54 papers reviewed were all RCTs and published in 21 different journals, of which the Journal of Clinical Oncology had the highest number of papers (12/54, 22%), the Lancet Oncology journal followed with 11 papers (21%), followed by the New England Journal of Medicine (6/54, 11%), European Journal of Cancer (3/54, 6%) and the remaining 17 journals published either one or two papers (Figure 3.2)

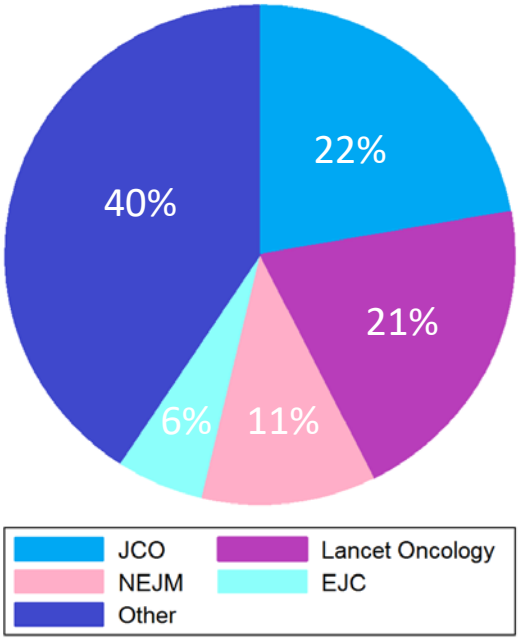

Figure 3.2: Pie chart of the journals in which the papers obtained from the literature review have been published.

The number of arms within these trials ranged from a minimum of two arms to a maximum of five arms; 46 (85%) papers were two arm trials, 6 (11%) were three arm trials, 1 (2%) was a four arm trial and 1 (2%) was a five arm trial (Table 3.1). Of the 54 papers reviewed, 38 papers (70%) were phase III trials, of which 35 of these were two arm trials, 2 of these were three arm trials and one four arm trial. Of the phase II trials, 10 trials had implemented two arms, two trials implemented three arms. One two arm trial and one three arm trial were implemented in phase IIb. Additionally, there was one seamless phase II/III trial with five arms in which patients were assigned either placebo or one of four propranolol regimens. A pre-planned interim analysis identified one of the four regimens and was carried forward into a phase III trial (Table 3.1).

Trial phase reported	**Number of arms**				**Total**
	2	**3**	**4**	**5**	
Phase II	10	2	0	0	12
Phase IIb	1	1	0	0	2
Phase II/III	0	1	0	1	2
Phase III	35	2	1	0	38
Total	46	6	1	1	54

Table 3.1: Two-way table comparing trial phase and trial arms

The majority of papers (50/54, 93%) incorporated drugs as the intervention for the study, whilst the remaining four interventions were a surgery/chemotherapy/ radiotherapy intervention (3/54, 6%) or a vaccine intervention (1/54, 2%) (Table 3.2). Additionally, the primary outcome measure for 45 out of 54 (83%) papers were time to event outcomes. The majority of primary outcomes were either PFS (22/54, 41%) or OS (13/54, 24%) (Table 3.3).

Intervention	Frequency
Vaccine	1
Surgery/Chemotherapy/Radiotherapy	3
Drug	50
Total	54

Table 3.2: List of categorised interventions of the papers.

Primary outcome	Frequency
Bowel movement/Flushing episodes	1
Disease Control Rate	1
Optimal dose	1
Overall pain response	1
Response to treatment	1
Cytogenic Response	2
Success/Failure	2
Time to event outcomes	45
DFS	1
EFS	1
OS and RFS	1
Time to neuropathy	1
RFS	2
OS and PFS	4
OS	13
PFS	22
Total	54

Table 3.3: List of primary outcomes of the papers. PFS = Progression-free survival; OS = Overall Survival; RFS = Relapse-free survival; EFS = Event-free survival; DFS = Disease-free survival.

There were 33 (61%) papers which published results obtained during an interim analysis, the remaining papers published results from either a subgroup analysis (1/54, 2%) or the final analysis (20/54, 37%). There were 26 (48%) papers that published results based on an interim analysis resulting in the trial reporting earlier than planned (Table 3.4), the reasons for reporting early were due to safety/efficacy/futility (21/26, 81%), met the primary endpoint (3/26, 12%) and

slow recruitment (2/26, 8%). All 26 papers had a pre-planned stopping criteria incorporated within the trial design for which an interim analysis would be performed that would check for safety/efficacy/futility i.e. group sequential methods (Table 3.4). A pre-planned analysis was incorporated within the majority of the papers (48/54, 89%). During the course of the trial 34 out of 48 papers planned for one interim analysis, 9 planned for two interim analyses, 3 planned for three interim analyses and the 2 remaining trials planned to perform an interim analysis annually.

Trial reported early	Pre-planned stopping criteria			Total
	No	Yes	Unknown	
No	5	22	1	28
Yes	0	26	0	26
Total	5	48	1	54

Table 3.4: Two-way table comparing early stoppage against pre-planned stopping criteria.

A single adaptive design method was applied within 44 out of 54 papers (82%), two methods applied within 9 out of 54 papers (17%) and three methods applied within 1 paper (2%) (appendix 4). There were no papers published from the same protocol. A total of 65 applications of adaptive design methods were implemented with adaptations using group sequential methods being the highest with 50 out of 65 applications (77%), followed by adaptations to treatment arm selection (8/65, (12%)), adaptations to sample size (4/65, 6%), adaptations to patient allocation (1/65, 2%), adaptations to endpoint selection (1/65, 2%) and adaptations to patient population (1/65, 2%).

Adaptive design methods were pre-determined within 49 out of 54 papers (91%), four papers had incorporated pre-determined and ad-hoc methods and one paper had applied ad-hoc adaptive design methods. The explicit use of the phrase 'adaptive design' to classify the use of adaptive design methods was stated in only two out of the 54 papers (4%); one paper incorporated multiple adaptive design methods including group sequential methods, sample size re-estimation and dose modification, the other paper stated the use of a 'Bayesian adaptive response design' which included dose modification and adaptive randomisation.

The different variables extracted from the papers have been categorised by adaptive design method applied (Table 3.5). Most studies using group sequential methods had two trial arms (42/50) and were in a phase III trial setting (35/50). All group sequential methods incorporated a planned stopping criteria as expected (see section 2.4) and 26 papers that reported an interim analysis resulted in early stopping of the trial. There were six trials that performed an interim analysis resulting in the trial continuing of which two were in the recruitment phase, two in the follow-up phase, one trial evaluated safety and one phase II trial concluded efficacy of a drug to be carried forward into a phase III trial.

Papers that implemented dose modification methods all had pre-determined the application of adaptive design methods, although only one explicitly stated the use of this method. All four papers that implemented sample size re-estimation methods incorporated a planned stopping criteria and led to the trial stopping early. Additionally, three out of the four papers that applied sample size re-estimation had both pre-determined and ad-hoc adaptive design methods.

3.4 Discussion

Various reviews have given different taxonomies to adaptive design methods, therefore the question arose as to how to classify these methods. It was decided to utilise the categories specified by the FDA as these categories are standardised (US Food and Drug Administration 2007, Chow et al. 2008, Bhatt et al. 2016, Hatfield et al. 2016, US Food and Drug Administration 2018). As mentioned in chapter two, group sequential methods have been used widely for a number of years so the robustness and the familiarity of these methods qualified them to be classified as 'well understood methods' by the FDA in the previous adaptive design guidance produced in 2010 (Todd 2007, US Food and Drug Administration 2010). Any papers from the literature search that incorporated interim analyses whereby treatment/trial related decisions were made were classed as group sequential methods.

A sensitivity analysis was performed to check if there were any papers published in 2015 by the known exemplar adaptive trials (section 4.3) that satisfied our

inclusion criteria, which were missed by the search. Two possible papers were identified relating to the STAMPEDE trial (James et al. 2015, James et al. 2016), but the first reported the results for the control arm as a cohort study and not the primary comparative results of the randomised trial and therefore was not eligible. The second paper (James et al. 2016) had an issue date of 2016 and therefore just fell outside of the search strategy, as the date of issue was used in the search strategy and not the online published date, which may be a limitation of the search strategy.

The review found that 53 out of 54 papers had pre-determined adaptive design methods which supported the definition specified by the FDA and many of these were phase III suggesting that adaptive design methods in RCTs are more frequently used in confirmatory trials. The extensive list of search terms (appendix 3) assisted in the capture of papers relating to various adaptive design methods, however only two papers explicitly defined the trial design to be an 'adaptive design' whilst it was inferred from the remaining papers.

One of the objectives of this literature review was to assess the reporting of the application of adaptive design methods within RCTs in the field of Oncology. This review reaffirmed the conclusions of the aforementioned reviews that there is a need to improve the reporting of adaptive design methods.

Furthermore, this review reiterated the need for an extension to the current CONSORT statement to ensure successful capture of adaptive design methods in RCTs as concluded by Stevely et al. (2015) and Hatfield et al. (2016). Thus, I proposed an extension to the CONSORT 2010 guidelines which has been published in the BioMed Central Medical Research Methodology journal (Mistry et al. 2017a). Thereafter, a consensus driven CONSORT extension for RCTs using adaptive design has been published (Dimairo et al. 2018) and a workshop was held by the ACE group in 2019 whereby the finalised guidelines were disseminated (appendix 5). Upon comparing the two extensions (Table 3.6), there were many similarities found, such as in section 2, where both CONSORT extensions emphasised the need for clearly defining the type of adaptive design used and reporting important changes in the design methods during the trial. In section 7 of

the proposed CONSORT extension guidelines, it was suggested that it is vital that any interim analyses that are performed are clearly shown and the decision making processes at this point are thoroughly explained.

Conversely there are items proposed in my extension which have not been mentioned in the CONSORT Statement by the ACE group such as stating in the title that it is an adaptive randomised trial to make it transparent that adaptive design method was adopted although the title word limit may prohibit this. However, the ACE group stressed the use of the term 'adaptive' as a keyword, which will be useful as an identifier in future literature searches. Additionally, I proposed that the rationale for implementing an adaptive design must be explained which has not been added in the CONSORT Statement extension proposed by the ACE group. Providing a rationale as to why a particular adaptive design has been implemented can be useful in giving justification and promoting adaptive designs.

The ACE group added sections to the original CONSORT 2010 checklist so it would cater more towards adaptive designs. In section 11, the ACE group emphasised that measures to ensure minimal interim information are reported to avoid any operational bias. An item in section 12 of the CONSORT 2010 has been added by the ACE group titled 'Statistical Methods' such that for any adaptive design feature, the statistical methods applied to make any key estimates or inferences were clearly shown. This was similarly proposed in my CONSORT extension whereby using simulation studies to show estimates or inferences has been suggested. A section reporting the results of any interim analyses has been added to section 17, 'Outcomes and estimation'. This decision has been regarded as very important as the decisions made at an interim analysis can dictate the way a trial runs and hence to report these are regarded as fundamental for any trial implementing interim analyses. The ACE group had experts in adaptive designs that provided input into the CONSORT extension which was a limitation to the CONSORT Statement I proposed. The extension to the CONSORT Statement published by myself and the ACE group (Mistry et al. 2017a, Dimairo et al. 2018) share a similar goal in striving to raise awareness of better and more transparent reporting of adaptive designs.

Data extracted	Adaptive method applied*					
	Group sequential methods (n=50)	Dose modifications (n=8)	Sample size re-estimation (n=4)	Adaptive randomisation (n=1)	Change in primary endpoint (n=1)	Change in patient eligibility (n=1)
Trial phase						
II	13	2	0	1	0	0
II/III	2	1	1	0	0	0
III	35	5	3	0	1	1
Number of arms						
2	42	7	2	1	1	1
3	6	1	0	0	0	0
4	1	0	1	0	0	0
5	1	0	1	0	0	0
Stage of reporting						
Interim analysis	33	2	2	0	1	1
Subgroup analysis	0	1	0	0	0	0
Final analysis	17	5	2	1	0	0
Planning of adaptive design method						
Pre-determined	45	8	1	1	0	1
Ad-hoc	1	0	0	0	0	0
Both	4	0	3	0	1	0
Explicitly labelled as an adaptive design						
Yes	1	1	1	1	0	0
No	49	7	3	0	1	1
Planned stopping criteria						
Yes	50	5	4	0	1	1
No	0	3	0	1	0	0
Trial stopped early						
Yes	26	1	4	0	1	0
No	18	6	0	1	0	1
Not yet	6	1	0	0	0	0

*Numbers based on the number of methods applied (65 adaptive methods) and not based on the number of papers.

Table 3.5: Data extracted split by adaptive method applied.

Section/Topic	Item No	Standard Checklist item	Proposed Extension for Adaptive Designs	ACE Group Extension for Adaptive Designs
Title and abstract				
	1a	Identification as a randomised trial in the title	Identification as an adaptive randomised trial if it is an adaptive design	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts) (Hopewell, Clarke et al. 2008, Hopewell, Clarke et al. 2008)	Include the term 'adaptive design' or 'adaptive methods'	**Trial design:** Description of the trial design (for example, parallel, cluster, non-inferiority); include the word 'adaptive' in the content or at least as a keyword **Outcome:** Clearly defined primary outcome for this report **Adaptation decisions made:** Specify what trial adaptation decisions were made in light of the pre-planned decision-making criteria and observed accrued data.
Introduction				
Background and objectives	2a	Scientific background and explanation of rationale	Rational for implementing an adaptive design	
	2b	Specific objectives or hypotheses		
Methods				
Trial design*	3a	Description of trial design (such as parallel, factorial) including allocation ratio	Define what adaptive design/ adaptive method will be applied	
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	Any changes during the trial should be reported as an adaptive method.	Important changes to the design or methods after trial commencement (such as eligibility criteria) outside the scope of the pre-planned adaptive design features, with reasons

Section/Topic	Item No	Standard Checklist item	Proposed Extension for Adaptive Designs	ACE Group Extension for Adaptive Designs
	3b	Pre-planned adaptive design features		Type of adaptive design used, with details of the pre-planned trial adaptations and the statistical information informing the adaptations
Participants	4a	Eligibility criteria for participants	Any changes in eligibility during the trial, should be classed as an adaptive design or adaptive method.	
	4b	Settings and locations where the data were collected		
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were administered		
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed		Completely define pre-specified primary and secondary outcome measures, including how and when they were assessed. Any other outcome measures used to inform pre-planned adaptations should be described with the rationale
	6b	Any changes to trial outcomes after the trial commenced, with reasons	Any changes during the trial are classed as an adaptive method and should be mentioned.	Any unplanned changes to trial outcomes after the trial commenced, with reasons
Sample size	7a	How sample size was determined	Any changes to sample size or power during trial classed as an adaptive design or adaptive method and should be mentioned.	How sample size and operating characteristics were determined

Section/Topic	Item No	Standard Checklist item	Proposed Extension for Adaptive Designs	ACE Group Extension for Adaptive Designs
	7b	When applicable, explanation of any interim analyses and stopping guidelines	Explain why the interim analysis will be taking place, if potential pre-planned adaptations during interim analysis taking place then these should be mentioned in the methods as well (3b). Include details of any planned stopping boundaries for either the trial or dropping any of the intervention arms.	Pre-planned interim decision-making criteria to guide the trial adaptation process; whether decision-making criteria were binding or nonbinding; pre-planned and actual timing and frequency of interim data looks to inform trial adaptations
Randomisation:				
Sequence generation	8a	Method used to generate the random allocation sequence		
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	Details if adaptive randomisation has been implemented.	Type of randomisation; details of any restriction (such as blocking and block size); any changes to the allocation rule after trial adaptation decisions; any pre-planned allocation rule or algorithm to update randomisation with timing and frequency of updates
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned		

Section/Topic	Item No	Standard Checklist item	Proposed Extension for Adaptive Designs	ACE Group Extension for Adaptive Designs
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions		
Blinding*	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how		
	11b	If relevant, description of the similarity of interventions		
	11c	Confidentiality and minimisation of operational bias		Measures to safeguard the confidentiality of interim information and minimise potential operational bias during the trial
Statistical methods*	12a	Statistical methods used to compare groups for primary and secondary outcomes	Details of how the adaptive design or the adaptive methods were applied Details of how the statistical methods were evaluated before implementation i.e. through the use of simulations?	
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses		
	12b	Estimation and inference methods		For the implemented adaptive design features, statistical methods used to estimate treatment effects for key endpoints and to make inferences

Section/Topic	Item No	Standard Checklist item	Proposed Extension for Adaptive Designs	ACE Group Extension for Adaptive Designs
Results				
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	Ensure any adaptations are shown on this diagram, such as dropping of arms, treatment switching.	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome and any other outcomes used to inform pre-planned adaptations, if applicable
	13b	For each group, losses and exclusions after randomisation, together with reasons		
Recruitment*	14a	Dates defining the periods of recruitment and follow-up		
	14b	Why the trial ended or was stopped	Any changes to recruitment during trial classed as an adaptive method, should be mentioned.	The reporting of reasons for stopping a trial early including circumstances leading to that decision could help readers to interpret results with relevant caveats.
	14c	Adaptation decisions		Specify what trial adaptation decisions were made in light of the pre-planned decision-making criteria and observed accrued data
Baseline data*	15	A table showing baseline demographic and clinical characteristics for each group		In the presence of marked differences in the numbers of randomised participants and those included in the interim or final analyses, authors are encouraged to report baseline summaries by treatment group for these two populations

Section/Topic	Item No	Standard Checklist item	Proposed Extension for Adaptive Designs	ACE Group Extension for Adaptive Designs
	15b	Similarity between stages		Summary of data to enable the assessment of similarity in the trial population between interim stages
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups		In the number of participants by treatment group should be reported for each analysis at both the interim analyses and final analysis whenever a comparative assessment is performed (for example, for efficacy, effectiveness, or safety).
Outcomes and estimation*	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)		Reporting of results as detailed in the CONSORT 2010 should mirror hypotheses of interest including subpopulations and full target populations.
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended		
	17c	Interim results		Report interim results used to inform interim decision-making
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory		
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms) (Ioannidis, Evans et al. 2004)		

Section/Topic	Item No	Standard Checklist item	Proposed Extension for Adaptive Designs	ACE Group Extension for Adaptive Designs
Discussion				
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses		For AD randomised trials, further discussion should include any deviations from pre-planned adaptations, interim analyses, protocol amendments on trial adaptations and results, potential bias and imprecision of treatment effects, potential heterogeneity in patient characteristics and treatment effects between stages.
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	If ad-hoc adaptive methods were implemented, at what point was it decided to implement this and why.	Authors should discuss the population to whom the results are applicable including any threats to internal and external validity which are trial dependent based on the implemented adaptations.
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence		
Registration	23	Registration number and name of trial registry		
Protocol*	24	Where the full trial protocol can be accessed, if available	First and last protocol, with a list of amendments made.	
	24b	Statistical analysis plan and other relevant trial documents		Where the full statistical analysis plan and other relevant trial documents can be accessed
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders		

*Sections have additional checklist items by the ACE group which have been coloured in grey.

Table 3.6: Proposed extension to the current CONSORT diagram (CONSORT 2010) compared to the ACE extension released in 2019 (Dimairo et al. 2019). ACE – Adaptive Design CONSORT Extension.

Clark (2013) reviewed unpublished research protocols submitted to research ethics committees in the UK in 2009 and found that out of a total of 446 protocols submitted, six reported an adaptive design using sample size re-estimation methods but an additional 88 protocols implemented a group sequential design (Clark et al. 2013). Therefore, trials that fit the definition of adaptive designs as specified within this literature review are encouraged to implement the proposed extension however trial stakeholders and methodologist incorporating stopping guidelines may wish to stick to the term 'group sequential' due to its familiarity and contribution to statistical methodology for many years (Todd 2007). The adaptive design extensions to the CONSORT 2010 will improve reporting and will ensure that more adaptive design based trials are captured in future literature searches and can help trial stakeholders to decide at what point a trial can be classed as adaptive. Furthermore, the greater capture of these adaptive methods through the CONSORT Statement extensions will greatly assist all those involved within clinical trials to ensure thorough planning and reporting of all adaptive natured RCTs, allowing full transparency of all adaptive methods performed during the trial.

3.5 Conclusion

In conclusion, from the literature review very few trials reporting in 2015 explicitly reported the use of adaptive design methods, which could imply that there may be many more trials using adaptive design methods without clearly stating so. The literature review also confirmed that the reporting of adaptive design methods within RCTs is poor and requires improvement. Following the adaptive design extension to the CONSORT 2010 guidelines may help to improve transparency of reporting the use of adaptive design methods within trials in the future.

4 Review of Adaptive designs used in Oncology clinical trials

4.1 Introduction

The literature review reported in chapter three suggested that adaptive methods were regularly implemented in oncology trials but not necessarily defined as 'adaptive trials'. This led to the decision to conduct a scoping exercise of the Warwick CTU trial portfolio, whereby on-going or completed cancer trials were assessed to see what trial design was implemented and if these trials used any adaptive methods. This exercise was expanded to examine exemplar oncology RCTs known to Warwick CTU that have implemented adaptive design methods.

The number of phase III Oncology trials has continually increased over the years, with most trials adopting a superiority hypothesis when comparing new cancer therapies to the standard. However, there are some novel cancer therapies that have been developed which are just as effective as the standard of care but are more convenient and less toxic (non-inferior) (Saad 2018). Hence, the choice of the hypothesis within phase III trials is dependent on the goals of the study. The choice of hypothesis will also be explored within this chapter.

4.2 Warwick CTU Scoping Exercise

Within Warwick CTU there were 13 on-going or completed phase II or phase III cancer trials in the portfolio in 2019 when the scoping exercise was undertaken. These 13 trials were assessed for the use of adaptive methods and choice of hypothesis with the results detailed below (Table 4.1). The Determination of Epidermal growth factor receptor-inhibitor versus Standard Chemotherapy early And Late Toxicity Events in Human Papillomavirus-positive oropharyngeal squamous cell carcinoma (De-ESCALaTE) trial had no changes, and the Optimal Personalised Treatment of early breast cancer using Multi-parameter Analysis (Optima) trial (including Optima prelim– the feasibility phase) was developed with an adaptive design. The remaining 11 out of 13 trials showed some form of pre-planned adaptation. The following adaptive methods were identified through this exercise; changes in the sample size including changes to the power of the study, changes in the eligibility criteria, and performing interim analyses to assess for

safety, efficacy or futility of drugs. Details of each trial and adaptations incorporated will be discussed in the following sections.

4.2.1 ARTemis trial – Pre-planned interim analysis

The Avastin Randomised Trial with neo-adjuvant chemotherapy for patients with early breast cancer (ARTemis) trial was a randomised phase III trial in patients diagnosed with human epidermal growth factor receptor 2 (HER2)-negative early breast cancer (Earl et al. 2015). The trial aimed to determine whether the addition of an anti-angiogenic agent bevacizumab to neo-adjuvant chemotherapy is superior to the standard neo-adjuvant chemotherapy in terms of both short and long-term outcomes.

The power calculations assumed a 70:30 split in the sample size between people with Oestrogen receptor (ER) positive and ER negative tumours and a pathological complete response on the standard arm of 10% and 25% respectively. A total of 800 patients (400 patients in each arm) were recruited into the trial to allow an absolute difference of 10% to be detected between the treatment groups in the primary outcome of pathological complete response with 85% power at the 5% two-sided significance level. The assumptions used for the sample size calculations were checked by the DMC. A pre-planned interim safety analysis was performed on the first 200 patients completing treatment to check for any potential safety issues related to the bevacizumab drug. The results of this interim analysis were considered by the DMC, who reported no safety concerns. As an interim analysis due to safety had been incorporated within this trial, it could be classed as a group sequential adaptive design.

Trial	Status*†	Adaptive Method	Hypothesis	Summary
ARTemis	Long term follow-up	Interim analysis	Superiority	Planned interim analysis for safety. Interim results showed no safety concerns, trial continued as planned. Superiority hypothesis implemented; absolute difference of 10%.
AVAST-M	Long term follow-up	Interim analysis, change in eligibility	Superiority	Planned interim analysis for safety and efficacy. Interim results showed no concerns hence continued as planned. Final analysis when all patients had been on study for five years. Eligibility criteria broadened to help increase rate of recruitment. Superiority hypothesis implemented; absolute difference of 8%
COUGAR-02	Completed	Sample size change, change in eligibility	Superiority	Sample size decreased from 320 to 164 patients due to new external information obtained and change of assumptions in the control arm. Eligibility criteria broadened to help increase rate of recruitment. Superiority hypothesis implemented; increase in median survival of two months.
LIHNCS	Completed	Sample size change, change in eligibility	Superiority	Sample size increased from 300 to 419 patients to allow for increased power in the pre-planned subgroup analysis. Eligibility criteria broadened to help increase rate of recruitment. Primary outcome looked at rate of dysplasia hence superiority hypothesis implemented to detect absolute difference of 13%.
MAMMO-50	Follow-up phase	Closing of cohort	NI	Closure of cohort study due to recruitment of patients being detracted from main RCT. NI hypothesis implemented; absolute NI margin of 3%.
Neo-Escape	Completed	Interim analysis	Superiority	Phase II two arm trial implementing 'Pick the winner' approach. Planned interim analysis resulting in closing of an arm due to futility.
Optima	Recruitment phase	Addition of trial arm	NI	Adaptive design implemented - plan to add additional multi-parameter test arms. NI hypothesis implemented; absolute NI margin of 3%.

Trial	Status*†	Adaptive Method	Hypothesis	Summary
Persephone	Long term follow-up	Interim analysis	NI	Unplanned interim analysis performed due to results released of another similar trial. Interim analysis results showed no concerns; trial continued as planned. NI hypothesis implemented; absolute NI margin of 3%.
PET-NECK	Completed	Change in eligibility Extension of follow-up time	NI	A minimum of 140 deaths were required at the end of the planned two-year follow-up period; this was not met resulting in extending the follow-up period to 5 years. Eligibility criteria was forced to be broadened due to a change in practice at many centres; this could have resulted in a significant reduction in recruitment if the change was not made. NI hypothesis implemented; absolute NI margin of 10%.
Select-D	Completed	Sample size change	Superiority	The second randomisation was closed due to futility. Sample size reduction from 530 patients to 400 patients in first randomisation. Pilot study estimating VTE recurrence rate originally with a width of the 95% CI of 8% but increased to a width of 9%.
TEAMM	Completed	Sample size change	Superiority	Sample size increased from 800 patients to 1000 patients. Superiority hypothesis implemented; amended from 10% difference with 90% power to 7% difference with 80% power.
VICTOR	Completed	Trial closure, sample size change	Superiority	Early closure of trial due to withdrawal of study drug. Sample size was 8000 but forced to analyse the 2434 recruited. Superiority hypothesis implemented; amended from 20% difference with 85% power to 25% difference with 87% power.

*Completed meaning that primary outcome has been reported and no longer following-up patients. †Status of the trials at December 2020.

Table 4.1: Summary of the scoping exercise performed of the trials at Warwick CTU. NI = Non-inferiority; OS = Overall Survival; CI = Confidence Interval

4.2.2 AVAST-M – Pre-planned interim analysis, change in eligibility

Adjuvant Avastin trial in high risk melanoma (AVAST-M) was a randomised phase III trial with a superiority hypothesis that investigated whether bevacizumab after standard surgery would help patients diagnosed with melanoma who were at high risk of recurrence and to see if this population live longer. Patients were allocated to either bevacizumab or regular checks post-surgery. The trial required a minimum of 1320 patients (660 patients in each arm) to detect an absolute increase in five-year OS from 40% to 48% with 85% power at the 5% two-sided significance level. The trial recruited 1343 patients; 671 were allocated bevacizumab and 672 had regular check-ups (Corrie et al. 2014).

A pre-planned interim analysis of safety and efficacy took place once all patients had finished treatment and had been in the study for at least a year which resulted in the continuation of the trial as planned. During the conduct of the trial, the eligibility criteria were broadened to extend the potential pool of patients suitable for the trial and improve recruitment. The inclusion criteria were extended to include patients greater than 16 years rather than 18 years, those with stage IIB disease and patients that may have received prior adjuvant therapy at an earlier stage of their disease. A final analysis of the AVAST-M trial was performed when all patients had been on the study for five years.

4.2.3 COUGAR-02 – Sample size decrease, change in eligibility

COUGAR-02 was a phase III open-label RCT with a superiority hypothesis that investigated the use of a chemotherapy drug, docetaxel, versus active symptom control for patients with advanced oesophago-gastric adenocarcinoma who relapsed within six months of previous chemotherapy (Ford et al. 2014).

The initial sample size calculated that 320 patients were required to be recruited within two years with a follow-up period of approximately six months to detect a difference in median survival from four months to six months with 90% power and 5% two-sided significance level. A minimum of 133 events in the control arm would be obtained by this time. However, the obtainment of new external

evidence from a similar trial showed that the actual OS was lower (2.4 months median OS) than initially assumed (Thuss-Patience et al. 2011). Hence after two years of recruitment and with 57 patients randomised, the sample size was amended using this new external data from 320 patients to 164 patients to be recruited within 3.5 years with six months for follow-up. The new sample size would require a minimum of 81 events to detect an increase in median survival from 3.5 to 5.5 months with 80% power and 5% two-sided significance level.

Furthermore, the eligibility criteria were extended to include patients with oesophageal cancer and include patients that had more than one prior chemotherapy to increase the potential pool of eligible patients. Hence to adjust for the new type of cancers, stratification was adjusted from type of surgery (previous gastrectomy versus stomach in situ) to site of disease (oesophagus versus oesophago-gastric junction versus stomach).

4.2.4 LIHNCS – Sample size increase, change in eligibility

The Lugol's Iodine in Head and Neck Cancer Surgery (LIHNCS) phase III RCT with a superiority hypothesis investigated the use of Lugol's iodine staining to improve clear resection margin rates in head and neck cancer surgery for oral and oropharyngeal carcinoma. Patients in this trial were allocated to either surgical treatment using Lugol's iodine or the control group which was surgical treatment without the use of Lugol's iodine (McCaul et al. 2013).

The primary outcome for this trial was the rate of dysplasia at margins. Initially the planned total sample size was 300 patients to achieve 90% power with a 5% two-sided significance level to detect a decrease of the rate of dysplasia from 20% in the control to 7% in the Lugol's Iodine group. However, the samples size was increased to 419 patients to give more power to the pre-planned subgroup analysis of 300 patients with stage I/II: the good prognostic subgroup.

4.2.5 MAMMO-50 – Closing of a cohort

MAMMO-50 is an ongoing phase III RCT in women diagnosed with early breast cancer. It is investigating whether less frequent mammograms i.e. two yearly mammography for conservation surgery patients or three yearly for mastectomy

patients is not inferior to annual mammography (Mammo-50 2016). Furthermore, trial patients were over the age of 50 at initial diagnosis, previously treated for invasive or non-invasive breast cancer and were three years post diagnosis. Alongside this trial, eligible patients who did not want to be randomised had the option to be registered into a sister cohort.

The total planned sample size of 5000 patients randomised were required to allow the detection of NI, where NI is defined as no worse than 3% below the control arm, assuming 5-year disease specific survival rate of 89%, with 90% power and a 2.5% one-sided significance level. The cohort study was closed to recruitment after the feasibility phase as it detracted from recruitment into the main RCT. Recruitment of the target 5000 patients to the main RCT was completed in September 2018. As of December 2020, follow-up for these patients continues.

4.2.6 Neo-escape – Pick the winner approach, pre-planned interim analysis

The Neo-Escape trial was a two-arm randomised phase II trial to assess the feasibility of extended chemotherapy; comparing arm one: Neoadjuvant carboplatin, then surgery followed by adjuvant paclitaxel and gemcitabine versus arm two: Neoadjuvant gemcitabine and carboplatin, then surgery followed by adjuvant paclitaxel in patients with inoperable ovarian cancer (Poole et al. 2012).

The primary feasibility outcome was the percentage of patients completing 12 cycles of chemotherapy; >80% deemed an acceptable completion rate and <60% unacceptable. Fleming's single stage procedure was used to calculate the required sample size of 44 patients for each trial arm to test feasibility with a type I error of 5% and type II error of 10%. If both arms met the pre-specified feasibility criteria then the study would be extended to a comparative randomised trial using a 'pick the winner' approach (see section 2.9 for more information) for which an additional total of 36 patients would be recruited. Following an interim analysis, one of the trial arms was closed to recruitment on the grounds of futility, which suggested less than 5% probability of it reaching the pre-set criteria for feasibility.

The trial continued as a single arm phase II trial and closed to recruitment with 47 patients recruited to the remaining trial arm.

4.2.7 Optima trial – Add additional multi-parameter test arms

The Optima trial is a partially blinded randomised phase III trial with a NI design comparing test directed therapy using a multi-parameter test with standard treatment of chemotherapy and endocrine therapy in early breast cancer patients (Stein et al. 2016). This trial has planned for the addition of other multi-parameter tests hence it has been classed as an adaptive design. It included a feasibility phase, Optima prelim, which evaluated various multi-parameter tests with a view to adapt the 'gold-standard' test after this preliminary stage if needed. One of the key conclusions from the Optima prelim trial was that although there is significant value to the National Health Service (NHS) to research into all multi-parameter tests, the Prosigna test was considered as the highest priority. This was a key reason for changing from the 'gold standard' Oncotype DX test in Optima prelim to the Prosigna test in the Optima main trial to determine which treatment patients would receive from the test results.

The Oncotype DX test looks at tumour samples for a group of 21 genes (Paik et al. 2004). The result of this test gives a recurrence score which calculates the risk of distant recurrence following endocrine therapy of ER positive node negative breast cancer and determines whether patients are low, intermediate or high risk. Various studies have supported the use of this test and have shown that it can predict chemotherapy sensitivity within an adjuvant setting (Gianni et al. 2005, Chang et al. 2008, Marchionni et al. 2008, Smartt 2010, Hornberger et al. 2012, Ward et al. 2013). However the restrictions that have been highlighted by this test include the limited amount of data supporting the use of this test as the data supporting these claims were based on small patient cohorts (Smartt 2010, Hornberger et al. 2012, Ward et al. 2013). Furthermore the effect of the Oncotype DX test on long term outcomes (e.g. overall survival) have not been shown (Stein et al. 2016).

The Prosigna test looks at a group of 58 genes to approximate the risk of recurrence which can be scored from 0 to 100 (Breast Cancer 2016) and categorised as low, intermediate or high risk (Sestak et al. 2013, Filipits et al. 2014, Breast Cancer 2016). Various studies have proven that through the use of Prosigna testing, the response to adjuvant chemotherapy can be predicted and can be differentiated between high risk and low risk patients that have ER positive disease (Parker et al. 2009, Esserman et al. 2012, Prat et al. 2016).

The sample size required for the Optima prelim trial was 300 randomised patients to assess agreement between the various multi-parameter test. The Optima prelim trial registered 350 participants from 34 centres; of which 313 were randomised.

The Optima main trial aims to recruit a sample size of 4500 patients to demonstrate NI of test-directed treatment (defined as no worse than 3% below the estimated 85% 5-year invasive DFS rate for the control arm) with 80% power and a 5% one-sided significance level. Furthermore, the adaptive designs of the Optima trial allows for the addition of other multi-parameter genomic tests in the future (Figure 4.1).

The vital question that will be answered through this trial will be if the use of multi-parameter genomic testing can accurately differentiate between high risk and low risk breast cancer patients and hence reduce the number of patients that receive chemotherapy.

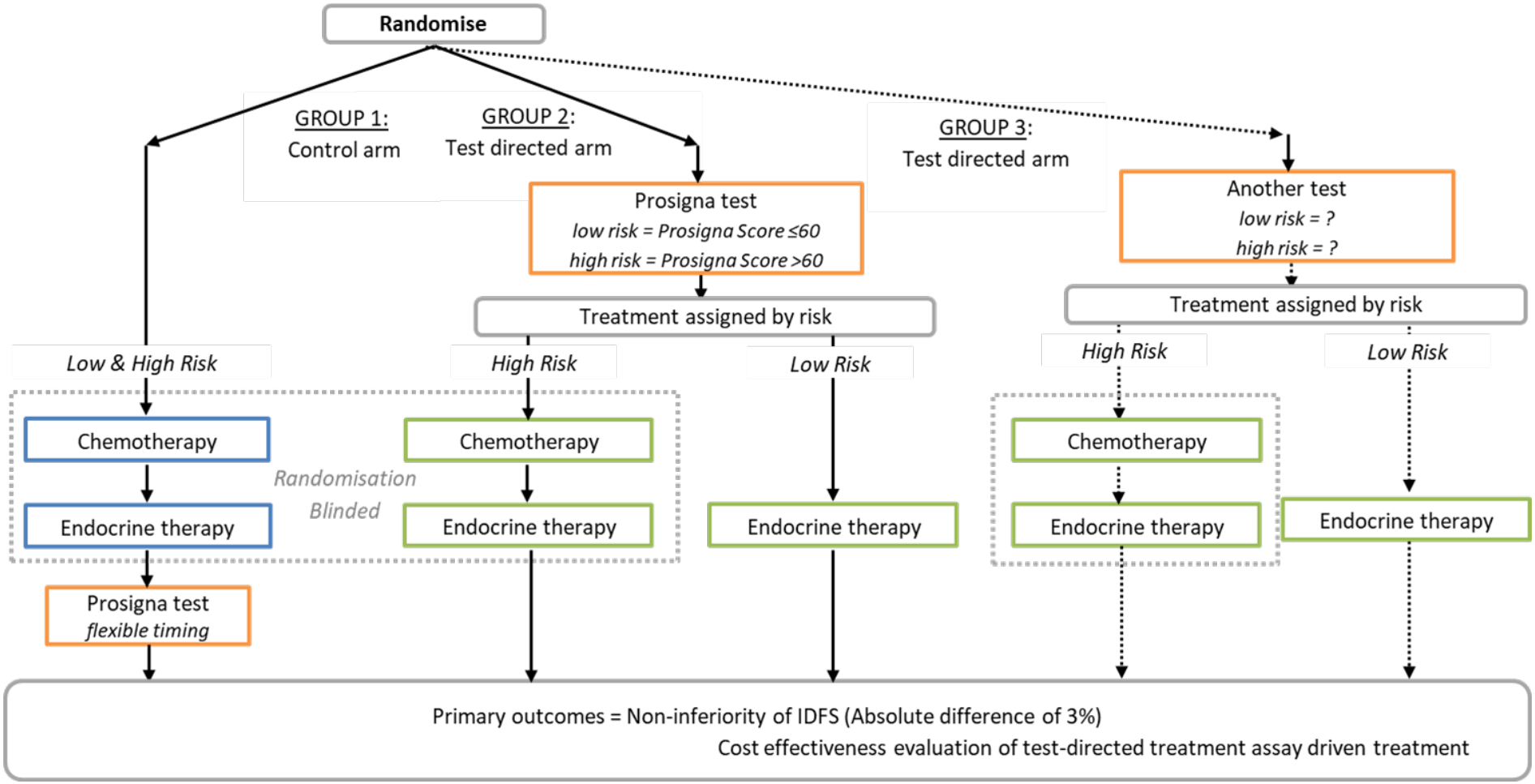

Figure 4.1: Proposed trial schema to add more multi parameter test to the Optima Trial (Stein et al. 2016). IDFS – Invasive disease-free survival.

4.2.8 Persephone – Pre-planned interim analysis

The Persephone trial was a randomised phase III trial comparing two durations of Herceptin; six months versus 12 months to understand how well they work in treating women with HER2-positive early breast cancer with a NI design (Earl et al. 2019, Earl et al. 2020). 4000 patients were required to evaluate an absolute NI margin of 3% for DFS for six months Herceptin verse 12 month Herceptin with 85% power and 5% 1-sided significance, assuming a 4-year disease-free survival rate of 80% on the 12-month control arm.

Initially the timing of randomisation was from the start of treatment however this was changed to any time between diagnosis and up to having received six months of Herceptin. At the beginning of the trial, only intravenous therapy administration was allowed, however this was subsequently changed to allow the use of subcutaneous administration as this form was introduced nationally. Patients may prefer subcutaneous administration as it takes significantly less time than intravenous therapy. Furthermore an unplanned interim analysis of the data was presented to the DMC following the results from the Herceptin Adjuvant (HERA) trial and Protocol of Herceptin Adjuvant with reduced exposure (PHARE) trials (Goldhirsch et al. 2013, Pivot et al. 2013). The DMC found no reason to change the conduct of the Persephone trial and encouraged the continuation of the trial without any change. The trial closed to recruitment with 4088 patients randomised. The trial published the primary outcome results and is now in long term follow-up (Earl et al. 2020). A meta-analysis is planned with other trials that have assessed the optimal duration of Herceptin in treating patients diagnosed with early breast cancer (Joensuu et al. 2009, Pivot et al. 2013, Mavroudis et al. 2015, Joensuu et al. 2018).

More information about this trial is provided in chapter seven.

4.2.9 PET-NECK – Change in eligibility, follow-up extension

Pet-Neck was a two-arm pragmatic phase III RCT assessing whether PET-CT (Position emission tomography - Computed tomography) guided need for surgery

was not inferior to immediate neck dissection in head and neck cancer patients in terms of OS (Mehanna et al. 2016).

560 patients were required to evaluate an absolute NI margin of 10% assuming a survival rate of 75% at two years in the control arm (planned neck dissection) with 90% power using a 5% one-sided significance level with 3 years recruitment and 2 years follow-up. Only 111 of the 140 deaths required for the analysis had been observed towards the end of the original 2 year follow-up period and therefore the follow-up period was extended to 5 years to reach the target number of events and planned 90% power. If additional follow-up had not been sought, then the power would be 76% (Mehanna et al. 2017).

During the conduct of the trial, the eligibility criteria were broadened to allow occult head and neck tumours as well as allowing neck dissection surgery to be performed either before or after chemoradiotherapy. Previously it was allowed only before chemotherapy. The reason for this amendment was that a change of practice was taking place at some centres during the trial and thus would have prohibited recruitment. Both the randomisation and primary analysis were stratified by the intended timing (before and after chemotherapy) to minimise bias and adjust for this choice in the comparison across trial arms.

4.2.10 Select-D – Sample size decrease

The Select-D trial was a randomised phase III pilot trial of Dalteparin and Rivaroxaban for the treatment of cancer patients. In addition, there was a second double-blind randomisation assessing the duration of anticoagulation treatment (Young et al. 2018). The trial had a pre-planned safety analysis which required 220 patients (110 on the Dalteparin arm and 110 on the Rivaroxaban arm) to detect an excess of 10% assuming a rate of 5% on the control arm with 80% power and 5% one-sided significance level. The trial was designed as a roll through into the full phase III trial if pre-defined criteria were met and sufficient funding was available.

The sample size was initially calculated to recruit 530 patients to provide estimates of venous thromboembolism (VTE) recurrence rate (primary endpoint) to be

within a width of the 95% CI of 8% assuming the six months VTE recurrence rate is 10%.

The second randomisation was closed on the grounds of futility as an insufficient number of patients were being randomised and therefore it would never reach target. The closure of the second randomisation resulted in a reduction of the total sample size as initially larger numbers were required in the first randomisation to ensure sufficient numbers of patients went through to the second randomisation. The sample size was reduced to 400 patients, which would allow the primary endpoint to be estimated within a width of the 95% CI of 9% assuming the six months VTE recurrence rate is 10%. Furthermore, the eligibility criteria were changed to exclude patients with primary oesophageal or gastro-oesophageal cancer based on safety, which was recommended by the DMC.

4.2.11 TEAMM – Sample size increase

The Tackling Early Morbidity and Mortality in Myeloma (TEAMM) trial was a phase III, randomised double-blind, placebo controlled trial assessing the use of levofloxacin as prophylaxis to reduce febrile episodes or deaths.

The primary endpoint of the trial was time to first febrile episode or death. An estimated proportion of 30% of patients experience a febrile episode or death within the first three months; the trial hypothesised that the use of prophylactic levofloxacin would reduce this to 20%. This required an initial sample size of 800 patients to detect a 10% difference with 90% power using a 5% two-sided significance level. This sample size was re-estimated to 1000 patients to allow for the detection of smaller difference of 8% instead of 10% with 80% power and to make reasonable adjustment for drop-outs. Although a 10% difference was meaningful with the initial expected recruitment rate, an 8% difference was felt necessary to change clinical practice. This decision was made prior to the completion of the original recruitment target, due to the quicker than expected recruitment rate and availability of study drugs, and argued that it would provide more power for the assessment of the secondary outcomes (Drayson et al. 2019). Furthermore, there were changes to patient eligibility to only include patients that

were no more than 7 days into starting anti-myeloma treatment instead of 14 days starting prior treatment which made the trial available to considerably more patients.

4.2.12 VICTOR – Early trial closure

VICTOR was a phase III, randomised double-blind, placebo controlled trial assessing the use of rofecoxib (VIOXX) in preventing recurrence in colorectal cancer patients following potentially curative therapy. The trial required 7000 patients to detect an increase in survival with a relative difference of 20% (HR = 0.80) with 85% power using a 5% two-sided significance level. This was a 2x2 factorial also assessing duration of two years versus five years rofecoxib.

Rofecoxib was withdrawn worldwide by Merck & Co. Inc. due to the Adenomatous Polyp Prevention on Vioxx (APPROVe) trial (Bresalier et al. 2005) finding an increased relative risk of cardiovascular incidents. This led to the early closure of the trial and all ongoing treatment stopped on the grounds of safety (albeit no safety concerns were raised by the VICTOR DMC). The statistical analysis plan was modified and the 2434 patients randomised were followed-up and the power calculations were modified with the trial now able to detect an increase in survival with a relative difference of 25% with 87% power using a 5% two-sided significance level (Midgley et al. 2010).

4.2.13 Summary

This scoping exercise of cancer trials within the Warwick CTU demonstrated that both superiority and NI hypotheses were implemented and that adaptive methods are regularly executed without classing them as adaptive designs. A summary from this exercise can be seen in Table 4.1. The key learnings are that broadening the eligibility in trials can help increase the rate of recruitment and allow subgroup analyses to be proposed if the criteria are being extended to specific populations. A time dependent indicator could be applied to indicate when these changes took place as a form of sensitivity to measure the impact of the change in eligibility and see if sufficient numbers were recruited. None of the trials have reported the impact the change in eligibility had on recruitment. Sample size re-estimation

methods were employed when new external evidence became available to ensure that the most appropriate assumptions for the control arm were used. When recruitment was doing well, sample sizes were increased to consider smaller differences and allow for subgroup analyses as well as accounting for drop-outs. Group sequential methods with pre-planned stopping rules for interim analyses allow the possibility of stopping the trial early on grounds of futility or efficacy, which can save time and money.

The scoping exercise was an exhaustive investigation for all oncology trials conducted at Warwick CTU and was further extended to consider exemplar adaptive design trials. This exercise has further supported the proposed extension to the CONSORT 2010 guidelines as concluded from the literature review reported in the previous chapter as there were many trials that implemented some form of adaptation. It is important to note these adaptations in the reporting of the trial results.

4.3 Examples of current adaptive designs trials outside of the Warwick CTU

The scoping exercise was extended to consider examples outside of Warwick CTU of adaptive trials within Oncology that were known to the CTU to explore how these trials were designed and are being conducted. The adaptive trials identified were STAMPEDE and FOCUS4 trials which are both being conducted by MRC CTU at University College London, CompARE (Comparing Alternative Regimens for escalating treatment of intermediate and high-risk oropharyngeal cancer) trial conducted by Cancer Research UK CTU in Birmingham, I-SPY 2 trial conducted by Quantum Leap Healthcare Collaborative in San Francisco and PARTNER (Platinum and PARP inhibitor for Neoadjuvant treatment of Triple Negative and/or BRCA positive breast cancer) trial conducted by the Cancer Research UK Cambridge Institute.

4.3.1 STAMPEDE trial

The STAMPEDE trial is a phase II/III RCT, which initially began as a MAMS trial and later evolved into a platform trial with the addition of six treatment arms in 2011[1].

[1] http://www.stampedetrial.org

This trial is for patients with locally advanced prostate cancer who are commencing long-term Androgen Deprivation Therapy (ADT) (STAMPEDE trial 2005, Sydes et al. 2009, Sydes et al. 2011, Sydes et al. 2012, Choodari-Oskooei et al. 2020). This trial is one of the largest exemplars of an innovative MAMS platform design (Carthon et al. 2016) as it allows for the evaluation of multiple arms at different interim stages and the addition and termination of experimental treatments during the trial (see section 2.10 on platform trials). This innovative design implemented by the STAMPEDE team has been used as an exemplar of adaptive design methodology by the FDA because it has efficiently evaluated several treatments simultaneously compared to several individual trials (US Food and Drug Administration 2018) for which there are also cost saving implications.

Inclusion criteria for the trial involves patients being either high risk newly diagnosed non-metastatic patients or newly diagnosed metastatic patients or those treated with prior radical surgery and/or radiotherapy. The primary outcome of STAMPEDE is OS with the intermediate outcome being FFS. All sample sizes were based on a final stage power of 90% to detect a hazard ratio (HR) of 0.75 with a one-sided 2.5% significance level.

This trial began in 2005 with one control arm (Androgen deprivation therapy ± radiotherapy), which was continuously extended, versus five experimental arms (Arms B, C, D, E, and F) with subsequent additions of four extra arms at various points throughout the trial; Arm G – introduced in November 2011, arm H – introduced in March 2013, arm J – introduced in January 2014, arm K – introduced in January 2016 and arm L – introduced in June 2017. The control arm consists of contemporaneous patients, meaning that patients allocated to a new experimental arm are compared to control arm patients randomised after the new experimental arm was introduced. To date, the control arm has changed twice; first in 2017 to abiraterone then again in 2019 to radiotherapy for metastatic patients.

The trial was powered to detect a difference in relative improvement at each of the interim analysis stages by performing pairwise comparisons for each of the initial five experimental arms against the control arm when 113, 216 and 334 PFS

events were reported in the control arm. More emphasis was on controlling the PWER rather than the FWER because different hypotheses were being tested in each of the five experimental arms (Bratton et al. 2016). Stopping rule guidelines for discontinuing the specified experimental arms were produced (Table 4.2). At each increasing stage, the significance level becomes more stringent and requires a higher power. If the null hypothesis (i.e. effect of experimental arm treatment is the same or worse) is rejected then the treatment continues to the next stage where a greater amount of evidence is gathered (Sydes et al. 2009). At each stage the experimental treatments need to show evidence of a benefit over the control arm to continue to the next stage, i.e. the obtained HR should be below the critical HR to continue. It was approximated that for the first five experimental arms between 2800 and 3600 patients were required to be recruited within seven years. The primary outcome analysis for the first five experimental arms was planned when 403 deaths in the control arm have been observed.

After the second interim analysis in April 2011, the evidence suggested that there was no evidence of benefit in arm D (HR = 0.94), which led to the DMC recommending discontinuing treatments in arms D and F (both arms used celecoxib) (James et al. 2011). Subsequently the additional research arm G was added in November 2011 and had equal allocation to control. The guidelines for stopping arm G were different to arms B-F (Table 4.2). The accrual of patients would be stopped when either 1500 patients were recruited or after three years. The primary outcome analysis would be performed when 267 deaths were observed in the control arm. The sample size for this arm increased from 1500 to 1800 due to the proportion of non-metastatic patients expected to be recruited was higher.

In March 2013 arm H was activated, which only recruited those patients with newly diagnosed metastatic cancer that will be on ADT for the first time. Experimental arm H had a similar approach to arm G; it would take the same number of events to trigger the interim analyses at the different stages (Table 4.2). The experimental arm H was applicable to around 60% of the STAMPEDE patient population. It was anticipated that around 1250 patients were required over 4

years of recruitment to observe 267 control arm deaths, which was expected to be reached 5.25 years after the activation of arm H.

For the experimental arm J, which was activated in January 2014, two interim analyses were planned before reaching the final stage (Table 4.2). It was anticipated that it would require approximately 1800 patients within 3.5 years of recruitment to observe 267 deaths within six years after the activation of arm J.

The experimental arm K was activated in January 2016. The key inclusion criteria for randomisation between the experimental arm K versus the control arm were patients who were non-diabetic. The timing of the interim and final analysis was determined by the number of deaths driven by patients with metastatic cancer (Table 4.2). OS was used as both the intermediate outcome measure and the final primary outcome measure. It was anticipated that it would require approximately 1800 patients (1100 would be patients with metastatic cancer) recruited over three years to observe 473 deaths, which would trigger the primary outcome analysis at approximately eight years after the activation of arm K.

Guidelines for comparison of control arm to experimental arms B-F					
Stage	**Sig. level**	**Power**	**Target HR**	**Number of control arm events**	**Consider discontinuation if HR_1 (observed) is…**
Activity Stage I	0.500	95%	0.75	~113 FFS	>1.000
Activity Stage II	0.250	95%	0.75	~216 FFS	>0.920
Activity Stage III	0.100	95%	0.75	~334 FFS	>0.890
Efficacy Stage IV*	0.025	90%	0.75	~403 OS	>0.850
Guidelines for comparison of control arm to experimental arms G and H					
Activity Stage I	0.500	95%	0.75	~75 FFS	>1.000
Activity Stage II	0.250	95%	0.75	~142 FFS	>0.920
Activity Stage III	0.100	95%	0.75	~221 FFS	>0.890
Efficacy Stage IV*	0.025	90%	0.75	~267 OS	>0.850
Guidelines for comparison of control arm to experimental arms J					
Activity Stage I	0.400	95%	0.70	~66 FFS	>0.957
Activity Stage II	0.120	95%	0.70	~139 FFS	>0.869
Efficacy Stage III*	0.025	90%	0.75	~267 OS	>0.85
Guidelines for comparison of control arm to experimental arms K					
Activity Stage I	0.40	92%	0.80	~121 M1 deaths	>0.965
Efficacy Stage II*	0.025	92%	0.80	~473 M1 deaths	>0.881
Guidelines for comparison of control arm to experimental arms L					
Activity Stage I	Completed in the PATCH trial (2013) with 206 PFS events.				
Activity Stage II	Planned formal interim upon reaching 410 PFS events in the control arm				
Efficacy Stage III*	Performed upon reaching approximately 565 deaths and 815 PFS events				

*Outcome of treatment concluded at this stage

Table 4.2: Guidelines for interim analysis at each of the activity stages (http://www.stampedetrial.org).HR = Hazard Ratio, M1 = Patients diagnosed with new metastatic cancer. FFS = Failure free survival; OS = Overall Survival.

The experimental arm L was added around June 2017 (Gilbert et al. 2018). The inclusion criteria for this comparison included patients starting long-term hormone therapy for high-risk non-metastatic or metastatic prostate cancer that have had eight weeks of anti-androgens and no more than approximately a months' worth of luteinizing hormone releasing hormone. This comparison allocates eligible participants to the control arm or arm L (combination of transdermal oestradiol ± radiotherapy ± docetaxel). The unique aspect of this comparison is that the evaluation of clinical efficacy will be based on a combined meta-analysis using data from the Prostate Adenocarcinoma: Transcutaneous Hormones (PATCH) trial. The STAMPEDE trial had originally planned for two activity stages, of which the first activity stage was incorporated in the PATCH trial in 2013 when 206 PFS events were reached, which led to the PATCH DMC recommending further recruitment into a phase III trial[1]. A formal interim analysis was planned for activity stage II when approximately 410 PFS events were observed in the control arm for both the PATCH and STAMPEDE trial, however, the number of events were accrued slower than originally anticipated and therefore this was removed from the analysis schedule (Langley 2020). The final comparison requires approximately 2000 patients in total (500 patients from STAMPEDE) with co-primary endpoints of PFS and OS. A breakdown of the STAMPEDE trial can be seen in Figure 4.2.

To date, for arms B to F, the primary results have been published and all active follow-up has been discontinued (James et al. 2016, Mason et al. 2017). For arm G and H, the primary results have been published and the trial remains on active follow-up to allow for long-term analysis (James et al. 2017, Parker et al. 2018). Arm J is in follow-up and arms K and L were both recruiting but were temporarily paused due to the COVID-19 pandemic[1, 2].

[1] http://www.stampedetrial.org
[2] https://www.ctu.mrc.ac.uk

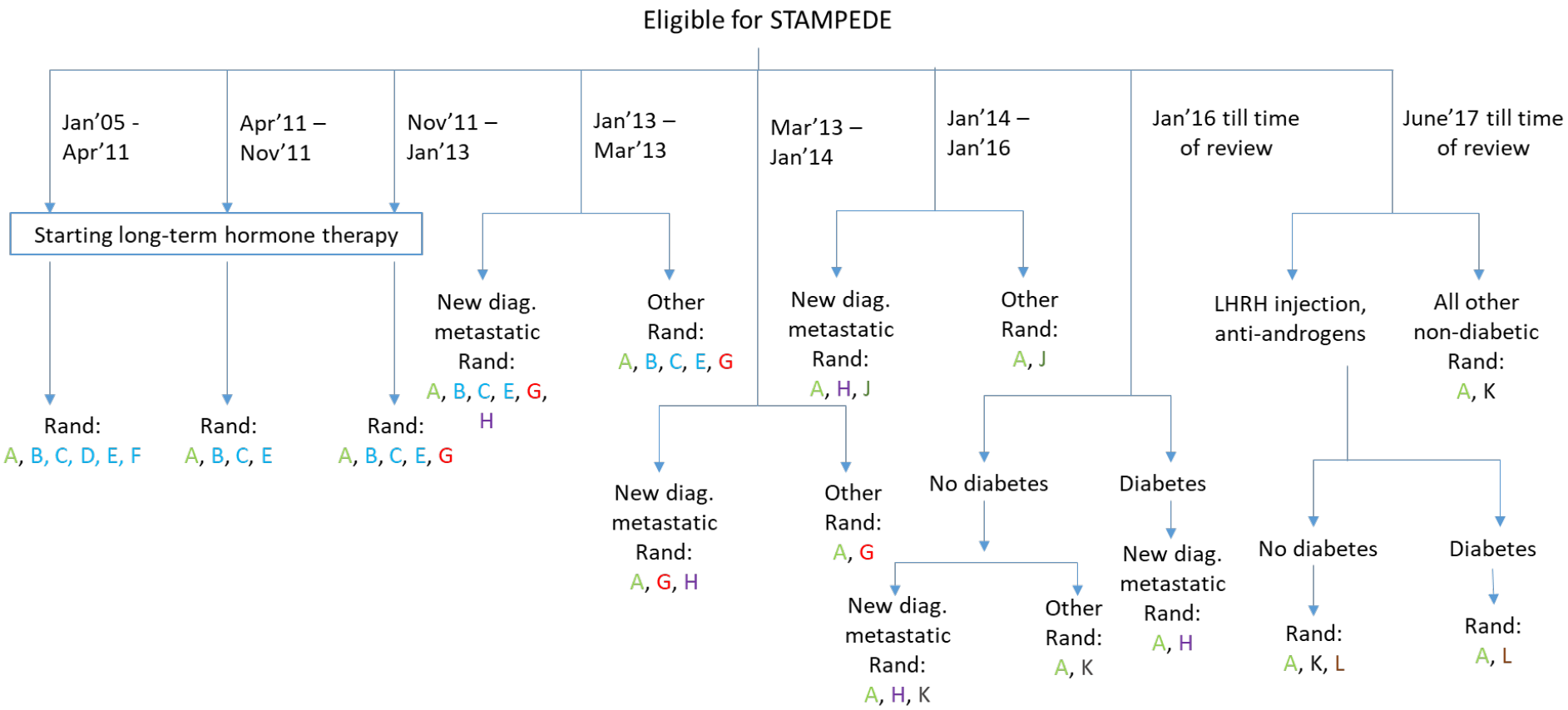

Figure 4.2: Diagram showing the STAMPEDE trial design over time. Green (A) – control arm, blue (B, C, D, E, F) – First experimental arms, other colours (G, H, K, J, L) – Added experimental arms.

4.3.2 FOCUS4

The FOCUS4 RCT utilises an umbrella MAMS design, including biomarker-stratified and non-stratified comparisons within one master protocol (see section 2.10) for patients with advanced or metastatic colorectal cancer (FOCUS4 2019). The primary outcome for this trial is PFS and an additional outcome of OS may be considered at the final stage. The intermediate outcome at each stage is also PFS.

At the initial stage, all patients receive standard first-line chemotherapy treatment for approximately 16 weeks. During this period, biomarker testing is performed on the original tumour to identify which drug/s would be best suited and hence patients are stratified into one of five cohorts.

The FOCUS4 trial utilises a MAMS design whereby there are four stages at which interim analyses will be performed on each of the first five cohorts. The reason for the analyses at each stage are as follows: stage I - safety, stage II – lack of sufficient activity, stage III – efficacy for PFS and a potential stage IV – efficacy for OS. The decision to continue the trial to assess stage IV is dependent on the resources required to achieve the necessary sample size and follow-up along with the adequate supply of the treatment. Stages I and II are regarded as a phase II study with stages III and IV classed as a phase III study. An example of the operating characteristics for a specific cohort can be seen in Table 4.3.

Parameters	**Stage I**	**Stage II**	**Stage III**	**Stage IV**
	Safety	**LSA**	**Efficacy PFS**	**Efficacy OS†**
	Phase II		**Phase III**	
One-sided Alpha	0.3	0.1	0.025	0.025
Power	0.95	0.95	0.95	0.85
Target HR	0.65	0.65	0.65	0.70
Critical HR	0.91	0.83	0.79	0.80

†Stage IV only undertaken if sufficient control arm events reached.

Table 4.3: Possible generic operating characteristics for possible comparisons in all wildtype cohort. LSA = Lack of Sufficient Activity; PFS = Progression free survival; OS = Overall Survival; HR = Hazard ratio.

The duration of recruitment is seven years with the aim to recruit over 1500 patients in total. Timing of interim analyses at each stage is determined by a maximum number of events required. Continuation on to the next stage is based on reaching a pre-specified critical HR, and corresponding maximum number of patients for each

stage required. The trial implements an allocation ratio of 2:1 in favour of the experimental arm as this allows for more information to be obtained on early safety and toxicity. A summary of this information can be found in Table 4.4. Specific details for each comparison have been detailed in separate protocols (FOCUS4 2019).

Molecular cohort	Randomised allocation ratio	Phase	Outcome and stage	Target HR	Max number of events required: total (control arm)	Estimated cumulative analysis time (months)	Max number of pts required
BRAF mutation	2:1	2	PFS - I	0.5	41 (16)	20.4	61
			PFS - II	0.5	76 (28)	32.5	97
		3	PFS - III	0.5	118 (42)	46.5	139
			OS - IV (potential)	0.65	217 (79)	100.4	301
PIK3CA mutation and/or PTEN loss	2:1	2	PFS - I	0.65	107 (40)	17.0	170
			PFS - II	0.65	197 (71)	26.5	264
		3	PFS - III	0.65	303 (107)	37.2	373
			OS - IV (potential)	0.7	289 (109)	54.6	546
KRAS or *NRAS* mutation	2:1	2	PFS - I	0.65	109 (41)	16.1	177
			PFS - II	0.65	198 (72)	22.8	273
		3	PFS - III	0.65	302 (107)	31.4	378
			OS - IV (potential)	0.7	287 (109)	50.6	574
All wild type	2:1	2	PFS - I	0.65	109 (41)	20.0	180
			PFS - II	0.65	198 (72)	30.6	275
		3	PFS - III	0.65	301 (107)	42.3	381
			OS - IV (potential)	0.7	289 (109)	60.8	547

Table 4.4: Summary of the operating characteristics and timelines for each of the cohorts, extracted from FOCUS4 protocol (FOCUS4 2014). OS – Overall survival, PFS – Progression free survival; HR = Hazard ratio.

To date, recruitment has been suspended to the trial in March 2020 due to the COVID-19 pandemic and follow-up closed to all patients in October.

4.3.3 CompARE trial

The CompARE trial is a MAMS phase III RCT trial investigating which treatment is the most effective for patients who have high-risk oropharyngeal cancer (Mehanna 2015). The primary outcome is OS and the intermediate outcome measure is DFS. It was anticipated that approximately 650 patients would be recruited into the trial.

Patients were originally randomly allocated into either the control arm or one of three experimental arms:

- Arm 1: Concomitant cisplatin chemotherapy plus radiotherapy (Control arm)
- Arm 2: Induction chemotherapy followed by arm 1
- Arm 3: Dose-escalated radiotherapy plus concomitant cisplatin
- Arm 4: Resection of primary and selective neck dissection followed by arm 1

The timing of the interim analyses is determined by the number of events in the control arm. Analysis for stage I is planned after 70 control DFS events have been reported, analysis at stage II will be after 114 control DFS events and analysis at stage III will be after 169 control DFS events. The final analysis in the trial is scheduled to take place when 128 control patients have died in the trial; it is anticipated that the analysis of the primary outcome will be analysed 6.6 years after the start of recruitment. Table 4.5 below gives information about the power and significance level at each of the stages.

Stage	Number of control arm events	Power	1-sided significance level	Consider discontinuation if HR_1 (observed) is…
Stage I	70 DFS	95%	0.5	>1.00
Stage II	114 DFS	95%	0.3	>0.92
Stage III	169 DFS	95%	0.15	>0.87
Final analysis	128 deaths	85%	0.1	

Table 4.5: Summary of the operating characteristics at each of the stages and final analysis. DFS – Disease free survival, HR – Hazard ratio.

A new arm has subsequently been added (durvalumab and chemoradiotherapy followed by durvalumab) and the original three experimental arms have been closed to recruitment (Cancer Research UK 2020) but no detail results have been published as to why these arms have closed. The Birmingham trials unit were contacted but no correspondence was received.

4.3.4 I-SPY 2

The I-SPY 2 is an adaptive phase II platform RCT that investigates the efficacy of drugs in combination with chemotherapy with the efficacy of standard therapy alone in the neo-adjuvant setting for women with large primary cancers of the breast (>3 cm)

(Barker et al. 2009). The primary outcome of the trial is the measure of pathologic complete response (pCR). The primary aim of the trial is to demonstrate the usefulness of specific treatment regimens for different patient subgroups based on their biomarker signatures.

The adaptive method of the study implements the use of Bayesian adaptive randomisation whereby specific drugs that show higher efficacy will be favourably assigned to patients. Bayesian predictive probabilities will be implemented at interim points, if treatments reach a pre-specified level of efficacy then it will 'graduate', otherwise after reaching a maximum sample size, recruitment to the treatment arm may be stopped for futility. At this point the randomisation probability will be revised; Figure 4.3 shows the procedure of how the I-SPY 2 process will work.

Implementing a Bayesian framework allows the trial to update the probabilities based on current data allowing more patients to be allocated to treatments that show promise and therefore accelerating the process of 'graduating' favourable treatment arms. This process continually allows new treatments to be added and evaluates favourable treatments efficiently. All treatment arms have a maximum sample size, where upon reaching this maximum, it would trigger the arm to be tested for futility. A limitation to this approach is that recruitment to poorly performing treatments will continue until a maximum sample size is reached. A solution would be to incorporate a futility boundary whereby at an interim stage if this boundary is crossed, recruitment to poorly performing treatments would be stopped.

To date, the I-SPY 2 trial continues to recruit to seven treatment arms but recruitment to 15 treatment arms has finished, of which the trial results for two treatments are pending, phase III trials have been conducted to evaluate two of the treatments and are ongoing for a further three treatments and recruitment was stopped early for the remaining eight treatments because of futility or toxicity[1].

[1] www.ispytrials.org

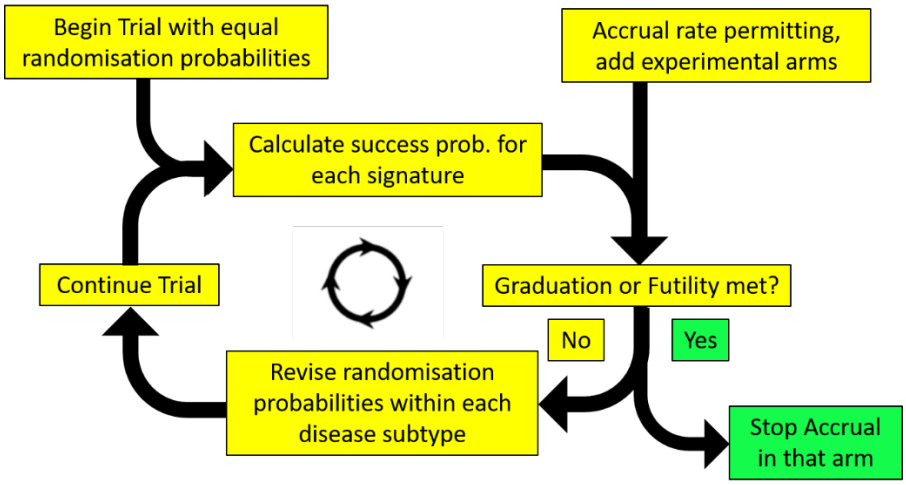

Figure 4.3: Flow chart showing I-SPY 2 adaptive process (Berry 2015)

4.3.5 Partner

The Partner trial is a phase II/III three stage multi-centre RCT that "evaluates the safety and efficacy of the addition of Olaparib to platinum-based neoadjuvant chemotherapy in breast cancer patients" (Abraham 2017). The trial aimed to recruit at least 527 patients over the course of three stages with a minimum of 220 germline breast cancer gene (gBRCA) patients recruited. Patients are randomly allocated to one of three treatment arms (control arm and two experimental arms) in a 1:1:1 ratio at stages I and II and 1:1 at stage III. The trial implements a 'pick the winner' design (see section 2.7 for more information) whereby at stage II, one of the two experimental arms will be the 'winner' and hence taken forward into stage III. The concepts applied to this design are like that of a MAMS design because there are multiple arms and the decision to drop an arm is based on the results at an interim stage. The duration of the trial has not been given however patients will receive 21 weeks of chemotherapy followed by breast surgery after which patients will be followed up for ten years.

Stage I will determine the safety of the two research arms. The total sample size required at this stage is 75 patients (25 patients in each arm). Safety analysis will be performed when the first 25 patients in each of the research arms have received at least one dose of Olaparib, if the dose needs modifying then those initial patients will not be analysed in the final phase III primary analysis.

At the end of stage II, one research arm will be selected to continue onto stage III. The sample size for stage III will be 159 patients (53 patients in each arm). The outcome for stage II is the pCR rate and the analysis will be performed when this rate is available for 53 patients in each of the research arms. The research arm that will be selected to continue to stage III will be determined by the arm with the highest observed pCR rate.

At the end of stage III, an efficacy analysis using the pCR rate will be performed. A total sample size of 444 patients is required with equal allocation for stage III between the control arm and the selected experimental arm to detect an improvement of 15% from approximately 45-55% in the control arm for all patients and a 20% improvement for gBRCA patients, with 80% power and an overall 5% two-sided significance level. The trial has incorporated a futility analysis for the gBRCA patients and plans to create a set of stopping guidelines to terminate the trial on the grounds of efficacy.

The DMC recommended that the trial continue without any change at stage one. In April 2018, recruitment to stage two was completed, the results for this were to be reviewed early in 2019 but no results have been published (Abraham et al. 2019).

4.3.6 Summary

This exercise examined some RCTs that have implemented adaptive design methods. A summary of the key points from these examples can be seen in Table 4.6. All trials shared a common theme to ensure correct decisions are made at the right time in an efficient manner. This was demonstrated by the multi-arm designs utilised in all trials by ensuring experimental arms that showed promise continued through the trial. The STAMPEDE, FOCUS4 and CompARE trials have implemented the MAMS design using the 'nstage' command in Stata (Barthel et al. 2009, Blenkinsop et al. 2019). For every stage, a high power has been used (≥ 85%) and all significance levels at intermediate activity stages are lenient but with a strict significance value used at the final stage to determine efficacy. A different outcome measure was used for the intermediate activity stages compared to OS, which was used at the final efficiency stage for the STAMPEDE and CompARE trial. All three of these MAMS trials have evolved to

'platform' trials by adding more treatment arms however only the STAMPEDE trial is labelled with this terminology. Additional comparisons are treated like a two-arm multi-stage trials to obtain the control arm events required to trigger the analysis at the end of each stage (Sydes et al. 2011). The repeated contributions of control arm patients demonstrate the efficiency and perhaps cost-savings involved from utilising these designs. Furthermore, allocating more patients to the experimental arm allows for more information to be obtained to take decisions with regards to safety and toxicity as seen in the FOCUS4 trial.

This exercise demonstrates that implementing adaptive trials are ideal as they are efficient and effective, however the resource required to execute these large trials cannot be underestimated. As mentioned in chapter two, the work of Hague et al (2019) and Schiavone et al (2019) must be taken into consideration so Clinical Trial Units can fully understand the resource and forward planning required prior to any thought of implementing adaptive trials of this scale.

Trial	Summary
STAMPEDE	Implemented a platform MAMS design. Selected values for stagewise significance 0.5, 0.25, 0.1, 0.025 regularly used for the comparisons. Stagewise power value of 95% and final value power of 90% regularly used. Incorporated a new treatment arm with the aim to perform a meta-analysis using another trial. Exemplar for adaptive design methodology as shown in latest FDA guidance.
FOCUS4	Incorporates an Umbrella MAMS design. Trial uses a master protocol as within this trial there are multiple trials with their own designs. Selected values for stagewise significance of 0.3, 0.1, 0.025, 0.025. Stagewise power value of 95% and final power of 85%. Allocation ratio 2:1 in favour of the experimental arm.
Compare	Implements a MAMS design with one arm as a surgical arm. Uses same stagewise power and final power values as the FOCUS4 trial.
I-SPY 2	Incorporates a Bayesian Adaptive randomisation whereby drugs that show promise will be allocated more patients. Interim analyses performed to make the decision of whether to continue or stop the drugs.
Partner	A three arm seamless (phase II/III) trial incorporating a pick-the-winner approach - like a MAMS whereby one of the two experimental arms will not continue onto the phase III trial.

Table 4.6: Summary from investigating examples of adaptive trials within Oncology.

4.4 Discussion

This chapter captured adaptive methods applied within oncology trials at Warwick CTU but also considered exemplar adaptive trial designs external to Warwick CTU. The Warwick CTU scoping exercise found four trials (ARTemis, AVAST-M, Neo-Escape and Persephone) implemented interim analysis due to safety, efficacy or futility. Five trials (COUGAR-02, LIHNCS, Select-D, TEAM and VICTOR) changed the sample size which was not pre-planned. The VICTOR trial was unique as it made ad-hoc changes to the trial due to the withdrawal of the study drug. The Persephone trial ran simultaneously with other similar trials (more information in chapter seven) investigating the use of Herceptin duration hence the release of results for these other sister trials influenced the conduct of the Persephone trial. The Optima trial is the only trial at Warwick CTU that explicitly states that it is an 'adaptive design' as this trial plans to incorporate additional multi-parameter tests to compare against a contemporaneous control, it can be considered as a platform trial. Pre-planned adaptations should be strongly encouraged although the scoping exercise found that ad-hoc changes may be necessary to adapt to the current evidence and situation.

All the exemplar trials considered in section 4.3 have multiple arms and have incorporated at least one interim analysis for safety, efficacy or futility. The STAMPEDE trial is the only trial labelled as a platform trial (section 2.10), whereas CompARE and I-SPY 2 trial have added new treatments into the trial like a platform design. I-SPY 2 is the only trial out of the examples in section 4.3 that is based in the pharmaceutical industry. There are other adaptive design based trials within the pharmaceutical industry, for example within Janssen Pharmaceuticals there are two actively running platform trials and a basket trial within oncology (Kyle 2021).Within Janssen Pharmaceuticals, there is an active support group that helps with the design and execution of adaptive designs and internal presentations are conducted related to adaptive designs to encourage and create an awareness of the advantages of implementing adaptive designs.

It can be clearly seen from all these adaptive trial examples that detailed planning is required to implement the novel designs. For example, within the MAMS designs the number of stages, the significance level and power at each stage, the intermediate

and primary outcome, the alternate hypothesis needs to be pre-specified. These parameters will determine the number of control arm events required and the timing at each stage. Furthermore, when deciding on the choice of outcome measures, it is assumed that the intermediate outcome measure occurs more frequently than the primary outcome measure, however there can be situations where the intermediate outcome is the same as the primary outcome (Sydes et al. 2009). The intermediate outcome measure provides a base for the amount of 'activity' emerging whereas the primary outcome is the base for definitive conclusions to be made. It is important to perform a variety of calculations for the sample size as within MAMS trial some parameters are fixed and some can be variable which can affect the power, significance level and sample size. For example, the number of arms that are available to trial stakeholders may be fixed however the rate at which patients are accrued or the relative difference between treatments can vary.

It is evident that multi-arm methods, such as variations of the MAMS design (Wason 2015), or seamless phase II/III design that assess multiple treatments or drop the loser/pick the winner design in the field of Oncology are increasingly popular. Parmar et al urges that more multi-arm randomised trials should be implemented in all areas of clinical trials as multiple treatments can be assessed simultaneously, which can result in a simpler, quicker and cheaper trials in comparison to several standard two-arm trials (Parmar et al. 2014). The MAMS trials discussed in this chapter involved poor prognostic populations, for example the STAMPEDE trial assesses OS in a patient population with metastatic prostate cancer, where events can accrue quickly. However, within early breast cancer trials, time to event outcomes require much longer follow-up, typically greater than five years. Hence, if there were multiple treatments that needed to be evaluated within breast cancer, would a MAMS design still be more efficient than other trial designs to answer the same number of questions? How would a MAMS design change with a NI hypothesis rather than assessing superiority?

In my next chapter I will investigate the use of MAMS designs in cancer sites with varying survival rates with a superiority hypothesis.

5 Comparison of hypothetical MAMS designs with long-term outcomes

5.1 Introduction

A MAMS design, as mentioned in chapter two, consists of pairwise comparisons of multiple experimental arms against a single control arm. These designs are increasing in popularity due to their efficiency and effectiveness in testing several different treatments at once and are supported by patients and clinicians willingness to participate in these studies (Parmar et al. 2014). These MAMS designs have been employed in trials such as STAMPEDE, CompARE and FOCUS4 (James et al. 2008, Shiu et al. 2013, Mehanna 2015), all of which have been discussed in chapter four. The MAMS trials have up to now, as seen in the exemplar studies presented in previous chapter, been conducted in poor prognostic populations where the events accrue quickly, for example the STAMPEDE trial assesses OS in a patient population with metastatic prostate cancer that has a five-year OS rate of approximately 50% (Sydes et al. 2009). Hence, the application of a MAMS design in a good prognostic population e.g. breast cancer (five-year OS rate ≈ 85%), where there is a lower chance that a patient will experience an event(s) has not been explored.

Various hypothetical MAMS designs were evaluated within this chapter in three cancer sites: breast, colon and lung cancer, primarily focusing on the longer-term outcome of OS. These three cancer sites have the highest incidence rates but varying survival rates (Cancer Research UK 2018) which helped to further understand if it was feasible to conduct MAMS designs in different scenarios. The trial designs were compared through three phases using a superiority hypothesis.

The first phase investigated the use of MAMS designs using the survival rates obtained from the Office for National Statistics (ONS) for patients in England between 2013 to 2017 and followed up to 2018 (Office for National Statistics 2019). The stagewise significance levels and power values were estimated such that an overall significance level less than 5% and power of at least 80% was reached. The aim of the first phase was to gain insight into the different MAMS designs and

evaluate within which setting a MAMS design was more feasible using a superiority hypothesis.

The second phase of the MAMS designs aimed to compare the designs obtained from the first phase with those implementing the method recommended by Royston et al (2011) for the stagewise significance levels and power for each cancer site to see which method proved to be more robust and hence most appropriate for use in the third phase.

The third phase investigated four trials that had published primary outcome results for each of the three cancer sites. The aim of this phase was to understand the feasibility of applying a MAMS design as well as understanding its efficiency compared to standard two-arm trials.

5.1.1 Choice of parameters

The Cancer Research UK website (Cancer Research UK 2018) was used to screen all phase III clinical trials that had published results related to the primary endpoint to help with the choice of parameters to be used in all three phases of the calculations. The reason for selecting breast, colon and lung cancer were because they have the highest incidence rates and distinctively vary in their OS rates, thus the results will be more heterogenous (Cancer Research UK 2018, Office for National Statistics 2019).

Trials that involved the testing of treatments and had results published in a medical journal were further explored for eligibility. The trials were eligible if the following results were provided:

- An OS rate at any time point;
- The number of deaths for the control arm;
- The minimally clinical important difference (MCID) the trial aimed to detect;
- Patient eligibility information;
- The number of treatment arms;
- The initial planned sample size, power value and level of significance;

- The actual sample size, time taken to recruit and the follow-up period.

Twelve trials were selected (four from each cancer site) that fulfilled the eligibility criteria and gave diverse results. This would help in understanding the application of MAMS designs in different trial settings. Details of these 12 trials are provided in the following section and a summary of these trials can be found in Table 5.1.

The Clinicaltrials.gov website identified 182 clinical trials that are actively recruiting within the UK for each of the selected cancer sites (ClinicalTrials.gov 2020). The number of trials and the mean sample size within each phase split by cancer site was used to calculate an estimate of the proportion of patients that entered phase III clinical trials every year. This calculation would be part of the criteria to determine the feasibility of implementing a MAMS design in sections 5.3 and 5.4. The title, phase and sample size for all 182 trials can be found in appendix 6.

5.1.2 Description of the 12 CRC UK phase III trials used to inform the choice of parameters for the hypothetical MAMS designs

5.1.2.1 Breast cancer trials

Four breast cancer trials were found that fulfilled the eligibility criteria stated in section 5.1.1. Details of these trials that were used in the MAMS design calculations are provided below.

5.1.2.1.1 DEVA trial

The Docetaxel Epirubicin Adjuvant (DEVA) trial assessed the efficacy and toxicity of combining docetaxel after Epirubicin in postmenopausal node-positive early breast cancer patients which could be classified as either luminal B or HER2 enriched patients (Coombes et al. 2011). The DEVA trial was a phase III RCT with a partial 2 x 2 factorial design. Patients were randomly assigned to either Epirubicin (control) or Epirubicin followed by docetaxel (experimental). A subset of centres gave patients the option of being randomised a second time to assess the timing of tamoxifen treatment. The TTE outcomes for this trial were DFS for the primary outcome and OS as a secondary outcome. The sample size was calculated based on the log-rank test which required 792 patients with 90% power and 5% significance level to detect an

absolute improvement from 70% (control arm) to 80% (experimental arm) for five-year DFS. The DEVA trial recruited 803 patients in approximately eight years with a median follow-up of 64.7 months. The five-year DFS and OS rate for the control arm was 72.7% and 81.8% respectively with 75 deaths reported on the control arm. The five-year DFS and OS rate for the experimental arm was 79.5% and 88.9% respectively.

5.1.2.1.2 SoFEA trial

The SoFEA trial was a phase III RCT which aimed to assess the use of steroidal anti-oestrogen fulvestrant by incorporating it with continued oestrogen deprivation in postmenopausal women with hormone-receptor-positive breast cancer, which could be classified as either luminal A or luminal B patients (Johnston et al. 2013). Patients were randomised to one of three arms: exemestane (control arm), fulvestrant plus anastrozole (FA) or fulvestrant plus placebo (FP). The TTE outcomes for this trial was PFS for the primary outcome and OS as a secondary outcome. A planned sample size of 750 patients would be able to detect an improvement in median PFS from 5.5 months in the FA arm to 7.5 months in the FP arm and detect an improvement from four months in the control arm to 5.5 months in the FA arm with 90% power and 5% two-sided level of significance. The trial recruited 723 patients in six years with a median follow-up of 37.9 months. The median PFS rates were 3.4 months in the control arm, 4.4 months in the FA arm and 4.8 months in the FP arm. The median OS rates were 21.6 months in the control arm with 173 events, 20.2 months in the FA arm and 19.4 months in the FP arm. The trial performed interim analysis for which no further details could be obtained. The MAMS design applied to this trial utilised the treatment difference between the control arm and the FA arm.

5.1.2.1.3 HERA trial

The HERA trial was a three-arm phase III RCT comparing Herceptin for one year, Herceptin for two years or observation (control arm) for patients diagnosed with HER2 positive early stage breast cancer that had completed locoregional therapy, which could be classified as either luminal B or HER2 enriched patients (Piccart-Gebhart et al. 2005, Goldhirsch et al. 2013, Cameron et al. 2017). Patients would begin the allocated treatment only after patients received either neoadjuvant

chemotherapy, adjuvant chemotherapy or both. The TTE outcomes for this trial was DFS for the primary outcome and OS as a secondary outcome. The sample size was calculated based on the log-rank test which required 4482 patients to reach 951 DFS events with 80% power and 2.5% two-sided significance level to detect an absolute improvement from 65% (control arm) to 71.8% (experimental arm) for five-year DFS. The HERA trial team performed three interim safety analyses to assess cardiac safety of Herceptin. The primary analysis of patients on Herceptin for one year against observation took place after patients had a median follow-up period of two years. The trial recruited 5102 patients in approximately 3.5 years. The three-year DFS and OS rates for the control arm were 74.3% and 89.7% respectively. The three-year DFS and OS rates for patients allocated to Herceptin for one year was 80.6% and 92.4% respectively. More details about the HERA trial are described in chapter seven.

5.1.2.1.4 NEAT trial

The National Epirubicin Adjuvant Trial (NEAT) and BR9601 trial were two joint phase III RCT that compares Epirubicin with cyclophosphamide, methotrexate and fluorouracil (CMF) against CMF alone (control arm) (Poole et al. 2006). These trials looked at the efficacy of anthracyclines in women with early breast cancer (all subtypes) that require adjuvant chemotherapy and could begin treatment within 10 weeks after having surgery. The TTE primary outcomes for this trial were OS and RFS. A combined sample size of 2000 patients aimed to detect a difference of 7% in OS and relapse-free survival (RFS) between the control arm and experimental arm with 85% power and 5% two-sided significance level. The NEAT and BR96001 trials recruited 2401 patients in approximately six years with a median follow-up of four years. The five-year OS rate was 75% with 171 deaths for the control arm and 82% for the Epirubicin plus CMF group. The five-year RFS rate was 69% and 76% for the control arm and the Epirubicin plus CMF group respectively.

5.1.2.2 Colon cancer trials

Four colon cancer trials were found that fulfilled the eligibility criteria stated in section 5.1.1. Details of these trials that were used in the calculations are provided below.

5.1.2.2.1 QUASAR 2 trial

The QUASAR 2 trial was a phase III RCT that investigated the use of bevacizumab in improving DFS for patients with colorectal cancer who were histologically proven to be high-risk stage II or stage III, had a primary resection four to ten weeks before randomisation and with a life expectancy of a minimum of five years (Kerr et al. 2016). The patients were assigned to receive either oral capecitabine alone (control arm) or oral capecitabine with bevacizumab. The TTE outcomes for this trial were three-year DFS for the primary outcome and three-year OS for the secondary outcome. A planned sample size of 2240 patients was required to find a 6% improvement in three-year DFS from the control arm group to the oral capecitabine with bevacizumab group with 90% power and 5% two-sided significance level. The QUASAR 2 trial recruited 1952 patients in approximately 5.5 years with a median follow-up of 4.92 years. The three-year DFS rate for control arm group was 78.4% and 75.4% for the oral capecitabine with bevacizumab group. The three-year OS rate for the control arm group was 89.4% with 169 deaths and 87.5% for the oral capecitabine with bevacizumab group.

5.1.2.2.2 COIN trial

The COIN trial aimed to assess the use of the EGFR targeted antibody cetuximab in combination with chemotherapy in patients with good organ function and with advanced colorectal cancer (stage IV) who had not received any prior chemotherapy (Maughan et al. 2011). The sample size and results used within this section for the calculations were based on patients with KRAS wild-type tumours as the primary outcome was OS in this group of patients. Patients were assigned to either oxaliplatin and fluoropyrimidine chemotherapy (control arm) or the control arm plus cetuximab. The other TTE outcome for this trial was PFS. This subgroup aimed to detect an improvement in two-year OS from 20% to 29.4% with 87% power and a 5% two-sided significance level. The COIN trial recruited 729 patients with KRAS wild-type tumours in approximately three years with a median follow-up of 21 months in the control arm and 23 months in the experimental arm. The median OS rates were 17.9 months with 257 deaths in the control arm and 17 months in the experimental arm and the median PFS rates were 8.6 months in both arms. During this trial, multiple interim

analyses were performed and reviewed by an independent DMC to advise on the safety of the trial treatment or recommend either continuation or closure of the trial.

5.1.2.2.3 FOCUS trial

The FOCUS trial investigated the use of different chemotherapy strategies to understand which regimens worked well to maximise the period of disease control with minimum adverse effects in patients with advanced or inoperable metastatic colorectal cancer (stage IV) (Seymour et al. 2007). In this phase III three-arm RCT, patients were allocated to either single-agent fluorouracil until failure then if patients were fit enough, they were given single-agent irinotecan (control group), fluorouracil until failure then either Irinotecan or oxaliplatin (arm B) or combination chemotherapy of either Irinotecan and fluorouracil or oxaliplatin and fluorouracil (arm C). The TTE primary outcome was OS and the secondary TTE outcome was PFS. A planned sample size of 2100 patients (700 in each arm) was required to detect a 7.5% improvement in two-year OS, from 15% in the control group to 22.5% in any pairwise comparisons of the control group versus either experimental arm. Hence, 1050 patients were required for any pairwise comparison with 80% power and 1% one-sided significance level. There were 2135 patients recruited into the trial over a period of 2.5 years with a median follow up of 26.5 months. The two-year OS rate was 22% for the control group with 617 deaths, 25% for arm B and 28% for arm C. The median PFS values for those given first-line treatment were 6.3 months for the control arm and arm B, 8.5 months for patients allocated to arm C that had Irinotecan and fluorouracil and 8.7 months for patients allocated to arm C that had oxaliplatin and fluorouracil.

5.1.2.2.4 PICCOLO trial

The PICCOLO trial aimed to combine panitumumab with irinotecan in patients that had advanced inoperable colorectal cancer (stage IV) which had progressed either during or after having chemotherapy containing fluoropyrimidine (Seymour et al. 2013). The TTE primary outcome was OS and the secondary TTE outcome was PFS. In this three-arm trial, patients were assigned to irinotecan (control arm), irinotecan plus ciclosporin or irinotecan plus panitumumab (IRPAN) however the sample size and results reported here are based on the comparison between the control arm and

the IRPAN arm for patients with KRAS wild-type tumours. A planned sample size of 466 patients was required to detect an anticipated improvement in median survival from 9 months in the control arm to 12.9 months in the IRPAN arm with 80% power and 5% two-sided significance level. The trial recruited 460 patients in two years with a median follow-up of 25.4 months. The median OS for the control arm was 10.9 months with 208 deaths and 10.4 months in the IRPAN arm. The median PFS for the control arm was 4.7 months and 5.5 months in the IRPAN arm. At the time of the interim analysis to test the superiority or inferiority of IRPAN against control using a p-value = 0.001, the IRPAN arm did not demonstrate superiority or inferiority hence continued to the final stage (Seymour et al. 2013).

5.1.2.3 Lung cancer trials

Four lung cancer trials were found that fulfilled the eligibility criteria stated in section 5.1.1. Details of these trials that were used in the calculations are provided below.

5.1.2.3.1 FRAGMATIC trial

The FRAGMATIC trial evaluated whether the use of LMWH improved survival for patients diagnosed with lung cancer (Macbeth et al. 2015). This two-arm phase III RCT assigned patients to receive LMWH or no LMWH (control arm). The TTE primary outcome was OS and the secondary TTE outcome was MFS. A planned sample size of 2200 patients was required to detect an improvement in one-year survival from 25% in the control arm to 30% in the experimental arm with 89% power and 5% two-sided significance level. The trial recruited 2202 patients in approximately four years with a median follow-up of 23.1 months. The one-year OS rates were 41.3% and 42.5% in the LMWH and control arms respectively, with 1020 deaths reported in the control arm. The one-year MFS rates were 16.2% and 14.9% in the LMWH and no LWMH arms respectively. During this trial, multiple interim analyses were performed and reviewed by an independent DMC to advise on safety of the trial treatment or recommend either continuation or closure of the trial.

5.1.2.3.2 FORTIS-M trial

The FORTIS-M trial investigated the use of talactoferrin which is an oral dendritic cell-mediated immunotherapy in patients with advanced (stage IIIb/IV) NSCLC

(Ramalingam et al. 2013). In this double-blind phase III RCT, patients were assigned to either receive talactoferrin or placebo (control arm). The TTE primary outcome was OS and secondary TTE outcome was PFS. A planned sample size of 720 patients was required to detect a 30% improvement in median OS from 4.6 months in the placebo arm (control arm) to six months in the talactoferrin arm with 85% power and 5% two-sided significance level. There were 742 patients recruited into the trial over a period of 2.5 years with a median follow-up of 18.1 months in the placebo arm and 19.6 in the talactoferrin arm. The median OS rate was 7.66 months in the placebo arm with 401 deaths and 7.49 in the talactoferrin arm. The median PFS rate was 1.64 months in the placebo arm and 1.68 months in the talactoferrin arm.

5.1.2.3.3 Big Lung trial

The Big Lung trial aimed to assess the benefits of adjuvant chemotherapy following a complete resection for patients with stage I-III NSCLC (Waller et al. 2004). For this phase III RCT, patients were assigned to either surgery alone (control) or surgery plus cisplatin-based chemotherapy. The TTE primary outcome was OS and the secondary TTE outcome was PFS. A planned sample size of 4000 patients was required to detect a 5% difference of OS at five years from 50% with surgery alone to 55% for surgery plus cisplatin-based chemotherapy. The trial team felt it was not possible to recruit such a large number of patients and hence accrued a sample of 500 patients, which could be added as an update to a meta-analysis. A total of 381 patients were recruited over a period of six years with a median follow-up of 34.6 months for the 183 patients that were alive, however the original sample size of 4000 patients was used as a comparator to deem whether or not the MAMS design was feasible. The two-year OS rate was 60% and 74% for the surgery and the surgery plus cisplatin group respectively. The two-year PFS rate was 51% and 53% for the surgery and surgery plus cisplatin group respectively. The event rate for this trial was not assessed as this trial would become part of a meta-analysis. During this trial, multiple interim analyses were performed and reviewed by an independent DMC to advise on safety of the trial treatment or recommend either continuation or closure of the trial.

5.1.2.3.4 TOPICAL trial

The TOPICAL trial evaluated the use of erlotinib in improving the clinical outcome for patients diagnosed with advanced (stage IIIb/IV) NSCLC (Lee et al. 2012). This double blind, superiority phase III RCT allocated patients to receive either erlotinib or a matching placebo (control). The TTE primary outcome was OS and the secondary TTE outcome was PFS. The main aim of this study was to detect an increase in one-year OS from 10% in the placebo arm to 17.5% in the control arm. To achieve this, a target sample size of 664 patients was required with 90% power and 5% two-sided significance level. The trial recruited 670 patients over a period of four years with an additional two years for follow-up. The median OS rate was 3.6 months in the placebo arm with 314 deaths and 3.7 months in the erlotinib arm. The median PFS rate was 2.6 months in the placebo arm and 2.8 months in the erlotinib arm.

5.1.2.4 Summary

The 12 trials used for the different cancer sites have been summarised in terms of the number of arms, sample size, recruitment time, follow-up time and event rates in Table 5.1. Based on these results, three arm two stage, four arm three stage and five arm four stage hypothetical MAMS designs were calculated in section 5.5.

Name	Arms	Subtype/ Stage	TTE Outcomes	Observed Rec. time (Years)	Observed Med. FU time (Years)	MCID (Target HR)	Target Accrual	Observed Accrual	Control arm deaths	IA performed†
Breast cancer										
Deva	2	Luminal B/ HER2 enriched	OS, DFS	8	5.5	5-year DFS 70% to 80% (0.63)	792	803	75	No
SoFEA	3	Luminal A/ Luminal B	OS, PFS	6	3	Median PFS 4 to 5.5 months (0.73)	750	723	173	Yes
HERA	3	Luminal B/ HER2 enriched	OS, DFS	3.5	2	5-year DFS 65% to 72% (0.77)	4482	5102	90	Yes
NEAT	2	All four subtypes	OS, RFS	6	4	5-year RFS 45% to 52% (0.82)	2000	2401	171	No
Colon cancer										
QUASAR 2	2	Stage II/III	OS, DFS	5.5	5	3-year DFS 66% to 72% (0.79)	2240	1952	169	No
COIN	2	Stage IV	OS, PFS	3	1.75	2-year OS 20% to 29% (0.76)	784	729	257	Yes
FOCUS	3	Stage IV	OS, PFS	2.5	2.2	2-year OS 15% to 22.5% (0.79)	2100	2135	617	No
PICCOLO	3	Stage IV	OS, PFS	2	2	Median OS 9 to 12 months (0.75)	466	460	208	Yes
Lung cancer										
FRAGMATIC	2	Stage I/II/III/IV	OS, MFS	4	2	1-year OS 25% to 30% (0.87)	2200	2202	1020	Yes
FORTIS	2	Stage III/IV	OS, PFS	2.5	1.5	Median OS 4.6 to 6 months (0.77)	720	742	401	No
Big Lung	2	Stage I/II/III	OS, PFS	6	3	5- year OS 50% to 55% (0.86)	500	381	99	Yes
TOPICAL	2	Stage III/IV	OS, PFS	4	2	1-year OS 10% to 17.5% (0.76)	664	670	314	No

†Further details of the interim analyse have been presented in section 5.5.

Table 5.1: Details obtained from the different phase III cancer trials for breast cancer, colon cancer and lung cancer. TTE – Time to event; FU- Follow-up; SS – Sample size; IA – Interim analysis; OS – Overall Survival, DFS – Disease free survival, PFS – Progression free survival, RFS – Relapse free survival, MFS – Metastasis-free survival; MCID = Minimally clinical important difference

5.2 Methods for the first and second phase calculations

The hypothetical performance of a three arm two stage, four arm three stage trial using a superiority hypothesis were assessed for each of the three cancer sites. The first and second phase calculations based the recruitment, follow-up time and treatment difference on the results obtained from the selected 12 trials (Table 5.1). The total trial length (recruitment time plus follow-up time) would vary between 5.5 to 13.5 years for breast cancer, four to 10.5 years for colon cancer and four to nine years for lung cancer. Additional time for recruitment was given as the number of trial arms increased, i.e. a recruitment period of seven years for a three arm breast cancer trial and a recruitment period of eight years for a four arm breast cancer trial. The survival rates used for the first and second phase calculations were based on the rates obtained from the ONS in which one-year and five-year estimated survival rates for various types of cancers between the years 2013 to 2017 in England were stated (Office for National Statistics 2019). These figures represent the estimates for the general population. Although, it is unlikely that a clinical trial would be performed for the general population, this was implemented as a baseline for future trials implementing MAMS designs. The one-year and five-year estimated survival rates for breast, colon and lung cancer are displayed in Table 5.2.

Type of cancer	**Survival rate**	
	1-year	**5-year**
Breast (only Women)	95.8%	85.0%
Colon	76.0%	57.5%
Lung	40.8%	16.4%

Table 5.2: One-year and five-year survival rates for breast, colon and lung cancer from England between 2013-17 (Office for National Statistics 2019).

5.2.1 Scenarios

To better understand and appreciate the MAMS designs, each cancer site had up to three different scenarios that were considered based on the number of arms at the beginning of the trial (Table 5.3):

1) At every interim stage, the pairwise comparisons would show that all experimental arms were statistically significant from the control arm hence all experimental arms would continue to the final stage;
2) At every interim stage, pairwise comparisons would show all but one experimental arm to be statistically significant from the control arm using a superiority hypothesis hence progressively removing one experimental arm at each interim stage;
3) For the four arm three stage MAMS design, at the first interim stage the pairwise comparisons would show that one experimental arm would be statistically significant at the pre-specified level when compared to the control arm, i.e. the trial would begin with three experimental arms, and would be reduced to only one experimental arm after the first interim analysis.

Scenario	Number of arms at each stage (inc. control arm)
3 arm 2 stage	
1	3, 3
2	3, 2
4 arm 3 stage	
1	4, 4, 4
2	4, 3 2
3	4, 2, 2

Table 5.3: Number of arms at each stage (including control arm) for the different scenarios for a three arm two stage and four arm three stage MAMS design.

5.2.2 Assumptions

The following assumptions were used for the hypothetical calculations:

- The allocation ratio would be equal for all arms;
- Patients would be accrued at a uniform rate over the recruitment period;

- The sample size for each MAMS design would be fixed hence allowing for the total length of the trial to vary based on the different scenarios. This would be to ensure maximum information was obtained and not to minimise the sample size as in the VIETNARM trial (see appendix 1);
- The hypotheses would be all distinct therefore focus was on controlling the PWER as in the STAMPEDE trial;
- The first stage significance value would be 50% and all stagewise significance values thereafter descend in value.

5.2.3 Guidelines to determine a feasible MAMS design

The proportion of patients in the phase III trials for each cancer site were used to calculate an estimate for the maximum number of patients to enter phase III trials based on the estimated number of patients entering clinical trials each year. This estimation was based upon approximately 367,000 patients annual diagnosed with cancer in the UK between 2015-2017, of which approximately 55,000, 42,000 and 48,000 patients were diagnosed with breast, colon and lung cancer respectively (Cancer Research UK 2018). It is estimated that one in six patients diagnosed with cancer in the UK joins a clinical trial, which is approximately 61,000 patients (Ajithkumar et al. 2017). Therefore, by taking one in six patients for each of the selected cancer sites approximately 9170 patients enter a breast cancer trial, 7000 patients enter a colon cancer trial and 8000 patients enter a lung cancer trial. Furthermore, searching the ClinicalTrials.gov website (ClinicalTrials.gov 2020) identified a total of 182 clinical trials that are recruiting within the UK. The title, phase and the planned number of patients to be enrolled onto the study for all 182 clinical trials can be found in appendix 6. The number of trials and the mean planned sample size within each phase split by cancer site were recorded (Table 5.4).

Trial phase	Cancer site					
	Breast		Colon		Lung	
	No. of trials	Target Mean SS	No. of trials	Target Mean SS	No. of trials	Target Mean SS
Phase I	17 (23.6%)	203 (6.8%)	10 (33.3%)	204 (4.8%)	16 (20%)	267 (11.7%)
Phase I/II	16 (22.2%)	402 (13.5%)	8 (26.7%)	501 (11.9%)	22 (27.5%)	337 (14.7%)
Phase II	17 (23.6%)	287 (9.6%)	8 (26.7%)	305 (7.2%)	20 (25%)	275 (12%)
Phase II/III	4 (5.6%)	578 (19.4%)	0 (0%)	0 (0%)	2 (2.5%)	105 (4.6%)
Phase III	18 (25%)	1511 (50.7%)	4 (13.3%)	3199 (76%)	19 (23.8%)	705 (30.8%)
Phase IV	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1.3%)	600 (26.2%)
Total	72	2981	30	4209	80	2289

Table 5.4: Number of on-going clinical trials and the target mean sample size for each phase for breast, colon and lung cancer. SS – Sample size.

The maximum number of patients entering phase III trials for each cancer site is obtained using the proportion of patients entering phase III trials given in Table 5.4 and the estimated target number of patients that enter trials for the given cancer site which can be seen in Table 5.5. For example, from Table 5.4, the proportion of patients entering phase III breast cancer trials is 50.7%. Hence, using this information it is estimated that a maximum of 4649 patients (387 patients per month) are available to enter phase III breast cancer trials per year (calculated by taking 50.7% of 9170).

Cancer Site	Incidence for 2017	Patients entering clinical trials per year	Estimated max. number of Patients entering phase III trials per month
Breast	55000	9170	387
Colon	42000	7000	443
Lung	48000	8000	205

Table 5.5: The maximum number of patients that could be recruited into a trial per month for breast, colon and lung cancer.

For a MAMS design within the first and second phase calculations to be classed as feasible, the following must hold true:

- The number of patients recruited per month would be no more than the values stated in Table 5.5 for each of the cancer sites;
- A minimum duration of six months would be required between each stage as the time for operational tasks with this time period such as data entry, validating the data etc. needs to be accounted for;

- An achievable time given to recruit and follow-up patients which was based upon the 12 clinical trials mentioned in section 5.1.1 .

5.3 First phase of calculations

5.3.1 Methods

The first phase of the MAMS designs evaluated three cancer sites using an intermediate and primary outcome of five-year survival rate for breast and colon cancer and one-year survival rate for lung cancer (Table 5.2). The high mortality rate of patients diagnosed with lung cancer resulted in applying a one-year survival rate.

The HR within a superiority setting (calculation of these values are shown in chapter one), the number of years for recruitment for each hypothetical MAMS design and the follow-up period after recruitment for each of the cancer sites are displayed in Table 5.6. These parameters were selected based upon the trials shown in Table 5.1 and used for the first phase of calculations.

The time to perform the analysis for each stage is dependent upon the number of events reached in the control arm, hence the number of control arm events at each interim and final stage, the duration of each stage and the total sample size were recorded. Furthermore, the aim of this phase was to select an appropriate significance level and power at each stage such that the overall PWER was 5% and a pairwise power value of at least 80%.

Type of cancer	Rate (MCID Abs. difference)	HR	Total number of arms	Recruitment period (Years)	Approximate Follow-up (Years)
Breast	85.0% (7.5%)	0.48	3	7	4
			4	8	4
Colon	57.5% (7.5%)	0.78	3	4	3
			4	5	3
Lung	40.8 (5.0%)	0.87	3	4	2
			4	5	2

Table 5.6: HR calculated using five-year survival rates for breast and colon cancer with 7.5% absolute treatment difference. HR calculated using one-year survival rate for lung cancer with 5% absolute treatment difference. Parameters selected based upon trials shown in Table 5.1 and used for the first phase of calculations. HR – Hazard Ratio; MCID – Minimally clinical important difference.

5.3.2 Example – First phase calculations

Below is an example of a hypothetical four arm three stage MAMS design with a superiority hypothesis for patients diagnosed with breast cancer. Five-year OS survival rate of 85% will be used for the intermediate and primary outcome. A HR of 0.48 would be interpreted as an absolute improvement of 7.5% in survival from 85% to 92.5% at five years for any experimental arm compared to the control. For a four arm three stages MAMS design to have a pairwise power of approximately 80%, the MAMS design would require 95% power at stage one, 88% power at stage two and 86% power at the final stage for each pairwise comparison. To have a PWER of 5%, the MAMS design would require a 50% alpha value at stage one, 20% at stage two and 5% at the final stage. Approximately 640 patients would be recruited in eight years (accrued at a uniform rate of 7 patients per month) and allocated equally to all treatment arms, so that patients would be followed-up for approximately three years. This MAMS design was applied to the three different scenarios stated earlier. The parameters inputted into the 'nstage' command in Stata can be found in Table 5.7 and the syntax and results for scenario one is shown in Figure 5.1.

Parameter	Value
Total number of stages	3
Number of arms in each stage	4, 4, 4
Total accrual	640
Allocation ratio	1
Survival Probability for OS	0.85
Survival Time (years)	5
HR under H0	1
HR under H1	0.48
One-sided alpha for each stage	0.5, 0.2, 0.05
Power for each stage	0.95, 0.88, 0.86
Time units	1 (= one year)
Time of recruitment stopping (years)	8

Table 5.7: Parameters inputted for a hypothetical four arm three stage MAMS trial using the 'nstage' command in Stata, the intermediate and primary outcome was OS. OS – Overall survival, HR – Hazard ratio, H0 – Null hypothesis, H1 – Alternate hypothesis.

Applying the different scenarios resulted in variations in the lengths of follow-up and the number of cumulative experimental arm events across stages (Table 5.8) because the sample size and accrual period remained fixed, but the number of experimental arms varied for each scenario. For scenario one, the rate for which patients would be accrued in the control arm would be lower and therefore would take longer to reach the required number of control arm events to trigger the final analysis in comparison to the rate in scenarios two and three and therefore required more follow-up time.

Scenario	FU time	Experimental events at each stage	Total Trial length (years)
1	3.2	24, 30, 51	11.2
2	3	24, 20, 17	11
3	2.75	24, 10, 17	10.75

Table 5.8: The difference in FU time and cumulative experimental events required at each stage for the different scenarios. FU – Follow-up

From the results, the three scenarios for the MAMS design applied in breast cancer, 7 patients would be recruited per month satisfying the first criteria whereby a maximum of 387 patients could be recruited per month. The duration of each stage for all scenarios would be more than six months (Table 5.9) and recruitment and follow-up for patients would be achieved within a reasonable time when compared to the breast cancer trials (between 5.5 to 13.5 years for breast cancer stated in section 5.1.2) in Table 5.1, hence deeming these MAMS designs as feasible.

```
. nstage, accrue(80 80 80) arms (4 4 4) alpha(0.5 0.2 0.05) hr0(1 1) hr1(0.48 0.48)
>  omega(0.95 0.88 0.86) t(5 5) s(0.85 0.85) nstage(3) tstop(8)

n-stage trial design                        version 4.0.1, 2 Nov 2018
─────────────────────────────────────────────────────────────────────

Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
─────────────────────────────────────────────────────────────────────

Note: I outcome and D outcome are identical
Median survival time: 21.3 time units

Operating characteristics
──────────────────────────────────────────────────────────────────────────────
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
──────────────────────────────────────────────────────────────────────────────
1           0.5000    0.953    1.000    0.480    1.000    7.053    7.053
2           0.2000    0.881    1.000    0.480    0.761    0.923    7.976
3           0.0500    0.865    1.000    0.480    0.667    3.217   11.193
──────────────────────────────────────────────────────────────────────────────
Pairwise Error Rate          0.0421              Pairwise Power   0.8117
Familywise Error Rate (SE)   0.1008 (0.0006)
──────────────────────────────────────────────────────────────────────────────
Note: patient accrual stopped at time  8.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ──────────Stage 1──────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate       80       20       60
Patients*      564      141      423
Events**        39       15       24
           ──────────Stage 2──────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate       80       20       60
Patients*      639      160      479
Events**        49       19       30
           ──────────Stage 3──────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate       80       20       60
Patients*      640      160      480
Events**        84       33       51
           ──────────────────────────
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE
```

Figure 5.1: Syntax and output obtained for the first phase 4A3S MAMS design with a superiority hypothesis in a breast cancer setting for scenario one.

5.3.3 Results

A summary of the results from the first phase of the calculations is provided in Table 5.9. For the MAMS designs, it was evident that trials in lung cancer would require the most patients, followed by colon cancer then breast cancer. In comparison to the other cancer sites, the MAMS designs for lung cancer had the smallest treatment difference and had the smallest overall trial length which would result in requiring the most patients. Therefore, more events would be required and hence more patients to determine this difference. Thus, an improved survival rate indicated that fewer events would occur which meant that fewer events would be required to test the null hypothesis for futility or lack of benefit at each interim stage and to conclude superiority at the final stage.

There would be no difference to the overall length of the trial when comparing scenarios one and two for the three arm two stage breast cancer trial as the first interim stage would occur at seven years for both scenarios at which point all patients would be randomised. The only difference would be in the number of experimental arm events at the final stage as for scenario one both experimental arms would remain, whereas in scenario two, one experimental arm would be dropped and not analysed at the final stage. Follow-up would cease for those patients on the experimental arm which would be discontinued. The initial aim of the three arm two stage breast cancer trial was to have seven years for recruitment and four years for follow-up however these hypothetical trials would only need a maximum of 3.5 years follow-up to reach the required number of events with the estimated number of patients. A decrease in the number of patients to extend the follow-up period to four years would have resulted in the recruitment period completing before stage one. This design would have then been considered 'infeasible' as per the key principle of the MAMS design framework that the recruitment completes in between the interim stage before the final stage and the final stage (Blenkinsop et al. 2019). The results of these two scenarios satisfied all the criteria mentioned earlier whereby 5 patients would be recruited each month, the duration of each stage would be all greater than six months and recruitment and

follow-up would be completed between nine to 13.5 years - an achievable time, hence deeming all three arm two stage breast cancer trials as feasible.

For the four arm three stage breast cancer trial (used as an example in section 5.3.2), the second scenario would have a reduced follow-up period compared to the first scenario as an experimental arm would be discontinued whilst recruitment was on-going as the first interim stage occurred at seven years (recruitment period up until eight years). This would lead to an increase in the rate at which newly recruited patients would be allocated in the remaining arms which meant events occurred quicker and therefore reduced the duration of the follow-up period. For the third case, where it was hypothesised that an experimental arm was discontinued at each stage, the overall trial length would be further reduced. Hence, for this four arm three stage breast cancer trial, the overall trial varied from 10.75 years to 11.2 years based on the outcome of the pairwise comparisons at each stage. As mentioned in the previous section, the four arm three stage MAMS designs would be all deemed feasible.

The two scenarios of dropping arms for three arm two stage colon cancer trial would be like the three arm two stage breast cancer trial such that there would be no difference in the overall length of the trial as the first interim stage would occur at approximately before four years for both scenarios (four years for recruitment). The duration of each interim stage would be greater than six months and the recruitment period and follow-up period would be achievable (total trial length between four to 10.5 years for colon years to be deemed achievable) and the number of patients required would be less than the estimated maximum number of patients recruited per month (32 patients per month vs 443 estimated maximum patients per month).

The four arm three stage colon cancer trial like the four arm three stage breast cancer trial, would reduce the total length for each progressive case. The overall trial would vary from 7.5 years to 6.9 years based on the outcome of the pairwise comparisons at each stage. It was initially aimed to have a total trial length of eight years (five years recruitment and three years follow-up) however this would not be permitted by the ‘nstage’ command as any increase in patients would result in recruitment ceasing during an interim stage prior to the final stage. The four arm three stage trial

would require 40 patients per month to be recruited compared to the estimated maximum number of patients that could be recruited per month (443 patients) deeming this trial feasible.

The results of the two scenarios for the three arm two stage lung cancer trial would be like that of the breast cancer and colon cancer trials in terms of fewer experimental arm events required for the second case. However, the follow-up period would be reduced from two years in the first case to 1.6 years in the second case which would be because the interim analysis occurred at 2.6 years (whilst recruitment would be ongoing) unlike breast cancer and colon cancer where the first interim stage would take place at the end of the recruitment period. The three arm two stage lung cancer trial would require 45 patients per month which is less than the 205 estimated maximum patients that could be recruited per month deeming this trial as feasible.

The results of the three scenarios for the four arm three stage lung cancer trial would be also like that of breast cancer and colon cancer in that for each progressive case the follow-up time and the number of experimental arms required would decrease. However, compared to the other cancer sites, the durations of the interim stages would be less than a year (ranging from 0.7 years to 0.9 years). The total trial length ranged from five years for the third scenario to seven years for the first scenario, hence the third scenario would perform the final stage analysis at the end of recruitment which would mean there would be several patients that would have reduced or no follow-up. The four arm three stage MAMS designs for lung cancer would require 55 patients per month compared to 205 patients that could be recruited per month deeming this trial as feasible.

Scenario	Significance value		Power value		Total Patients (pts per month)	Control arm events		Experimental arm events	Accrual stop (Years)	Stage duration (Years)	Trial length (Years)
	Stagewise alpha	Overall alpha	Stagewise power	Overall power		Intermediate events	Final stage				
3 arm 2 stage - Breast											
1	0.5, 0.05	0.05	0.95, 0.82	0.81	427 (5)	15	28	16, 30	7	7, 3.5	10.5
2	0.5, 0.05	0.05	0.95, 0.82	0.81	427 (5)	15	28	16, 15	7	7, 3.5	10.5
4 arm 3 stage - Breast											
1	0.5, 0.2, 0.05	0.04	0.95, 0.88, 0.86	0.81	640 (7)	15, 19	33	24, 30, 51	8	7, 1, 3.2	11.2
2	0.5, 0.2, 0.05	0.04	0.95, 0.88, 0.86	0.81	640 (7)	15, 19	33	24, 20, 17	8	7, 1, 3	11
3	0.5, 0.2, 0.05	0.04	0.95, 0.88, 0.86	0.81	640 (7)	15, 19	33	24, 10, 17	8	7, 1, 2.75	10.75
3 arm 2 stage - Colon											
1	0.5, 0.05	0.05	0.95, 0.82	0.81	1540 (32)	97	216	156, 358	4	4, 3	7
2	0.5, 0.05	0.05	0.95, 0.82	0.81	1540 (32)	97	216	156, 179	4	4, 3	7
4 arm 3 stage - Colon											
1	0.5, 0.2, 0.05	0.041	0.95, 0.88, 0.86	0.8	2380 (40)	96, 138	246	234, 336, 609	5	4, 1, 2.5	7.5
2	0.5, 0.2, 0.05	0.041	0.95, 0.88, 0.86	0.8	2380 (40)	96, 138	246	234, 224, 203	5	4,1, 2.2	7.2
3	0.5, 0.2, 0.05	0.041	0.95, 0.88, 0.86	0.8	2380 (40)	96, 138	246	234, 112, 202	5	4, 1, 1.9	6.9
3 arm 2 stage - Lung											
1	0.5, 0.05	0.05	0.95, 0.82	0.81	2164 (45)	294	690	550, 1354	4	2.6, 3.4	6
2	0.5, 0.05	0.05	0.95, 0.82	0.81	2164 (45)	294	690	550, 664	4	2.6, 2	4.6
4 arm 3 stage - Lung											
1	0.5, 0.2, 0.05	0.041	0.95, 0.88, 0.86	0.8	3245 (54)	293, 433	782	825, 1236, 2310	5	2.8, 0.9, 3.3	7
2	0.5, 0.2, 0.05	0.041	0.95, 0.88, 0.86	0.8	3245 (54)	293, 433	782	825, 822, 751	5	2.8, 0.8, 1.5	5.1
3	0.5, 0.2, 0.05	0.041	0.95, 0.88, 0.86	0.8	3245 (54)	293, 433	782	825, 410, 750	5	2.8, 0.7, 1.5	5

Table 5.9: First phase of calculations. Sample size outcomes for a four arm three stage and three arm two stage hypothetical MAMS designs for Breast, Colon and Lung cancer with a superiority hypothesis.

5.4 Second phase calculations

5.4.1 Methods

The second phase calculations for the hypothetical MAMS trials evaluated the outcomes when the stagewise significance and stagewise power values as suggested by Royston et al (2011) were implemented. The authors within this paper suggest using for the significance level at interim stage i the formula $\alpha_i = 0.5^i$ $(i < f)$ and $\alpha_f = 0.025$ where f is the final stage. Furthermore, the power for each interim stage should be at least 95% and a power of at least 90% for the final stage. This method reduces the overall alpha value and increases the power, which in turn increases the robustness of these trials, additionally these values were implemented in the earlier comparisons of the STAMPEDE trial (see Table 4.2).

Based on the results of the first phase calculations, the recruitment period for the three arm two stage trials for the colon and lung cancer trials were extended from four years to six years and from five years to seven years in the four arm three stage trials to decrease the recruitment rate per month.

The results from this phase were compared to the results from the first phase calculations in terms of the differences between the overall significance values, the overall power values, the sample size, the number of events required for the control and experimental arms and the duration of the stages.

5.4.2 Example – Second phase calculations

The example shown in section 5.3.2 is continued here except the difference is that this MAMS design would provide 95% power at stages one and two and 90% power at the final stage for each pairwise comparison. A type I error rate of 50% at stage one, 25% at stage two and 2.5% at stage would be used as the significance level for each pairwise comparison. Assuming 872 patients are recruited in eight years (accrued at a uniform rate of 9 patients per month) and equally allocated, it is anticipated that patients will be followed-up for approximately 3.4 years. This hypothetical MAMS design was applied to the three different scenarios stated earlier. The parameters inputted into the 'nstage' command in Stata can be found in Table 5.10 and the syntax and results for scenario one are shown in Figure 5.2.

Parameter	Value
Total number of stages	3
Number of arms in each stage	4, 4, 4
Total accrual	872
Allocation ratio	1
Survival Probability for OS	0.85
Survival Time (years)	5
HR under H0	1
HR under H1	0.48
One-sided alpha for each stage	0.5, 0.25, 0.025
Power for each stage	0.95, 0.95, 0.9
Time units	1 (= one year)
Time of recruitment stopping (years)	8

Table 5.10: Parameters inputted for a hypothetical four arm three stage MAMS trial using the 'nstage' command in Stata, the intermediate and primary outcome was OS. OS – Overall survival, HR – Hazard ratio, H0 – Null hypothesis, H1 – Alternate hypothesis.

Applying the different scenarios for the second phase resulted in different lengths of follow-up and cumulative experimental arm events across stages compared to those seen in the first phase (Table 5.11).

Scenario	FU time	Experimental events at each stage
1	3.4	24, 42, 72
2	3	24, 28, 24
3	2.4	24, 14, 24

Table 5.11: The difference in FU time and cumulative experimental events require at each stage for the different scenarios for the second phase of calculations. FU – Follow-up.

From the second phase calculations, the three scenarios for the MAMS design applied in breast cancer, 9 patients would be recruited per month satisfying the first criteria whereby a maximum of 387 patients could be recruited per month. The duration of each stage for all scenarios were more than six months (Table 5.13) and recruitment and follow-up for patients would be achieved within a reasonable time when compared to the breast cancer trials in Table 5.1 hence deeming these MAMS designs as feasible.

```
. nstage, accrue(109 109 109) arms (4 4 4) alpha(0.5 0.25 0.025) hr0(1 1) hr1(0.48
> 0.48) omega(0.95 0.95 0.9) t(5 5) s(0.85 0.85) nstage(3) tstop(8)

n-stage trial design                        version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 21.3 time units

Operating characteristics
------------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
------------------------------------------------------------------------------
1          0.5000    0.953    1.000    0.480    1.000    6.009    6.009
2          0.2500    0.951    1.000    0.480    0.829    1.985    7.994
3          0.0250    0.902    1.000    0.480    0.665    3.384   11.378
------------------------------------------------------------------------------
Pairwise Error Rate          0.0226             Pairwise Power   0.8651
Familywise Error Rate (SE)   0.0577 (0.0005)
------------------------------------------------------------------------------
Note: patient accrual stopped at time  8.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ----------Stage 1----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      109       27       82
Patients*      655      164      491
Events**        39       15       24
           ----------Stage 2----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      109       27       82
Patients*      871      218      653
Events**        68       26       42
           ----------Stage 3----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      109       27       82
Patients*      872      218      654
Events**       118       46       72
           ------------------------------

 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE
```

Figure 5.2: Syntax and output obtained for the second phase four arm three stage MAMS design with a superiority hypothesis in a breast cancer setting for scenario one.

5.4.3 Results

A summary of the results from the second phase calculations are provided in Table 5.13. For the MAMS designs, like the first phase calculations, lung cancer would require the most patients, followed by colon cancer then breast cancer and for each hypothesized MAMS trial, the follow-up time would decrease for each progressive case. Furthermore, based on guidelines stated in section 5.2.3, the breast cancer and the colon cancer trials would be deemed as feasible MAMS designs that could be implemented using survival as the intermediate and primary outcome.

The number of control arm events and patients that would be required in the first phase and second of calculations can be seen in Table 5.12. A change in the stagewise significance values and stagewise power values would increase the events and patients required. Extending the recruitment period would reduce the number of patients required to be recruited into the trials per month for the colon cancer trials and the four arm three stage lung cancer trial but at the same time the change in the stagewise significance and power values would inflate the sample size.

	Control Events		**Sample Size**	
	First Phase	**Second Phase**	**First Phase**	**Second Phase**
Breast 3A2S	28	46	427	539
Breast 4A3S	33	46	640	872
Colon 3A2S	216	347	1540	2172
Colon 4A3S	246	347	2380	3115
Lung 3A2S	690	1111	2164	3420
Lung 4A3S	782	1111	3245	4536

Table 5.12: Number of control events and patients required in the first phase and second phase of calculations. 3A2S – Three arm two stage; 4A3S – Four arm three stage.

The use of the stagewise significance and power values as stated by Royston et al (2011) would give an overall lower alpha value and higher power value for each trial demonstrating the robustness within this phase compared to the first phase. Additionally, a high power at each interim stage would ensure that an experimental arm is not incorrectly discontinued from the trial as the futility boundary at the interim stages can be significantly large.

Case	Significance value		Power value		Total Patients (pts per month)	Control arm events		Experimental arm events	Accrual stop (Years)	Stage duration (Years)	Trial length (Years)
	Stagewise alpha	Overall alpha	Stagewise power	Overall Power		Intermediate events	Final stage				
3 arm 2 stage - Breast											
1	0.5, 0.025	0.02	0.95, 0.9	0.88	539 (6)	15	46	16, 48	7	6.2, 6.5	12.7
2	0.5, 0.025	0.02	0.95, 0.9	0.88	539 (6)	15	46	16, 24	7	6.2, 6	12
4 arm 3 stage - Breast											
1	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.87	872 (9)	15, 26	46	24, 42, 72	8	6, 2, 3.4	11.4
2	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.87	872 (9)	15, 26	46	24, 28, 24	8	6, 2, 3	11
3	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.87	872 (9)	15, 26	46	24, 14, 24	8	6, 1.9, 2.5	10.4
3 arm 2 stage - Colon											
1	0.5, 0.025	0.02	0.95, 0.9	0.87	2172 (30)	96	347	156, 580	6	4, 5	9
2	0.5, 0.025	0.02	0.95, 0.9	0.87	2172 (30)	96	347	156, 288	6	4.1, 4.2	8.3
4 arm 3 stage - Colon											
1	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.86	3115 (37)	96, 185	347	234, 453, 867	7	4.2, 1.8, 3	9
2	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.86	3115 (37)	96, 185	347	234, 302, 287	7	4.2, 1.7, 2.3	8.3
3	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.86	3115 (37)	96, 185	347	234, 150, 286	7	4.2, 1.6, 2	7.8
3 arm 2 stage - Lung											
1	0.5, 0.025	0.02	0.95, 0.9	0.87	3420 (48)	294	1105	550, 2180	6	2.5, 5.5	8
2	0.5, 0.025	0.02	0.95, 0.9	0.87	3420 (48)	294	1111	550, 1070	5.8	2.5, 3.3	5.8
4 arm 3 stage - Lung											
1	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.86	4536 (54)	293, 573	1105	825, 1647, 3276	7	2.8, 1.8, 4.4	9
2	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.86	4536 (54)	293, 573	1111	825, 1096, 1069	6.4	2.8, 1.6, 2	6.4
3	0.5, 0.25, 0.025	0.02	0.95, 0.95, 0.9	0.86	4536 (54)	293, 573	1111	825, 1647, 3276	5.8	2.8, 1.2, 1.8	5.8

Table 5.13: Second phase of calculations. Sample size outcomes for a four arm three stage and three arm two stage hypothetical MAMS designs for Breast, Colon and Lung cancer with a superiority hypothesis.

5.5 Methods for third phase calculations

The third phase of MAMS design calculations used the parameters obtained from the 12 phase III RCTs as summarised in Table 5.1 to:

1. Re-calculate sample size for a standard two-arm trial using OS as a primary outcome;
2. Create fictional scenarios to calculate sample sizes for MAMS designs where multiple treatments were available to compare against a common control using the 'nstage' command;
3. Use the guidelines (see section 5.5.1.1) to compare the efficiency of standard two-arm trials compared to MAMS designs.

Three arm two stage, four arm three stage and five arm four stage hypothetical MAMS designs were calculated.

5.5.1 Methods

The focus of the third phase was to assess the application of MAMS designs with longer term outcomes such as OS. These will be applied in a fictitious setting whereby multiple treatments would be available to evaluate against a common control by comparing the effectiveness of MAMS designs against the running of standard two-arm trials using the aforementioned phase III trials (Table 5.1). To ensure consistency and comparability against the MAMS designs, all trials had sample size re-calculated as a standard two-arm trial with the same assumptions as used in the MAMS design and were renamed with the suffix "_OS" to emphasise that OS had been used as the primary outcome. To recalculate these trials, the OS result obtained for the control arm, the planned treatment difference for the primary outcome in the original trial, power of 90% and one-sided significance of 2.5% were all used. The new sample size and number of events can be seen in Table 5.14.

Trial Name	OS (Target HR)	SS (Control SS)	Control events

Breast cancer trials			
Deva_OS	5 year 81.8% (0.43)	278 (139)	33
SoFEA_OS	Median 0S 21.6 mo. (0.73)	482 (241)	218
HERA_OS	3 year 89.7% (0.31)	448 (224)	19
NEAT_OS	5 year 75% (0.69)	1094 (547)	156
Colon cancer trials			
QUASAR 2_OS	3 year 89.4% (0.42)	344 (172)	32
COIN_OS	Median 17.9 mo. (0.76)	766 (383)	285
FOCUS_OS	2 year 22% (0.81)	950 (475)	368
PICCOLO_OS	Median 10.9 mo. (0.75)	718 (359)	259
Lung cancer trials			
FRAGMATIC_OS	1 year 42.5% (0.87)	2320 (1160)	1090
FORTIS-M_OS	Median 7.6 mo. (0.77)	908 (454)	313
Big Lung_OS	2 years 60% (0.84)	1980 (990)	728
TOPICAL_OS	Median 3.6 mo. (0.80)	842 (421)	420

Table 5.14: Re-calculated sample sizes using OS results. SS = Sample size.

The stagewise significance level and power values as suggested by Royston et al (2011) were used for all MAMS design in this phase of calculations. The primary and intermediate outcome for all trials were OS. It was assumed that for these MAMS designs all experimental arms would continue to the final stage hence maximising the number of patients and time required, therefore capitalising on any funding given to support the trial. This is the worst case scenario from a funding and time perspective however it is the best case scenario for hope for future patients as it means that there is a higher possibility that more treatments are worthwhile.

The survival probability and survival time used for each of the MAMS trials can be seen in Figure 5.3. The recruitment time, estimated follow-up time and the absolute treatment difference for each trial that was used in the calculations for the MAMS designs can be found in Table 5.15.

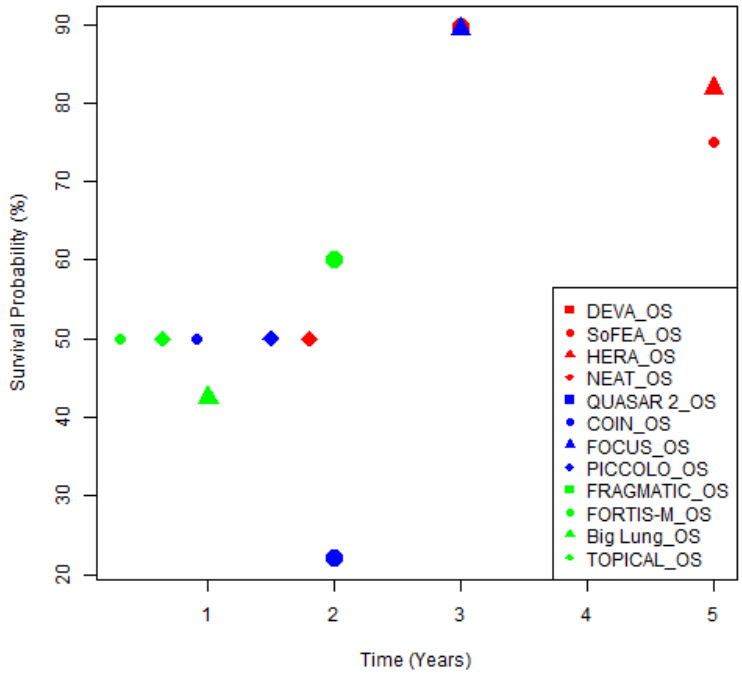

Figure 5.3: Scatter plot showing the survival probability and the time for each of the trials which was used to calculate sample sizes for the MAMS trials.

Trial Name	Rec. Time (Years)	Med. FU (Years)	Absolute Trt. Difference
Breast Cancer Trials			
Deva_OS	8	5.5	10%
SoFEA_OS	6	3	29.6 months
HERA_OS	3.5	2	7%
NEAT_OS	6	4	7%
Colon Cancer Trials			
QUASAR 2_OS	5.5	5	6%
COIN_OS	3	1.75	23.5 months
FOCUS_OS	2.5	2.2	7.5%
PICCOLO_OS	2	2	14.5 months
Lung Cancer Trials			
FRAGMATIC_OS	4	2	5%
FORTIS-M_OS	2.5	1.5	10 months
Big Lung_OS	6	3	5%
TOPICAL_OS	4	2	7.5%

Table 5.15: Summary of the parameters that were used to apply MAMS designs for each trial. FU = Follow-up; HR = Hazard Ratio; OS = Overall Survival; Trt = Treatment.

.

5.5.1.1 Guidelines to determine a feasible MAMS design for the third phase

Incidence rates of the subgroups for each cancer site were used to estimate the potential maximum number of patients that could be recruited per month for each of the 12 trials. Breast cancer subtypes are defined by three tumour markers; ER, PR which are both classed as hormone receptors (HoR) and HER2. The subtypes of breast cancer (formed by the joining of these markers) and the incidence rates for each of these subtypes have been classified in Figure 5.4 (Howlader et al. 2014, Komen 2017). The incidence rates of the different stages of colon and lung cancer in the UK between 2010-2014 (Cancer Research UK 2018) have been classified in Figure 5.5.

The incidence values for the different sub-types/stages for each cancer site and the maximum number of patients entering phase III cancer trials from Table 5.5 were used to calculate the maximum number of patients that could be recruited per month (Table 5.16). These values were used as one of the comparators to deem whether the results for the re-analysed trials from the third phase of calculations produced a feasible MAMS design.

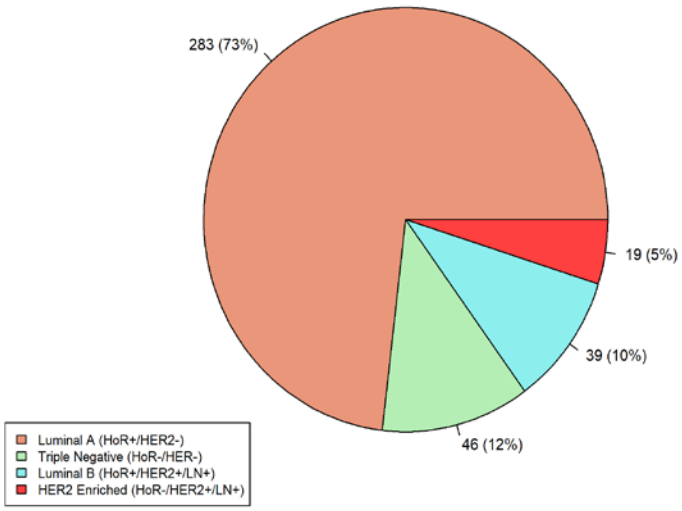

Figure 5.4: The estimated maximum number of patients to enter phase III trials per month spit by sub-type for Breast Cancer. HoR = Hormone Receptors; HER2 – Human Epidermal Growth Factor Receptor 2; LN = Lymph node.

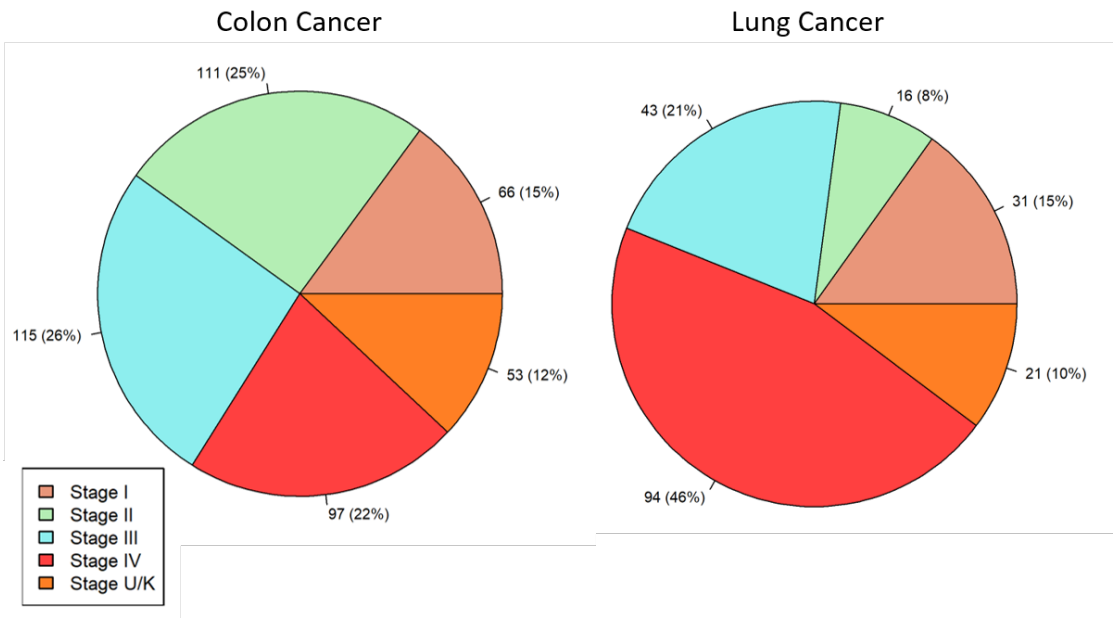

Figure 5.5: The estimated maximum number of patients to enter phase III trials per month spit by stage for Colon and Lung cancer. U/K – Unknown.

The hypothetical MAMS design for the trials were classed as feasible if all three of the following proved to be successful:

1. The MAMS design would require fewer patients per month when compared to the estimated maximum number of patients entering phase III trials per month (Table 5.16);
2. The number of control arm patients required in the MAMS designs would be no more than 10% greater than the two-arm trial;
3. If the number of control arm deaths from the MAMS design would be no more than 10% greater compared to those from the two-arm trial.

If all three criteria were met then the trials were assessed taking the work of Schiavone et al (2019) and Hague et al (2019), which has been mentioned in chapter two, into consideration to ensure the practicality aspect of employing a MAMS design (Hague et al. 2019, Schiavone et al. 2019). For example, a five arm four stage MAMS with a trial length of four years may show success with the criteria above but to have three stages within four years may not be practical.

Lastly, a further analysis was conducted comparing the sample size required for six separate standard two arm trials versus two separate four arm three stage trials versus three separate three arm two stage trials. All these trials had one control arm and six experimental arms. The aim of this comparison was to further understand if applying MAMS designs showed to be advantageous by requiring fewer patients to achieve the same number of answers.

Trial name	Max. number of potential patients per month (per year)	Recruitment time (years)	Total patients
Breast cancer trials			
DEVA_OS	58 (696)	8	5568
SoFEA_OS	322 (3864)	6	23184
HERA_OS	58 (696)	3.5	2436
NEAT_OS	387 (4644)	6	27864
Colon cancer trials			
QUASAR 2_OS	226 (2712)	5.5	14916
COIN_OS	97 (1164)	3	3492
FOCUS_OS	97 (1164)	2.5	2910
PICCOLO_OS	97 (1164)	2	2328
Lung cancer trials			
FRAGMATIC_OS	184 (2208)	4	8832
FORTIS-M_OS	137 (1644)	2.5	4110
Big Lung_OS	90 (1080)	6	6480
TOPICAL_OS	137 (1644)	4	6576

Table 5.16: The maximum number of potential patients that could be recruited into each of the trials per month and altogether based on the subtype of patients recruited in the original trial.

5.5.2 Example – Third phase calculations

To understand the process for the third phase calculations, the three arm two stage DEVA_OS trial will be used as an example. The DEVA_OS standard parallel trial would require 278 patients (139 patients for each arm) with 33 control arm events at 90% power with one-sided significance of 2.5% to detect an increase in survival at five-years from 81.8% to 91.8%.

The three arm two stage DEVA_OS MAMS trial would provide 95% power at stage one and 90% power at the final stage for each pairwise comparison. A type I error rate of 50% at stage one, 2.5% at the final stage would be used as the significance level for each pairwise comparison. A total of 336 patients (112 patients on each arm) would be required to be recruited in eight years with an anticipated follow-up for 5.5 years. The parameters inputted into the 'nstage' command in Stata can be found in Table 5.17 and the syntax and results can be seen in Figure 5.6.

Parameter	Value
Total number of stages	2
Number of arms in each stage	3, 3
Total accrual	336
Allocation ratio	1
Survival Probability for OS	0.818
Survival Time (years)	5
HR under H0	1
HR under H1	0.426
One-sided alpha for each stage	0.5, 0.025
Power for each stage	0.95, 0.9
Time units	1 (= one year)
Time of recruitment stopping (years)	8

Table 5.17: Parameters inputted for a hypothetical three arm two stage DEVA_OS MAMS trial using the 'nstage' command in Stata, the intermediate and primary outcome was OS. OS – Overall survival, HR – Hazard ratio, H0 – Null hypothesis, H1 – Alternate hypothesis

Using the guidelines specified in 5.5.1.1, to deem this MAMS design as feasible a maximum of 58 patients per month for eight years could be recruited (Table 5.16); the three arm two stage MAMS trial would require only three patients per month. For the control arm, the DEVA_OS trial would require 139 patients with 33 events compared to the MAMS trials which would require 109 patients with 35 events. Therefore, all three procedures were successful so the three arm two stage MAMS design was classed as feasible.

Furthermore, if in the situation where six experimental treatments were available to be compared against a common control for the DEVA_OS trial then three separate three arm two stage DEVA_OS trials would require a significantly fewer number of patients in comparison to performing six separate trials; 984 patients vs. 1668 patients respectively (Figure 5.10).

```
. nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1 1) hr1(0.426 0.426)
> accrue(42 42) arms(3 3) t(5 5) s(0.818 0.818) tstop(8)

n-stage trial design                        version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 3-arm 2-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 17.3 time units

Operating characteristics
-------------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
-------------------------------------------------------------------------------
1          0.5000    0.956    1.000    0.426    1.000    6.831    6.831
2          0.0250    0.903    1.000    0.426    0.626    6.602   13.433
-------------------------------------------------------------------------------
Pairwise Error Rate          0.0238             Pairwise Power   0.8801
Familywise Error Rate (SE)   0.0436 (0.0004)
-------------------------------------------------------------------------------
Note: patient accrual stopped at time  8.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ----------Stage 1----------
           Overall  Control   Exper.
Arms             3        1        2
Acc. rate       42       14       28
Patients*      287       96      191
Events**        24       12       12
           ----------Stage 2----------
           Overall  Control   Exper.
Arms             3        1        2
Acc. rate       42       14       28
Patients*      336      112      224
Events**        69       35       34
           ------------------------------
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at stages 1 and 2
END OF NSTAGE
```

Figure 5.6: Syntax and output obtained for the third phase three arm two stage MAMS design for the DEVA_OS trial.

5.5.3 Results

5.5.3.1 Breast Cancer Trials

Figure 5.7 shows the number of patients that would be required per month for each of the different MAMS designs for the breast cancer trials to achieve the required sample size. The five arm four stage MAMS DEVA_OS trial would require 11 patients per month to be recruited to reach the sample size. As per Table 5.16, an estimated maximum of 58 patients per month could be recruited for this trial hence deeming the first criteria successful. Similarly, it was found that the other MAMS breast cancer trials would require fewer than the estimated maximum number of patients that could be recruited per month (Figure 5.7). Thus, concluding that all these trials had the potential to carry out these MAMS designs.

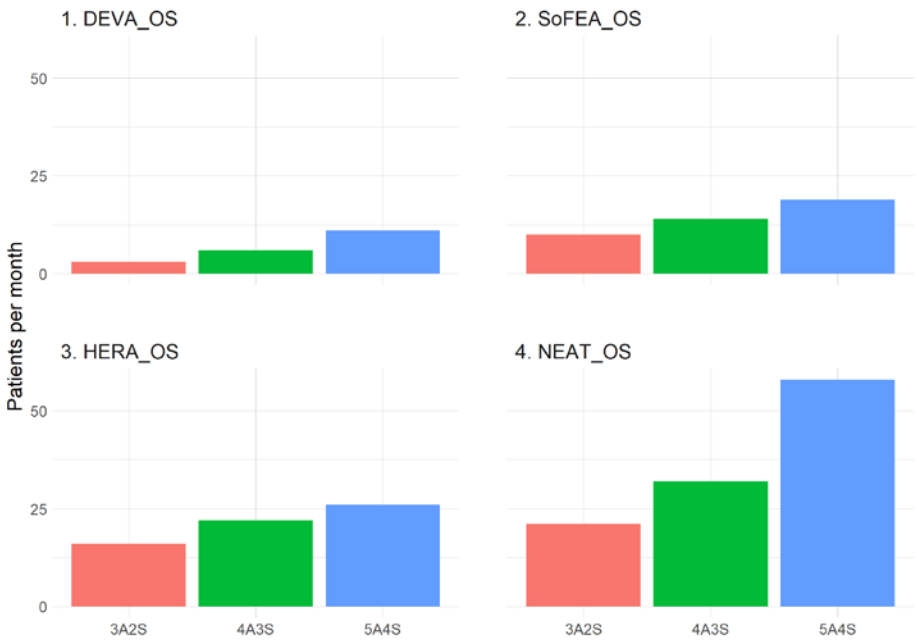

Figure 5.7: Total number of patients required assuming that all experimental arms reach the final stage for a three arm two stage, four arm three stage and five arm four stage MAMS design when applying the parameters obtained from the different breast cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

For the second criteria to be successful, the number of control arm patients required for the MAMS design would be similar to the standard two-arm trial. It can be seen in Figure 5.8 that more than 10% of control arm patients would be required for only the NEAT_OS five arm four stage trial in comparison to the two-arm trial. An increase

in control arm patients for these MAMS designs were observed because with each increase MAMS design, the follow-up length decreased, which indicated that more patients would require required to achieve the appropriate number of events. A simple solution to this would be to decrease the rate at which patients were recruited into the trial however this would lead to ceasing of the recruitment period in a stage prior to the final stage, which is a restriction in the 'nstage' program.

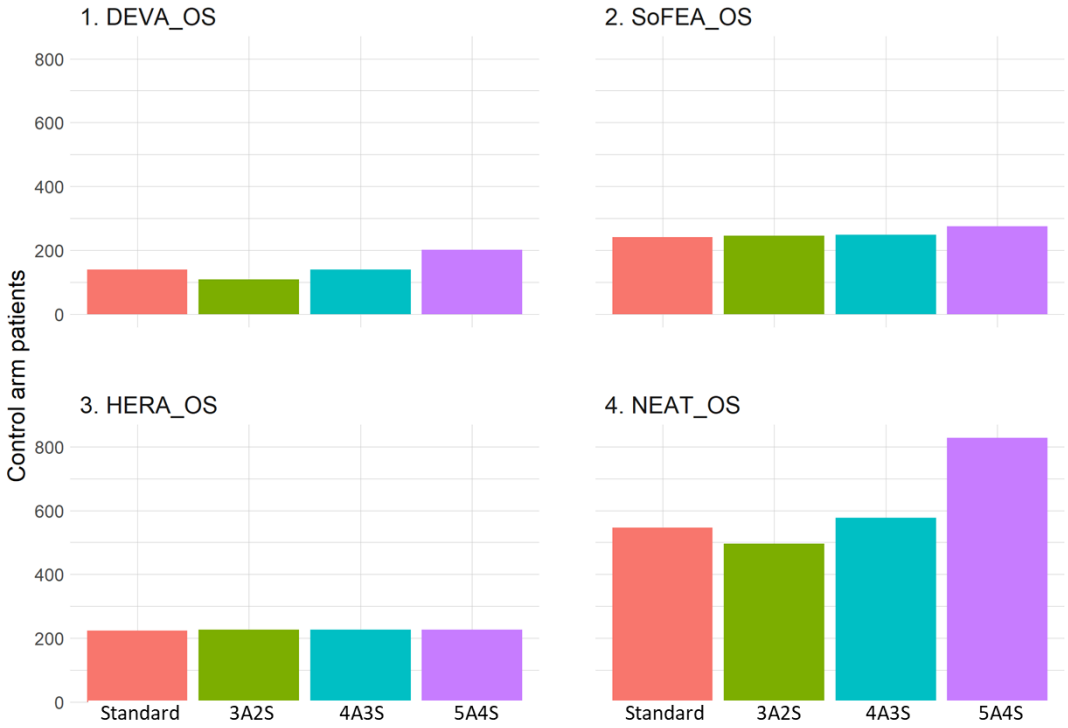

Figure 5.8: The number of control arm patients for the standard design compared to the number of control arm patients required for the different MAMS designs for the breast cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

The third criteria assessed the number of control arm events that would be required in the final stage of the MAMS design to conclude superiority compared to the total number of control arm events required in the standard two-arm trial. The change in the number of control arm events compared to the standard two-arm trial can be seen in Figure 5.9. The maximum decrease in the number of control arm events compared to the standard two-arm trial would be for the five arm four stage DEVA_OS trial which required approximately 12% fewer control arm events at the final stage. The NEAT_OS MAMS designs would require approximately 5% more

control arm events compared to the standard two-arm trial. Although, most of the MAMS designs would require more control arm events, the difference was not considered to be that large (i.e. fewer than nine events), resulting in deeming all MAMS designs as feasible.

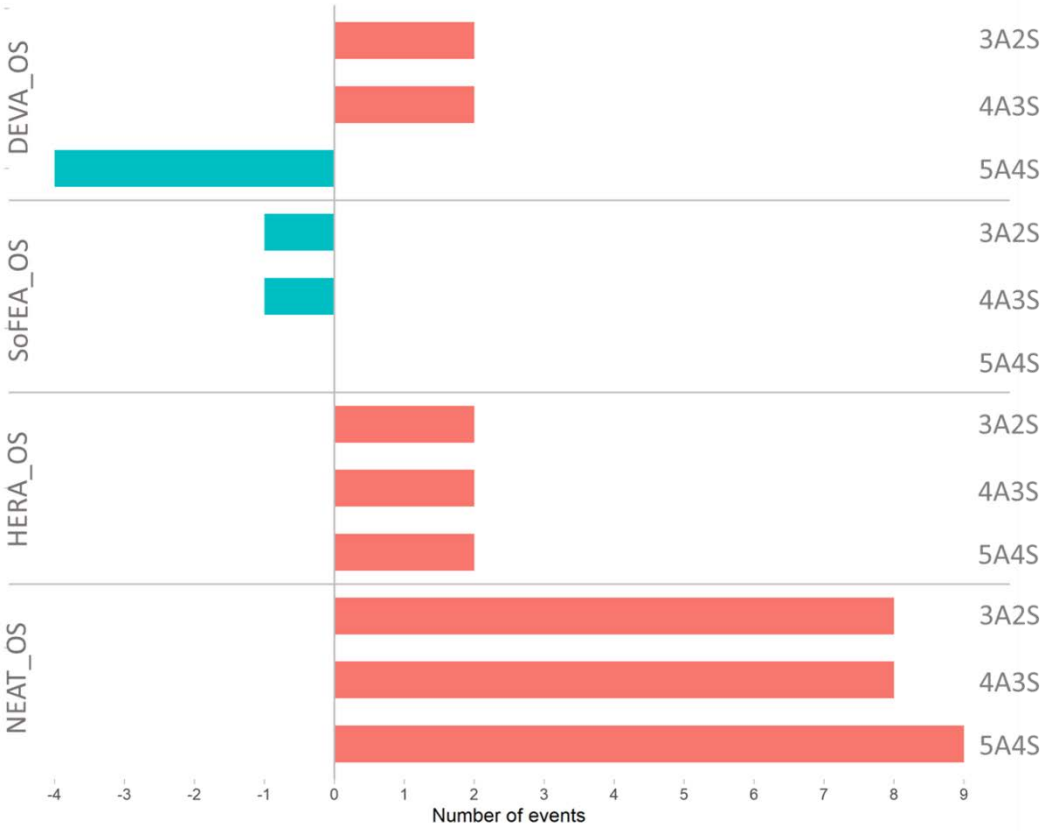

Figure 5.9: The change in the number of control arm events required for the different MAMS designs compared to the standard trial for breast cancer. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

A cross-check of the five arm four stage trials with the total trial length found that to perform the HERA_OS trial would not be viable. The total trial length for the five arm four stage HERA_OS trial would be five years and to plan and execute three interim analyses within this time would not be practical due to the operational tasks to consider prior to each interim analysis i.e. set-up of sites, allowing for sufficient follow-up bearing in mind that the treatment arms looked at different duration of one-year of Herceptin, collecting data, cleaning the data, validating data, arranging meetings with the DMC etc.

A summary of the results for all the breast cancer trials can be seen in Table 5.18. It is clear according to criteria one that all MAMS designs are feasible. The five arm four

stage MAMS designs for the DEVA_OS trial failed the second criteria as more control arm patients would be required compared to the standard design however when further investigated, it was found that the intended total trial length was reduced. To extend the trial length, fewer patients would be required however this would lead to ceasing of the recruitment period in a stage prior to the final stage (see section 2.6).

	Criteria			
	1	2	3	5A4S Cross-check
DEVA_OS				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	×	✓	✓
SoFEA_OS				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	✓	✓	✓
HERA_OS				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	✓	✓	✓
NEAT_OS				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	✓	✓	✓

Table 5.18: Summary of the outcomes for the MAMS designs for all breast cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

The sample size calculations for breast cancer found that if in the situation where six drugs are available for testing against a common control, then to achieve the same number of answers performing either three separate three arm two stage or two separate four arm three stage MAMS trials resulted in requiring fewer patients then conducting six separate standard two-arm trials (Figure 5.10). For example, for the NEAT_OS trial performing three separate three arm two stage trials (4464 patients altogether) or two separate four arm three stage trials (4620 patients) would require significantly fewer patients compared to six separate two-arm trials (6564 patients).

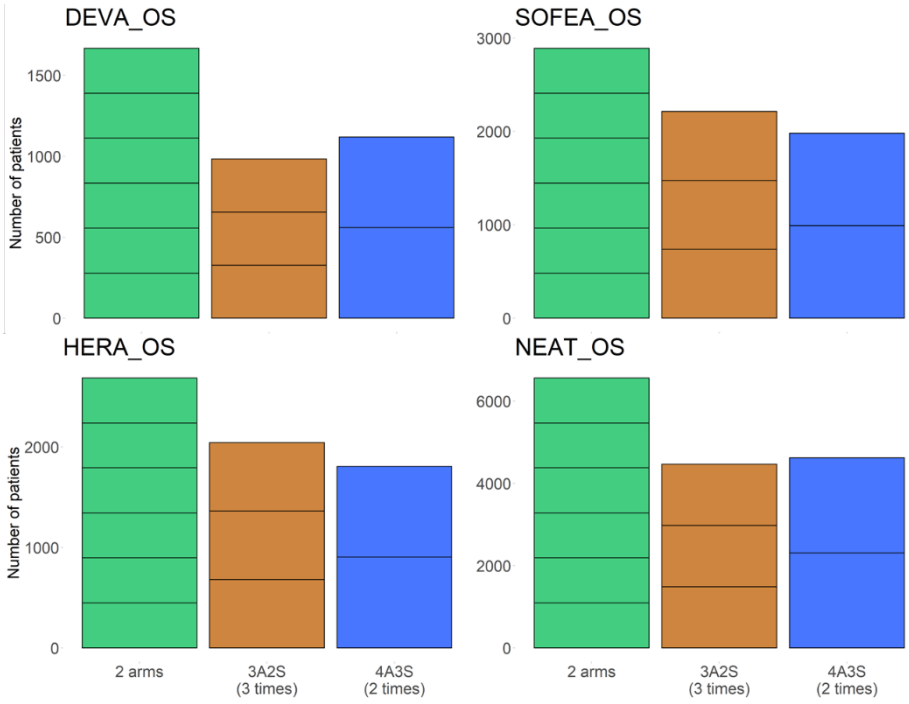

Figure 5.10: Comparing sample sizes to achieve the same number of answers for three separate three arm two stage versus two separate four arm two stage versus six standard parallel designs for breast cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage.

5.5.3.2 Colon Cancer Trials

The number of patients that would be required per month for each of the MAMS designs for the different colon cancer trials are displayed in Figure 5.11. To reach the calculated sample size for the five arm four stage COIN trial, 67 patients per month would need to be recruited for three years. An estimated maximum of 97 patients per month could be recruited for this trial hence deeming the first criteria successful. Similarly, all other MAMS colon cancer trials except the five arm four stage FOCUS_OS trial (total of 106 patients required per month compared to estimated maximum of 97 patients per month) required fewer patients than the estimated maximum number of patients that could be recruited per month.

Figure 5.11: Total number of patients required assuming that all experimental arms reach the final stage for a three arm two stage, four arm three stage and five arm four stage MAMS design when applying the parameters obtained from the different colon cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

Figure 5.12 shows more than 10% of control arm patients would be required compared to the standard two-arm trials for all QUASAR_OS, five arm four stage COIN_OS, FOCUS_OS and PICCOLO_OS trials. The five arm four stage MAMS design for the COIN_OS, FOCUS_OS and PICCOLO_OS had an increase in the number of patients as the length of follow-up decreased for these trials indicating that more patients would be required to obtain the number of events. The three arm two stage and four arm three stage MAMS designs for the COIN_OS and PICCOLO_OS trial showed that fewer control arm patients would be required compared to the standard trial.

Figure 5.12: The number of control arm patients for the standard design compared to the number of control arm patients required for the different MAMS designs for the colon cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

All MAMS designs would require similar control arm events in comparison to the standard two-arm trial (Figure 5.13). The greatest increase in the number of control arm events for the MAMS designs compared to their respective standard two-arm trial would be for the five arm four stage FOCUS_OS trial which would require an additional five control arm events. The five arm four stage MAMS trials for the COIN_OS, FOCUS_OS and PICCOLO_OS all had an increase in the number of control arm events that would be required resulting from a decrease in the total trial length leading to reduced times between each interim stage.

Although, most of the MAMS designs would require more control arm events, the difference would not be that large (i.e. fewer than five events and less than 10%). This resulted in deeming all MAMS designs as feasible.

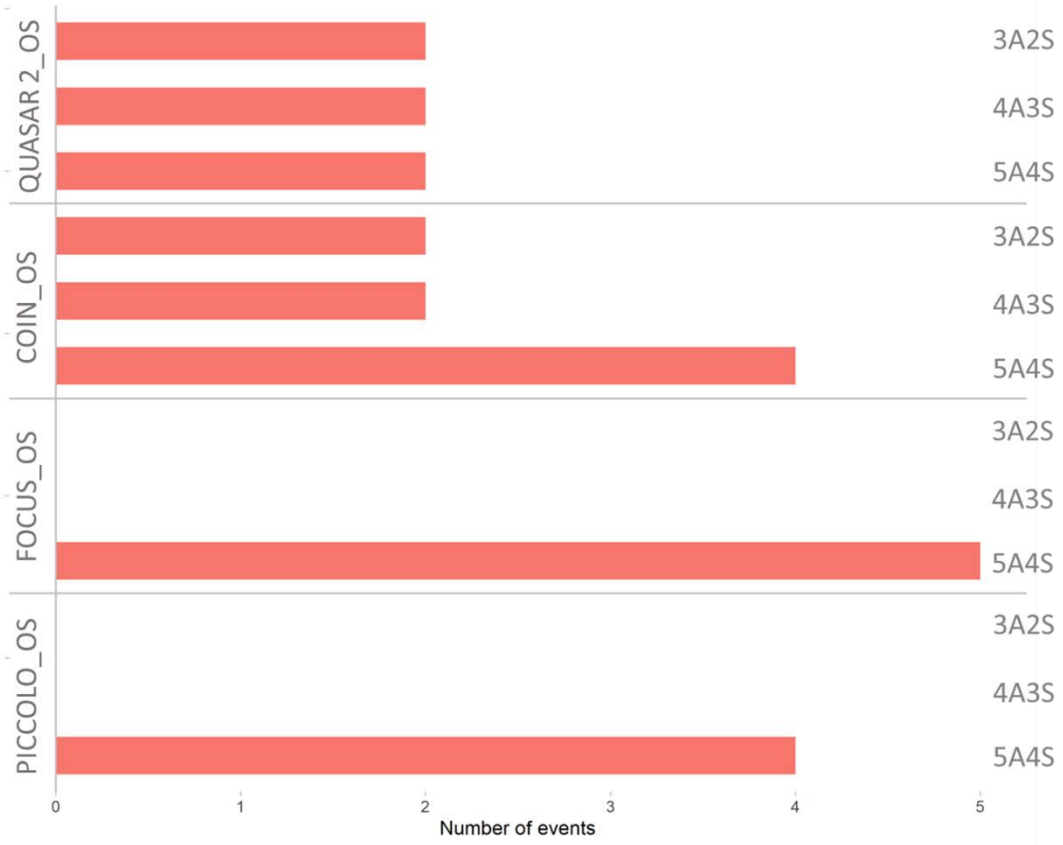

Figure 5.13: The change in the number of control arm events required for the different MAMS designs compared to the standard trial for colon cancer. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

A cross-check of the five arm four stage trials with the total trial length found that to perform the COIN_OS, FOCUS_OS and PICCOLO_OS trials would not be practical as the total trial length for these trials were 3.7, 3.16 and 2.5 years respectively. Hence, to accommodate three interim analyses within this short space of time would be impractical.

A summary of the results for all colon cancer trials can be seen in Table 5.19. According to first criteria from the guidelines, all MAMS designs would be feasible except the five arm four stage FOCUS_OS trial. There is an increase in the number of control arm patients for some of the MAMS designs compared to the two-arm trials.

	Criteria			
	1	2	3	5A4S Cross-check
QUASAR 2_OS				
3A2S	✓	×	✓	
4A3S	✓	×	✓	
5A4S	✓	×	✓	
COIN				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	×	✓	×
FOCUS				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	×	×	✓	×
PICCOLO				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	×	✓	×

Table 5.19: Summary of the outcomes for the MAMS designs for all colon cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

If in the situation where six drugs are available for testing against a common control, then to achieve the same number of answers, performing either three separate three arm two stage or two separate four arm three stage MAMS trials for colon cancer resulted in requiring fewer patients then conducting six separate standard two-arm trials (Figure 5.14).

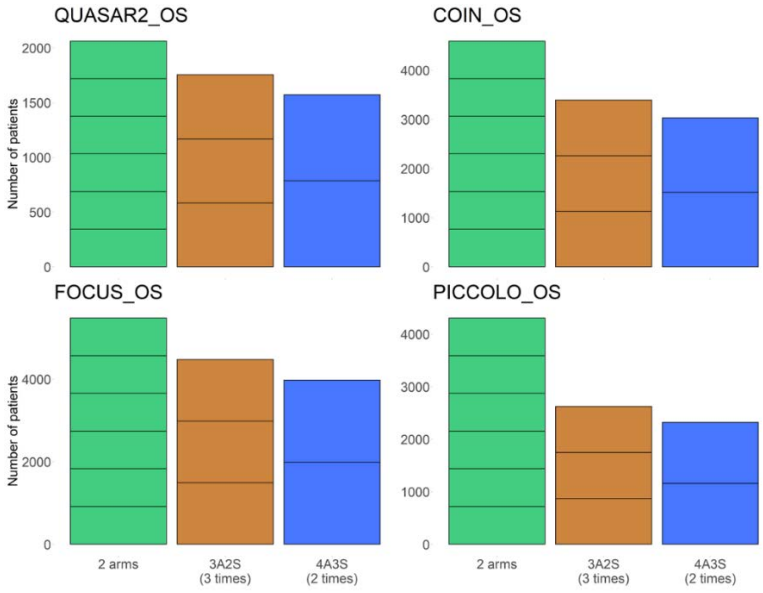

Figure 5.14: Comparing sample sizes to achieve the same number of answers for three separate three arm two stage versus two four arm two stage versus six standard parallel designs for colon cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage.

5.5.3.3 *Lung cancer Trials*

The number of patients that would be required per month for each of the MAMS designs for the different lung cancer trials are displayed in Figure 5.15. To reach the calculated sample size for the five arm four stage TOPICAL trial, 50 patients per month would need to be recruited for three years. An estimated maximum of 137 patients per month could be recruited for this trial hence deeming the first criteria successful. Similarly, the other MAMS lung cancer trials required fewer patients compared to the estimated maximum number that could be recruited per month concluding that these trials had the potential to carry out these MAMS designs.

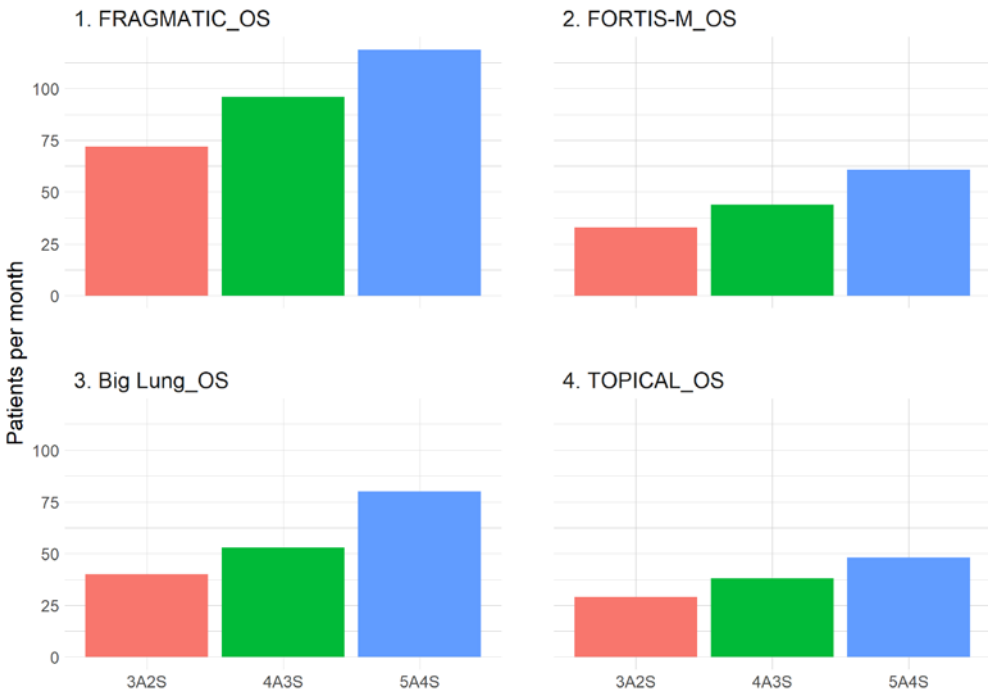

Figure 5.15: Total number of patients required assuming that all experimental arms reach the final stage for a three arm two stage, four arm three stage and five arm four stage MAMS design when applying the parameters obtained from the different lung cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

More than 10% of control arm patients would be required compared to the standard designs for the five arm four stage Big Lung_OS (Figure 5.16). All FORTIS_OS MAMS trials, the three arm two stage and four arm three stage Big Lung_OS MAMS designs showed that fewer control arm patients would be required compared to the standard trial. The five arm four stage Big Lung_OS trial would have the length of follow-up

resulting in an increase in the number of patients that would be required to reach the correct number of events.

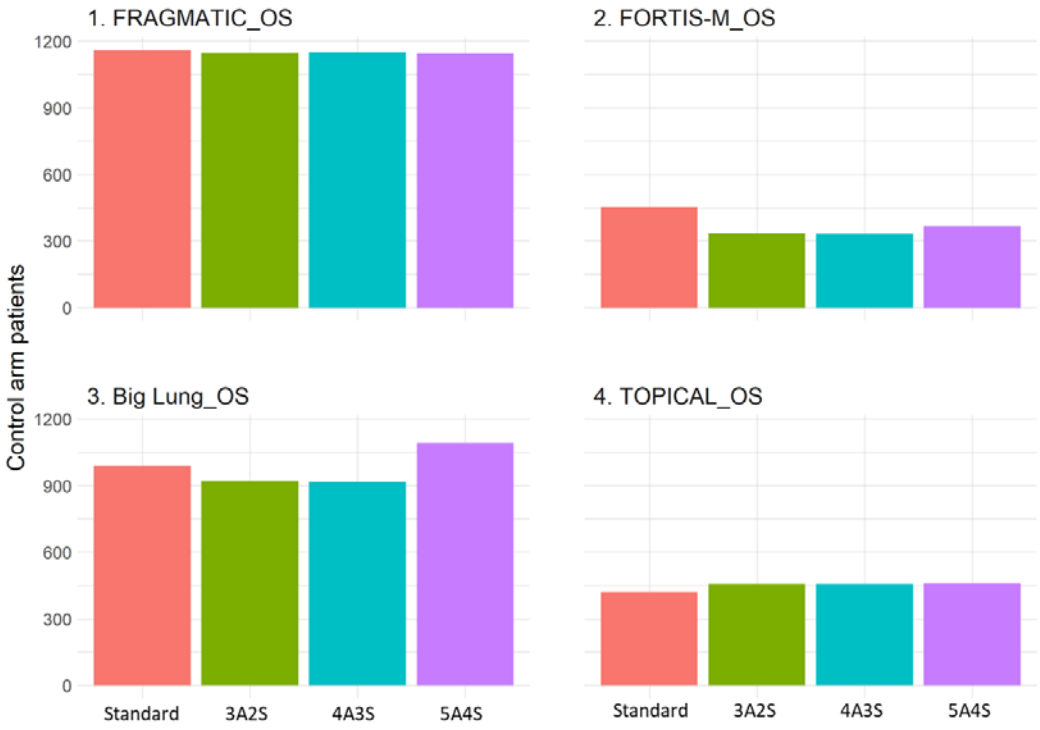

Figure 5.16: The number of control arm patients for the standard design compared to the number of control arm patients required for the different MAMS designs for the lung cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

The number of control arm events that would be required for the FRAGMATIC_OS, FORTIS-M_OS and TOPICAL_OS trials did not greatly differ from the events required for their standard two-arm trial (maximum increase/decrease of 2%) (Figure 5.17). The largest change of 11 events could be seen in the Big Lung_OS trial where the standard design required 728 deaths compared to 739 control arm deaths which is an increase of 1.5% required in the five arm four stage MAMS design. The difference in the number of control arm events between the standard two-arm trials and the MAMS trials was not considered to be that large. This resulted in deeming all MAMS designs as feasible in accordance with the guidelines.

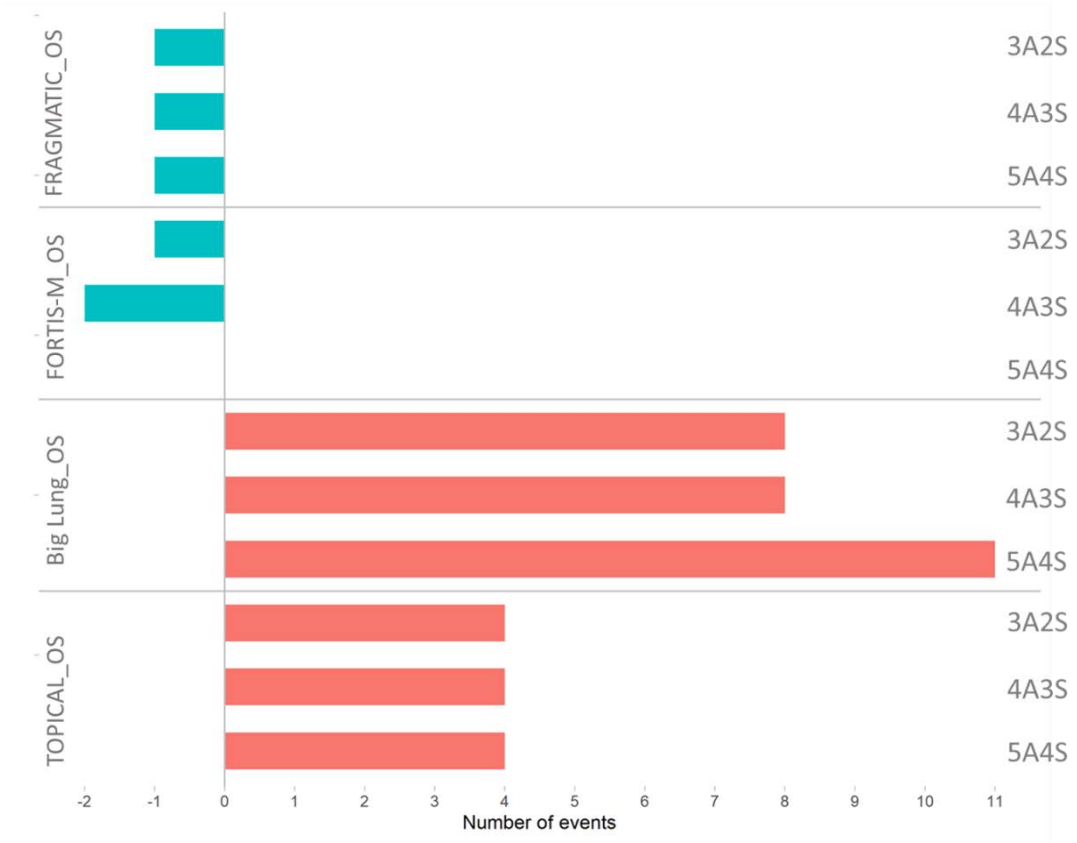

Figure 5.17: The change in the number of control arm events required for the different MAMS designs compared to the standard trial for lung cancer. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

Taking into consideration the administrative tasks required to execute MAMS designs, a cross-check of the five arm four stage trials with the total length concluded that these designs could be executed as the minimum length for these trials were six years.

The results for all the lung cancer trials can be seen in Table 5.20. According to first criteria from the guidelines, all MAMS designs are feasible. There is an increase in the number of control arm patients for the Big Lung_OS MAMS trial compared to their respective two-arm trials for the second criteria and the control arm events were all the trials were similar to the MAMS designs.

	Criteria			
	1	2	3	5A4S Cross-check
FRAGMATIC				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	✓	✓	✓
FORTIS-M				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	✓	✓	✓
Big Lung				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	×	✓	✓
TOPICAL				
3A2S	✓	✓	✓	
4A3S	✓	✓	✓	
5A4S	✓	✓	✓	✓

Table 5.20: Summary of the outcomes for the MAMS designs for all lung cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

If in the situation where six drugs are available for testing against a common control, then to achieve the same number of answers, performing either three separate three arm two stage or two separate four arm three stage MAMS trials for lung cancer would result in requiring fewer patients then conducting six separate standard two-arm trials (Figure 5.18).

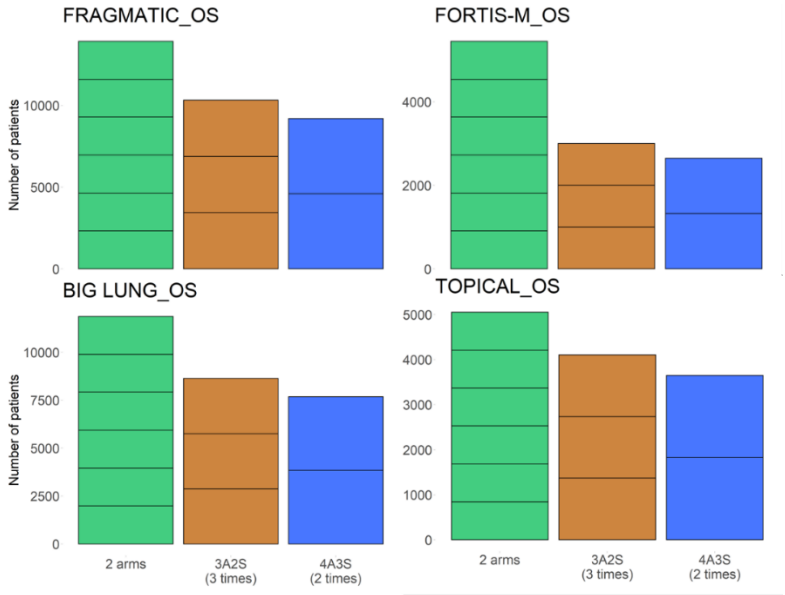

Figure 5.18: Comparing sample sizes to achieve the same number of answers for three separate three arm two stage versus two four arm two stage versus six standard parallel designs for lung cancer trials. 3A2S = Three arm two stage; 4A3S = Four arm three stage.

5.6 Discussion

The sample of phase III trials that were taken from the Cancer Research Centre UK were used to implement MAMS designs for all phases of the calculations. The findings from the calculations indicated that the application of MAMS designs within breast, colon and lung cancer trials would be feasible. It is difficult to establish the ideal situation where applying a MAMS design would be appropriate as there are many variables that can influence the outcome of these designs. The OS rate gives an indication of the number of events that would be accumulated over the course of the trial; hence an improved OS rate would indicate fewer events would occur so a MAMS design would need to recruit more patients. The longer the overall trial length (recruitment and follow-up period), the more time is given for the accrual of events hence reducing the number of patients required within a MAMS design. In the first phase of calculations, the first interim stage for a three arm two stage breast cancer trial took place at seven years, this may not be practical for trialists hence a solution could be to increase the number of patients recruited into the trial to reach the required number of events earlier or to use a surrogate endpoint as an intermediate outcome. Furthermore, fewer patients are required to detect a greater absolute treatment difference. Moreover, it was decided to use the alpha stagewise values as specified by Royston et al (2011) as these resulted in being more robust compared to the values used in the first phase calculations and had been implemented in the STAMPEDE trial. However, the choice of alpha and power at each stage can be decided collectively by the trial team based on the requirements of the trial. The choice of the alpha can be smaller, for example the FOCUS4 utilises a one-sided alpha value of 0.3 in the first stage. Using a smaller alpha value would mean a more rigorous futility boundary and therefore more events would be required to trigger the interim analysis. Trial stakeholders may be more familiar with an overall power of 80% and overall alpha of 5% in comparison to 88% power and 2% alpha, for which more stringent stagewise alpha and power values would be used.

The third phase of calculations considered the extreme case of MAMS designs where all experimental arms continue to the final stage because from a funding perspective, it would be wise to ensure that trial stakeholders apply for the maximum amount

possible to ensure the trial does not close due to lack of funds. It is evident from these calculations that the use of MAMS designs can be implemented within cancer sites with longer term outcomes. However, it is recommended that due diligence is taken for higher survival rates which result in fewer events. Trialist and patients may not be convinced that conclusions can be drawn with so little events. For example, the HERA_OS trial required approximately 20 deaths to deem superiority of trial drugs. In this situation, it would be better to use a surrogate outcome and use OS as a secondary endpoint. The third criteria to deem these MAMS designs as feasible was if the number of control arm deaths would be no more than 10% greater compared to those from the two-arm trial. The number of control arm deaths for all the MAMS trials did not vary much and the difference could be due to rounding error. Some of the five arm four stage trials considered in this chapter struggled to adhere to the intended trial length set out by the 'nstage' command resulting in an increase in the number of patients and therefore events required. For example, the five arm four stage COIN_OS, FOCUS_OS and PICCOLO_OS trials resulted in a total trial length of 3.7, 3.2 and 2.5 years respectively. Therefore, it is recommended to ensure various endpoints are considered with a range of trial lengths which will lead to the decisions being made in the most efficient, effective and ethical manner.

To add to the integrity and practicality of applying MAMS designs, the calculations could be further enhanced by implementing changes to the allocation ratio or additional eligibility criteria to ensure accuracy when gauging the maximum number of patients that could be recruited within a MAMS phase III trial. For example, the incidence rate of cancer by age group or ethnicity etc. Also, the MAMS designs within this chapter assumed that patients were accrued at a uniform rate, however in practice this may not be the case. For example, in the STAMPEDE trial, the initial recruitment began slowly which was due to the time taken in the opening of sites (Schiavone et al. 2019), as the momentum of the trial picked up, the rate at which patients were recruited increased. Therefore, changing the rate at which patients are accrued can add to the integrity when applying MAMS designs.

However, further research could be undertaken to investigate the practicality of performing these MAMS design as there are other factors that have not been

considered, such as further defining the incidence rates of cancer by taking baseline characteristics such as age, gender ethnicity etc. into consideration to estimate realistic sample size targets. Furthermore, only the incidence rates of the subgroups have been calculated and used for each of the trials in the third phase of calculations to estimate the maximum number of patients that can be recruited. For example, the DEVA trial assessed only women that were postmenopausal, which has not been taken into consideration. Additionally, there are many challenges when recruiting patients which needs to be taken into consideration, for example the poor health status for lung cancer patients may deem it difficult to recruit patients into studies (Schofield et al. 2008). However, the estimates calculated for the maximum number of patients recruited are limited to the UK hence extending recruitment to international sites could greatly benefit the recruitment into trials. It was essential that these estimates were calculated based upon recent data to ensure the validity of the comparisons made within this chapter as the incidence of cancer has significantly increased over the past 30 years with improvements in cancer detection but also the survival rates have improved with improvements in treatments (Cancer Research UK 2018).

The results from the third phase of calculations show that based on the estimated maximum number of patients, MAMS designs using up to four experimental arms are feasible in these common disease settings. It was evident that in certain situations more control arm patients are required but the number of events required for the control arm did not vary significantly compared to the standard two-arm trials. Additionally, when comparing the total sample size of various MAMS trials against repeating two-arm trials, it was evident that MAMS designs require fewer patients because all experimental arms for the MAMS design share a common control. Hence, MAMS designs answer many questions within one trial which can considerably reduce the cost associated with trials compared to conducting separate two-arm trials as fewer patients and fewer regulatory applications are required.

The outcome from this chapter support the results produced from a previous paper, whereby four trials in three cancer sites (renal, ovarian and colorectal cancer) with a standard parallel design were reanalysed implementing the MAMS methodology

(Barthel et al. 2009). The authors concluded that implementing MAMS designs can be both effective and efficient in evaluating if possible, many different treatments against a control, which supports the outcome of this chapter.

Phase III clinical trials within the field of Oncology are continually increasing in terms of size and expense, however there is a concern with regards to the rate of success and the cost associated with running these trials. Hence there is a need for novel adaptive designs that can answer multiple questions within one trial (Berry 2012, Reitsma et al. 2015, Wilson et al. 2015, Hind et al. 2017, Parmar et al. 2017). The use of MAMS designs within phase III trials can be implemented to aid the advancement and testing of new therapies (Parmar et al. 2017). Furthermore, research could be carried out to investigate the relationship between MAMS designs and Umbrella trials, Basket trials and Platform-based designs; all of which answer many questions within one trial.

5.7 Conclusion

The aim of this chapter was to develop a firm understanding of whether MAMS designs could work with longer term outcomes and to understand the development of these designs with varying survival rates. This chapter has shown that MAMS designs using a superiority hypothesis can be implemented in a variety of settings and not just in areas with poor survival. The next chapter will focus on implementing MAMS designs with a NI hypothesis.

6 MAMS designs extended to a NI hypothesis

6.1 Introduction

The many toxicities endured by patients from cancer therapies has encouraged trialists to find solutions that are just as effective but result in a reduced number of side effects (Riechelmann et al. 2013). Furthermore, the COVID-19 pandemic which began in 2019, has encouraged trial stakeholders to urgently seek-out treatments that are more convenient and require reduced time for patients to spend in hospital. Therefore, implementing trial designs with a NI hypothesis allows trial stakeholders to compare potential treatments that are less toxic, cost effective but just as efficient as the standard treatment.

This chapter extends the previous use of MAMS designs with a superiority hypothesis to a NI hypothesis and explains how the 'nstage' command can still be used to undertake these designs. Firstly, the underlying methodology applied to the MAMS design is extended for use with a NI hypothesis, thereafter simulations were performed to evaluate the statistical properties of this design using the parameters from the selected trials identified in the previous chapter. Lastly, like the third phase of calculations in the previous chapter, hypothetical MAMS designs were obtained using the selected trials and compared to running standard two-arm designs with OS as the primary endpoint.

6.2 Methodology to use a NI hypothesis within the MAMS framework

The MAMS framework specified by Royston et al (2011) and details provided in section 2.6 is intended to be implemented for a trial with a one-sided superiority hypothesis. However, it was shown in chapter one that the properties used in one-sided superiority designs will result in the same number of events required using the same parameters for a NI design. NI trials tend to be larger in size as they allow for a small loss of effectiveness and more events are required to detect these smaller treatment differences compared to superiority trials that seek to gain efficacy with larger treatment differences. The below sections demonstrate how the methodology for a MAMS design can be adapted for use with a NI hypothesis and how it can be applied using the 'nstage' command.

Let IO and PO be defined as the time-to event intermediate outcome and primary outcome respectively, where $IO = PO$ can also be a possibility for a MAMS designs using a NI hypothesis.

Let $j \geq 1$ denote the number of experimental arms for which pairwise comparisons will be made against a common control arm, C. The true HR, Δ_{ij} is the effect of the experimental arm, j on the control arm on the specified outcome at stage i, for $i = 1, \dots f$ where f is the final stage. The true HR is calculated using the IO from stage 1 to stage $f - 1$ and using the PO at stage f. It is assumed that the proportional hazards assumption holds for all treatment comparisons. The null and alternative hypotheses can be defined as:

$$H_{ij}^{0}: \Delta_{ij} = \Delta_{ij}^{0},$$
$$H_{ij}^{1}: \Delta_{ij} = \Delta_{ij}^{1}, \quad i = 1, \dots, f, j \geq 1$$

where Δ_{ij}^{0} is the value at the edge of the inferiority region for which the HR > 1 and Δ_{ij}^{1} is the no-difference value where the HR = 1. This differs from the superiority hypothesis which assumes at the null hypothesis there is no difference. Therefore, the probability to claim NI under the null and alternative hypothesis is desired to be as:

$$Pr(Claim\ non - inferior |\ inferior) = \alpha$$

$$\Pr(Claim\ non - inferior |\ non - inferior) = 1 - \beta = \rho$$

where α and β represent the type I and type II error rates respectively (see Figure 6.2).

A non-inferiority test of the cumulated data for each pairwise comparison is performed at stage i with nominal type I error rate α_i and power $\rho_i = 1 - \beta_i$ where β_i is the type II error rate at each stage i. If this is significant at level α_i (indicating NI) then just like the MAMS design using the superiority hypothesis the experimental arm continues onto the next stage up until stage $f - 1$, and at stage f, NI is concluded.

Under H_0 and H_1, the estimated log HR at stage i, $\ln \hat{\Delta}_i$ is distributed as:

$$H_0: \ln \hat{\Delta}_i \sim N(\ln \Delta_i^0 , v_i^0)$$

$$H_1: \ln \hat{\Delta}_i \sim N(\ln \Delta_i^1 , v_i^1)$$

where the estimated variances under H_0 and H_1 are v_i^0 and v_i^1 respectively. The one-sided significance level for each of these hypotheses can be specified such that:

$$\begin{aligned} \alpha_i = \quad & \Pr\left(\ln \hat{\Delta}_i > \ln \delta_i | \, H_0\right) \\ & \Pr\left(\frac{\ln \hat{\Delta}_i - \ln \Delta_i^0}{\sigma_i^0} > \frac{\ln \delta_i - \ln \Delta_i^0}{\sigma_i^0} | H_0\right) \\ & \Phi\left(\frac{\ln \delta_i - \ln \Delta_i^0}{\sigma_i^0}\right) \\ & \Phi\left(z_{\alpha_i}\right) \end{aligned}$$

where σ_i^0 is defined as square root of v_i^0, the critical HR at stage i is defined as δ_i and $\Phi(\,)$ is the standard normal distribution function.

The power for each of these hypotheses can be defined as follows:

$$\begin{aligned} \rho_i = \quad & \Pr\left(\ln \hat{\Delta}_i > \ln \delta_i | \, H_1\right) \\ & \Pr\left(\frac{\ln \hat{\Delta}_i - \ln \Delta_i^1}{\sigma_i^1} > \frac{\ln \delta_i - \ln \Delta_i^1}{\sigma_i^1} | H_1\right) \\ & \Phi\left(\frac{\ln \delta_i - \ln \Delta_i^1}{\sigma_i^1}\right) \\ & \Phi\left(z_{\rho_i}\right) \end{aligned}$$

where σ_i^1 is defined as square root of v_i^1.

Therefore, the probability under the null or alternative hypothesis of reaching stage $i+1$ or concluding NI of either the control or experimental arm are based on normal approximations such that the appropriate critical HR and events are given by the multivariate tail areas, which is the same when using a superiority hypothesis.

6.3 NI hypothesis application using 'nstage'

In practice, for the methodology discussed above to work using the 'nstage' command in Stata, the values Δ_{ij}^{0} and Δ_{ij}^{1} are entered as 'hr0' and 'hr1' (Figure 6.1). Using the example given in chapter two (section 2.6.8) where a four arm three stage MAMS trial was designed with an assumed five-year OS rate for the control arm of 50.5% to find an increase in OS of 7% in the experimental arm which resulted in a relative HR of 0.81. This example was amended to reflect the use of a NI hypothesis instead of a superiority hypothesis to demonstrate the interchangeability of this command (Figure 6.2). Therefore, the five-year OS rate for these patients were 57.5% with an absolute NI margin of 7% which resulted in a relative HR of 1.23. Using these assumptions resulted in 4368 patients required to obtain the appropriate number of events in the same amount of time using a NI hypothesis compared to 3750 patients required using a superiority hypothesis as shown in the example in section 2.6.8. The reason why there is a difference in the number of patients is because there is an improved survival rate in the control arm when applying a NI hypothesis which requires more patients to reach the required number of events when the duration of the study is fixed.

```
. nstage, nstage(3) accrue(728 728 728) arms(4 3 2) alpha(0.5 0.25 0.025) hr0(1.23 1.23)
> hr1(1 1) omega(0.95 0.95 0.90) t(5 5) s(0.575 0.575) aratio(1) tunit(1) tstop(6)

n-stage trial design                    version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 6.3 time units

Operating characteristics
------------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
------------------------------------------------------------------------------
1           0.5000    0.950    1.230    1.000    1.230    3.800    3.800
2           0.2500    0.950    1.230    1.000    1.158    1.610    5.410
3           0.0250    0.900    1.230    1.000    1.085    2.408    7.818
------------------------------------------------------------------------------
Pairwise Error Rate          0.0214              Pairwise Power   0.8577
Familywise Error Rate (SE)   0.0547 (0.0005)
------------------------------------------------------------------------------
Note: patient accrual stopped at time  6.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ----------Stage 1----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      728      182      546
Patients*     2767      692     2075
Events**       508      127      381
           ----------Stage 2----------
           Overall  Control   Exper.
Arms             3        1        2
Acc. rate      728      243      485
Patients*     3938     1082     2856
Events**       756      252      504
           ----------Stage 3----------
           Overall  Control   Exper.
Arms             2        1        1
Acc. rate      728      364      364
Patients*     4368     1297     3071
Events**       982      491      491
           ------------------------------
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE
```

Figure 6.1: Syntax and output obtained from implementing the 'nstage' command in Stata for a four arm three stage trial with a NI hypothesis.

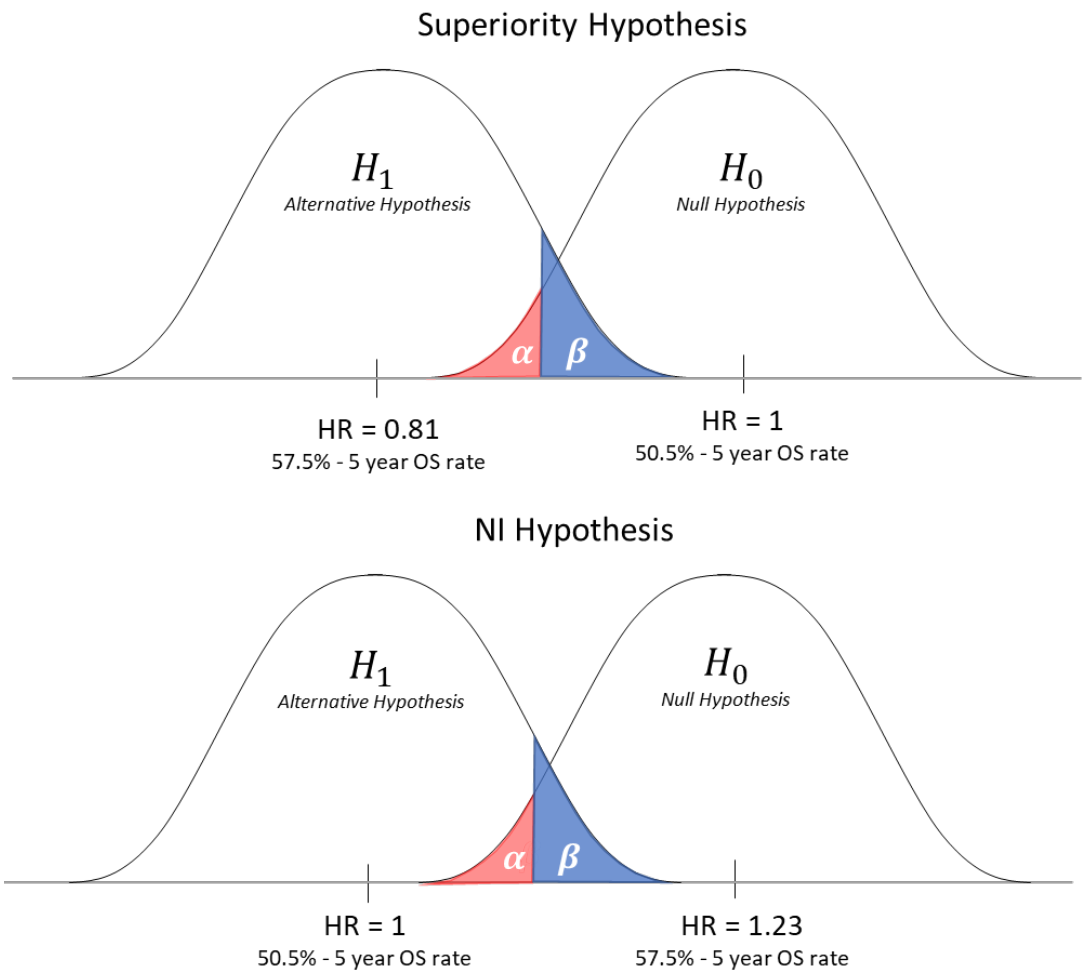

Figure 6.2: Demonstrating the interchangeability between a superiority hypothesis and NI hypothesis from section 2.6.8. α - Type I error rate; β - Type II error rate; HR – Hazard Ratio; OS – Overall Survival,

This design would require 4368 patients recruited for six years and followed up for approximately two years for this fictional sample size calculation (Figure 6.1). It was assumed that one experimental arm was inferior (obtained HR greater than the critical HR) at each stage. The first interim analysis would take place when 127 control arm events are obtained at approximately 3.8 years. The second interim analysis would take place when 252 control arm events are obtained at approximately 5.4 years and the final analysis would take place when 491 control arm events are obtained at 7.8 years. The critical HR at stages one and two are 1.23 and 1.16 respectively. At these interim stages, pairwise comparisons are made for each experimental arm against the control arm; if the obtained HR is less than the critical

HR than that experimental arm would continue to the next stage (Figure 6.3). The critical HR at the final stage is 1.09, it was assumed that one experimental arm would reach this stage and if the obtained HR was less than the critical HR then NI would be declared.

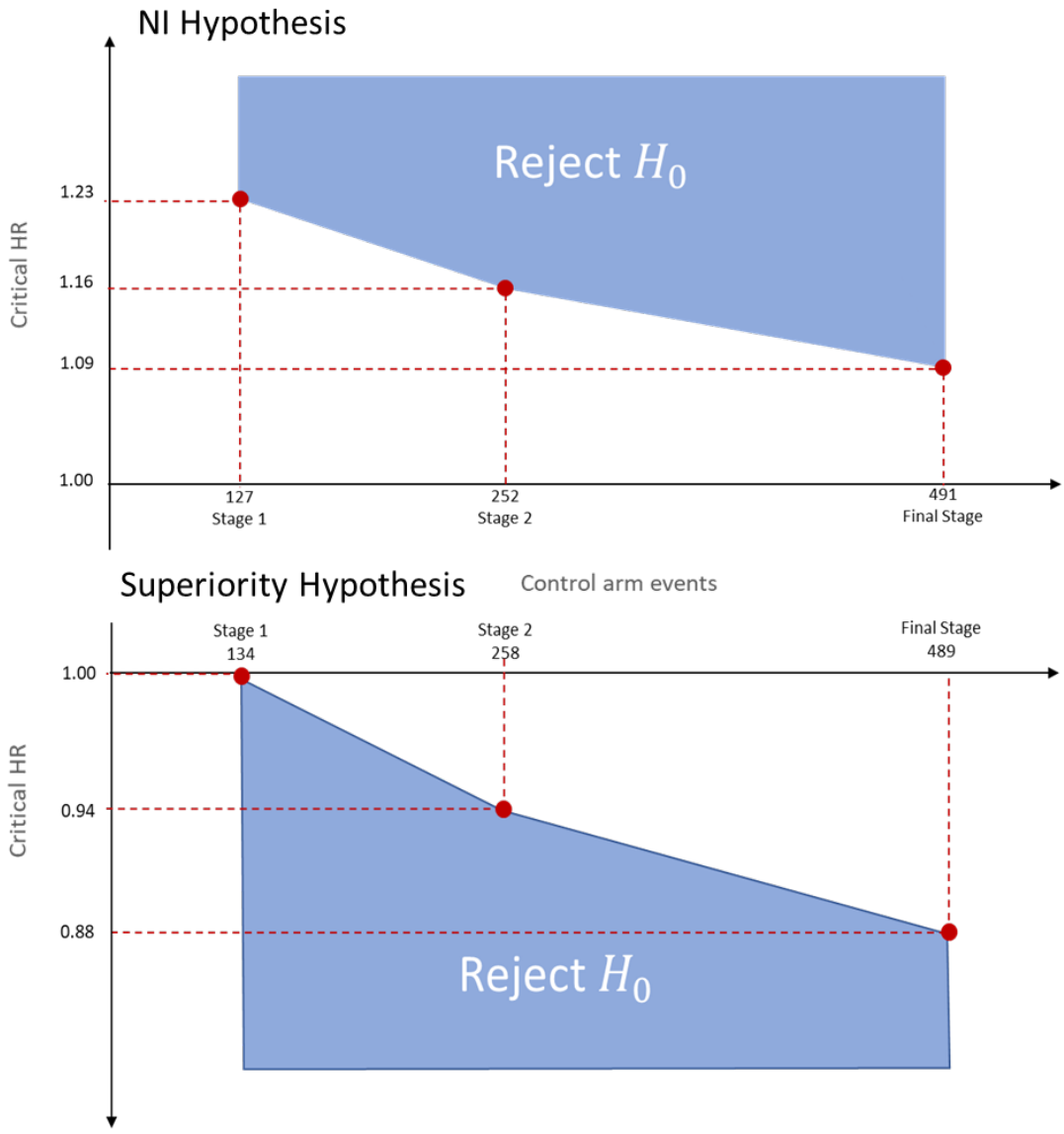

Figure 6.3: Region for which H_0 is rejected for the pairwise comparisons at each stage for the example four arm three stage MAMS design using a NI hypothesis and comparing it against the example used in chapter two with a superiority hypothesis. NI = Non-inferiority; HR = Hazard Ratio.

The switching from a one-sided superiority hypothesis used in chapter two to a NI hypothesis resulted in similar number of control arm events as anticipated from the theory provided in section 1.4.5 and previous section within this chapter. The similarity of the number of events when changing between both the one-sided

superiority and NI hypotheses supports the use of the MAMS framework by Royston et al (2011) for MAMS designs with a NI hypothesis. However, to confirm this empirical finding, simulations were performed to validate the statistical properties of the MAMS design with a NI hypothesis.

6.4 Simulations to validate MAMS design with NI hypothesis

Simulations studies can be implemented to evaluate the performance of innovative statistical methods in pre-defined situations (Burton et al. 2006). Simulations were performed to evaluate the statistical properties of the NI designs obtained using the 'nstage' command. The number of events for each treatment arm, the type I error rate and the power obtained from the simulations were compared with the corresponding parameters from using the 'nstage' command.

The first and second phase calculations in the previous chapter were exploratory to help further understand the MAMS designs hence the calculations applied within this chapter use the methods applied for the third phase of calculations. Due to the poor survival rates of patients diagnosed with lung cancer, it would not be appropriate to apply a NI hypothesis. The results from the third phase of calculations in the previous chapter for the four different trials within each cancer site gave similar conclusions therefore, hypothetical MAMS designs with a NI hypothesis were applied only to the DEVA trial (breast cancer) and the QUASAR 2 trial (colon cancer).

6.4.1 Aims

The aims of the simulations were (1) to evaluate the stagewise and pairwise type I error and power and (2) to compare operational characteristics to the output from 'nstage' command for a three arm two stage MAMS design when using a NI hypothesis. Three scenarios using different alpha and power values for the DEVA and QUASAR2 trials were used to validate the statistical properties of these designs (Table 6.1). These two trials will be referred to as DEVA_NI and QUASAR 2_NI respectively to emphasise that a NI hypothesis has been implemented.

Scenario	Alpha		Power	
	Stage One	Final Stage	Stage One	Final Stage
Scenario 1	0.5	0.025	0.95	0.9
Scenario 2	0.5	0.05	0.9	0.9
Scenario 3	0.25	0.05	0.9	0.85

Table 6.1: The different stage one and final stage alpha and power values used to evaluate the simulations.

6.4.2 Method

The sample size for the three arm two stage DEVA_NI trial and QUASAR 2_NI trial was calculated using a 3% absolute NI margin and the parameters outlined in Table 6.2 using the 'nstage' command.

Trial	Parameters	
DEVA_NI	H0	1.188
	H1	1
	Accrual rate per year	845
	Survival rate	5-year OS - 81.8%
	Recruitment period (years)	8
QUASAR 2_NI	H0	1.305
	H1	1
	Accrual rate per year	579
	Survival rate	3-year OS - 89.4%
	Recruitment period (years)	5.5

Table 6.2: Parameters used to calculate the sample size using the 'nstage' command.

The results of the sample size calculations which were used in the simulations can be seen in Table 6.3 with full details of the results in appendix 7. From these results the critical HR, the time (when no arms dropped and when 1 arm dropped) and the total sample size were used in the simulations. The type I error and power at stage one (Table 6.1), the PWER, pairwise power and the number of control arm events at each stage were all validated by the simulations.

Trial	Results	Scenario 1		Scenario 2		Scenario 3	
		Stage		Stage		Stage	
		1	2	1	2	1	2
DEVA_NI	PWER		0.023		0.042		0.044
	Pairwise Power		0.87		0.83		0.81
	Critical HR	1.188	1.070	1.188	1.078	1.119	1.069
	Time (no arms dropped)	5.9	13.5	4.6	11.5	7.1	10.2
	Time (1 arm dropped)	5.9	12.6	4.6	10.4	7.1	10.0
	Sample Size	4995	6760	3857	6760	5988	6760
	CA Events	183	710	111	579	259	486
QUASAR 2_NI	PWER		0.023		0.042		0.044
	Pairwise Power		0.87		0.83		0.81
	Critical HR	1.305	1.111	1.305	1.124	1.191	1.109
	Time (no arms dropped)	4.8	11.6	4.4	11.9	5.7	8.2
	Time (1 arm dropped)	4.8	11.1	4.4	10.8	5.7	8.2
	Sample Size	2755	3185	1808	2665	3297	3474
	CA Events	77	298	47	243	109	204

Table 6.3: Results from the sample size calculation using the 'nstage' command and used in the simulations. PWER = Pairwise Error Rate; HR = Hazard Ratio; CA = Control arm

Data were simulated based on the total sample size for each of the trials, i.e. for the three arm two stage DEVA_NI trial, the number of observations, $n^{obs} = 6760$. Data were generated under the null hypothesis and the alternative hypothesis for each scenario as detailed in Table 6.1.

Initially, data for the first stage were generated and then based on the decision at the first stage, data for the second stage were generated. The time-to-event data for the first stage were simulated for the number of patients denoted, $\mathrm{n}_{i1}^{\mathrm{obs}}$, where $i \in (1, 2, 3)$ denotes the treatment arm. Patients were equally allocated to one of the three treatment arms. It was assumed that the patients' survival followed an exponential distribution (see section 1.3) with hazard rate $\lambda_i = -\frac{\log \mathrm{S}_i(\mathrm{t})}{\mathrm{t}_i}$. For $i = 1, 2$, the hazard rate was generated based on the control arm survival rate and for $i = 3$, the hazard rate was based on the alternative hypothesis. It was assumed that patients were recruited at a uniform rate during the whole recruitment period. Therefore the time at which patients were recruited was generated using a uniform distribution from the start of the trial to the first interim stage as this occurred prior to the end of the recruitment period, i.e. start of trial < first interim stage < recruitment period < final stage. If the recruitment time plus the survival time was greater than the time of the first stage, patients were censored at the first interim

analysis stage. Pairwise comparisons between the control arm ($i = 1$) and the two experimental arms ($i = 2, 3$) were made using a Cox proportional hazards model. If the resulting HR for each pairwise comparison was less than the critical HR obtained from the output of the 'nstage' command, then the experimental arm continued to the next stage. If both experimental arms resulted in hazard ratios greater than the critical HR, then the trial was stopped, and inferiority of both experimental arms was declared.

If at least one experimental arm continued to the second stage, time-to-event data was again simulated using an exponential distribution for the number of patients required in the second stage denoted, $\mathrm{n}_{i2}^{\mathrm{obs}}$, where $i \in$ (1,2,3), (1,2) or (1,3) depending on the results of the first stage. Recruitment time was generated using a uniform distribution from the end of the first stage to the end of the recruitment period. The data simulated for the second stage were appended to the data generated in the first stage ($n^{obs} = \mathrm{n}_{i1}^{\mathrm{obs}} + \mathrm{n}_{i2}^{\mathrm{obs}}$) with those experimental arms that continued to the next stage. If the recruitment time plus the survival time was greater than the time of the final stage, then patients were censored. Pairwise comparisons were performed between the control and the experimental arm(s) that continued to the second stage. If the hazard ratio(s) were less than the critical HR at the final stage, then NI was declared.

The type I error and power at stage one was calculated by the proportion of simulations that had a HR less than the critical HR at stage one when comparing treatment one against treatment two and treatment one against treatment three respectively. Similarly, the PWER and pairwise power were calculated by the proportion of simulations that had a HR less than the critical HR at the final stage when the comparing treatment one against treatment two and treatment one against treatment three respectively. It was hypothesized that with 10,000 simulations, the 95% CI of the estimated rates should be within the true type I error, power, PWER and pairwise power (true values were obtained from 'nstage' output).

The number of events in the control arm at stage one and the final stage were stored and the average number of events were compared to the 'nstage' output. Histograms

were plotted for the hazard ratios for each pairwise comparison to assess normality. Data were simulated 10,000 times using Stata 16 with input seed '130933' and the code used to implement these simulations can be found in appendix 8.

6.4.2.1 Worked example

The steps performed to execute scenario one of the simulations using the DEVA_NI three arm two stage trial is shown below using the results from Table 6.3:

1. Data were simulated for 4995 patients; patients were equally allocated to one of three treatment arms using a uniform distribution.
2. Time-to-event data were simulated using an exponential distribution for each treatment arm with the following lambda values, $\lambda_1 = \lambda_2 = 0.0402$ and $\lambda_3 = 0.0476$ so that the properties under the null and alternative hypothesis could be assessed.
3. Part of the recruitment period was generated using a uniform distribution beginning at zero to 5.9 years (time of first stage).
4. If recruitment time plus the survival time was greater than 5.9 years then patients were censored using an indicator variable (0 = censored, 1 = event).
5. Pairwise comparisons were performed using Cox proportional hazards model between the control arm ($i = 1$) and each experimental arm ($i = 2, 3$). If the resulting hazard ratios was less than 1.188 then the respective experimental arm continued onto the next stage, this resulted in one of four possibilities:
    i. Both experimental arms continued
    ii. Only experimental arm two continued
    iii. Only experimental arm three continued
    iv. Confirmed inferiority of all experimental arms resulting in trial stoppage

6. Data were simulated for the remaining 1765 patients (Total sample size minus patients recruited in first stage), patients were allocated equally to the remaining treatment arms using a uniform distribution.
7. The remaining recruitment period was generated using uniform distribution beginning at 5.9 to eight years (end of recruitment period)
8. Data were appended to the data generated in the first stage, for which there were two possible outcomes:
    i. Both experimental arms continued to the next stage and if the recruitment time plus survival time was greater than 13.527, patients were censored using an indicator variable.
    ii. If only one experimental arm continued to the next stage and if the recruitment time plus survival time was greater than 12.586, patients were censored using an indicator variable.
9. Pairwise comparisons were performed between the control and experimental arm(s) that continued to the second stage. If the resulting hazard ratio(s) were less than 1.07 then NI was declared.
10. The events in each arm at the interim and final stage were stored for each simulation alongside the hazard ratios at each stage for each pairwise comparison.

The simulations were used to calculate the type I error and power at stage one and the PWER and pairwise power at the final stage. The proportion of simulations at stage one for the pairwise comparison between treatment arm 1 vs 2 and treatment arm 1 vs 3 that had a HR less than 1.188 provided the type I error and power respectively at this stage. The proportion of simulations at the final stage for the pairwise comparison between treatment arm 1 vs 2 and treatment arm 1 vs 3 that

had a HR less than 1.07 provided the PWER and pairwise power respectively at this stage.

6.4.3 Results

Table 6.4 shows the results obtained from the simulations of the type I error, power, PWER, pairwise power and the average number of control arm events. It can be clearly seen for the three arm two stage DEVA_NI trial, that the type I error rate, PWER, power and pairwise power of the true rates were contained within the 95% CI of the simulated values. When comparing the mean number of control arm events for the 10,000 simulations at each stage for the DEVA_NI at each scenario, majority have the same control arm events as obtained from the 'nstage' output. The average final stage control arm events obtained through the simulations for scenario one required an additional five events and scenario two showed an additional four more events required compared to the 'nstage' output which most probably is due to rounding error.

Figure 6.4 shows a plot of the hazard ratios obtained from the pairwise comparisons between the control arm (treatment one) and treatment two or treatment three for the DEVA_NI trial. As expected, the hazard ratios for the pairwise comparison between treatment one and treatment two are normally distributed around the null hypothesis (HR = 1). Similarly, the hazard ratios for the pairwise comparison between treatment one and treatment three were normally distributed around the alternative hypothesis (HR = 1.188).

Similarly, the QUASAR 2_NI trial showed that the type I error rate, PWER, power and pairwise power of the true rates were contained within the 95% CI of the simulated values for majority of the scenarios (Table 6.4). However, the true PWER and pairwise power for scenario two were somewhat greater than the upper boundary of the 95% CI of the simulated values.

A plot of the hazard ratios obtained from the pairwise comparisons between the control arm (treatment one) and treatment two or treatment three for the QUASAR 2_NI trial can be seen in Figure 6.5. The hazard ratios for the pairwise comparison between treatment one and treatment two were normally distributed just below the

null hypothesis (HR = 1) for scenario one and two which is also reflected by the PWER estimates obtained in Table 6.4. The hazard ratios for the pairwise comparison between treatment one and treatment three were normally distributed around the alternative hypothesis (HR = 1.305).

To ensure the correctness of these simulations, similar simulations were performed using a superiority hypothesis (as intended by the 'nstage' command) for the DEVA_SUP trial using the parameters from chapter five (Appendix 9).

Trial	Scenario	Stage	Nstage Type I error	Simulated Type I error (95% CI)	Nstage Power	Simulated Power (95% CI)	Nstage Control arm events	No. Control arm events
DEVA_NI	1	1	0.500	0.499 (0.489 – 0.509)	95.0%	94.6% (94.2% - 95.0%)	183	183
		Final (PWER/Pairwise power)*	0.023	0.021 (0.018 – 0.024)	87.0%	86.6% (85.9% - 87.3%)	710	715
	2	1	0.050	0.507 (0.497 – 0.517)	90.0%	89.6% (89.0% - 90.2%)	111	111
		Final (PWER/Pairwise power)*	0.042	0.038 (0.034 – 0.042)	83.0%	82.3% (81.6% - 83.0%)	579	583
	3	1	0.250	0.247 (0.239 – 0.255)	90.0%	90.0% (89.4% - 90.6%)	259	259
		Final (PWER/Pairwise power)*	0.044	0.042 (0.038 – 0.046)	81.3%	81.4% (80.6% - 82.2%)	486	485
QUASAR 2_NI	1	1	0.500	0.501 (0.491 – 0.511)	95.0%	94.9% (94.5% - 95.3%)	77	77
		Final (PWER/Pairwise power)*	0.023	0.021 (0.018 – 0.024)	87.1%	86.7% (86.0% - 87.4%)	298	297
	2	1	0.500	0.499 (0.489 – 0.509)	90.0%	89.6% (89.0% - 90.2%)	47	47
		Final (PWER/Pairwise power)*	0.042	0.036 (0.032 – 0.040)	83.2%	82.0% (81.2% - 82.8%)	243	240
	3	1	0.250	0.240 (0.232 – 0.248)	90.0%	90.1% (89.5% - 90.7%)	109	109
		Final (PWER/Pairwise power)*	0.044	0.040 (0.036 – 0.044)	81.5%	81.7% (80.9% - 82.5%)	204	202

*Showing final pairwise error rate and pairwise power for the nstage and simulated outputs.

Table 6.4: Results of the simulations performed for the DEVA_NI and QUASAR 2_NI three arm two stage compared to the true values and the 95% CI given by the 'nstage' output. PWER = Pairwise error rate.

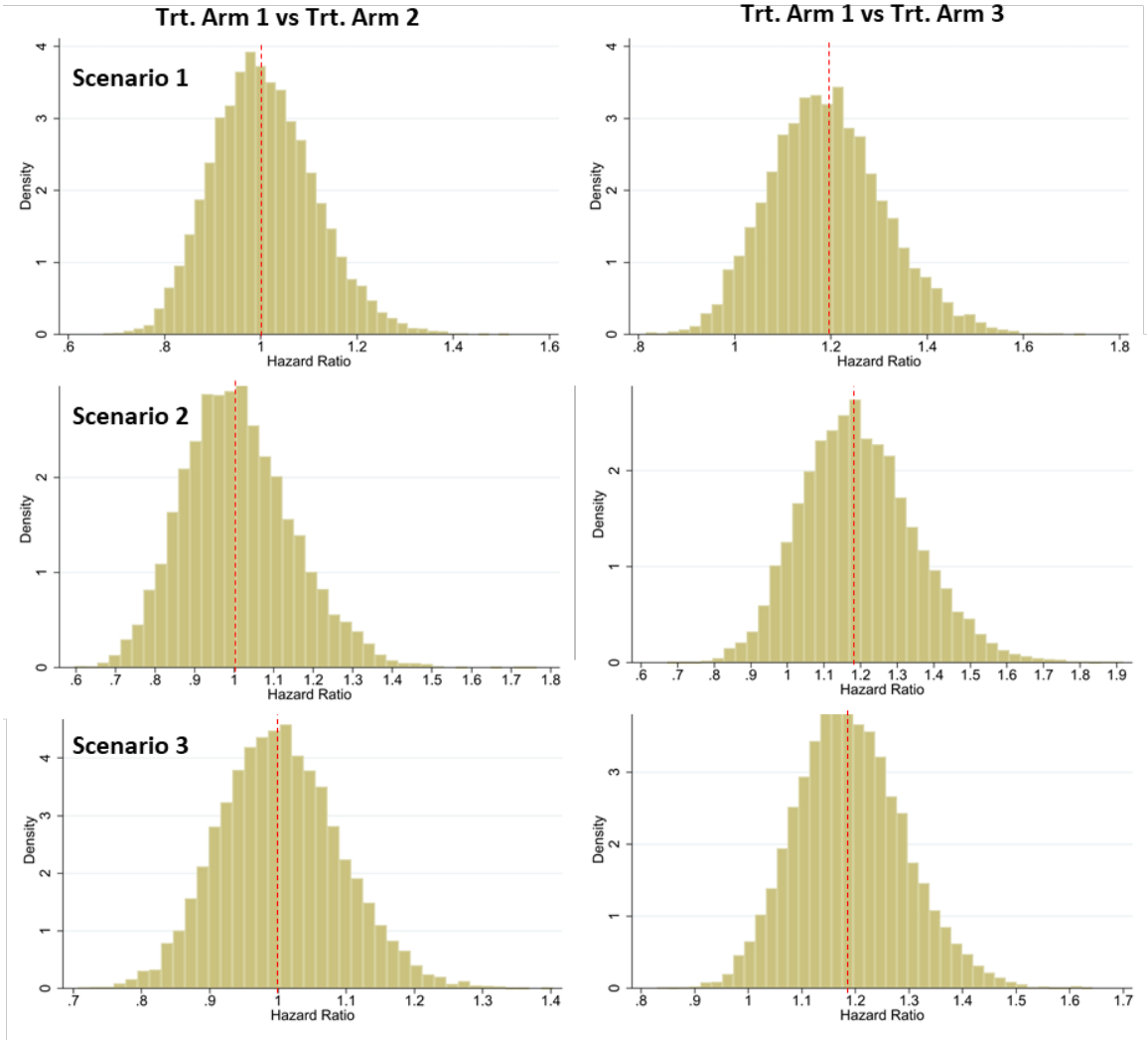

Figure 6.4: Plot of the hazard ratios from the simulations for the pairwise comparisons[1] for each scenario for the DEVA_NI three arm two stage trial.

[1] Data generated for treatment arm two using same survival rate as the control arm, data generated for treatment arm three using the alternative hypothesis.

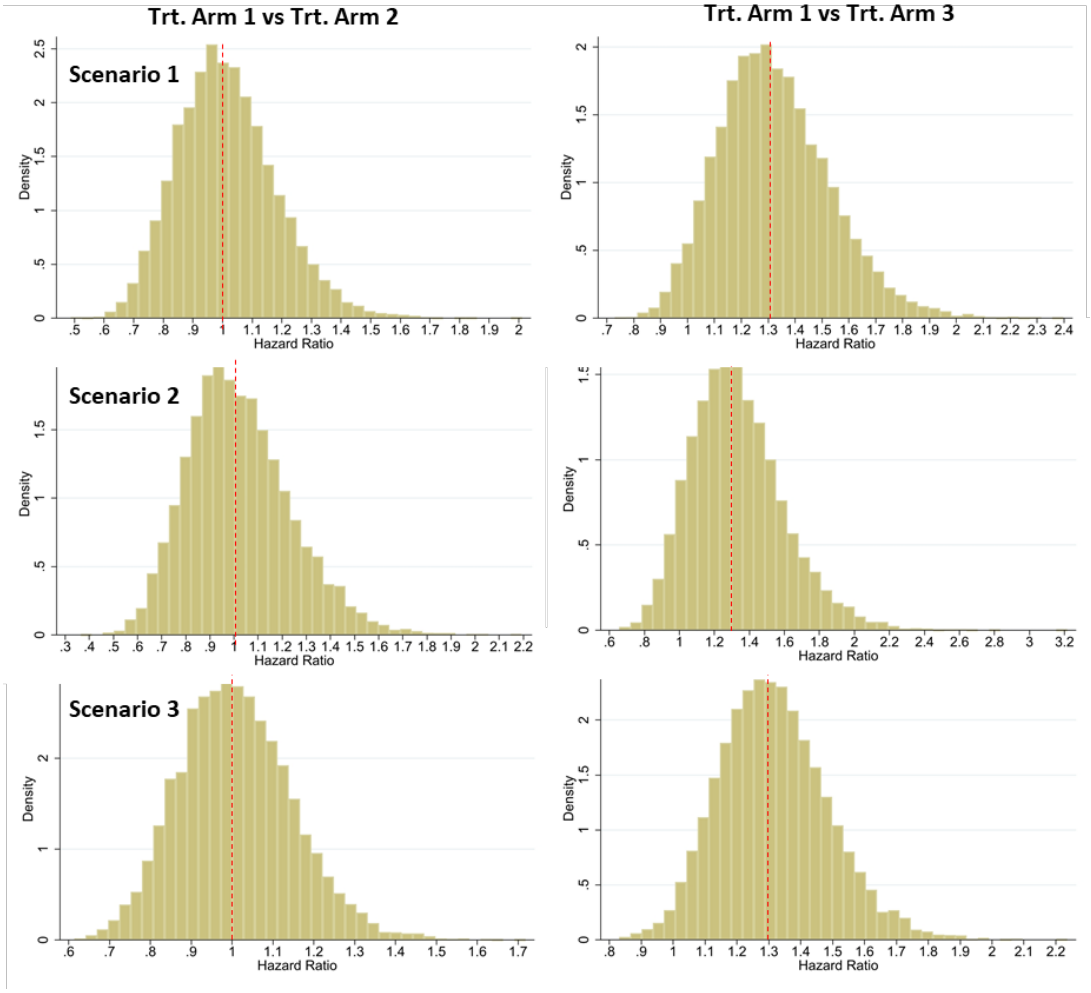

Figure 6.5: Plot of the hazard ratios from the simulations for the pairwise comparisons[1] for each scenario for the QUASAR_NI three arm two stage trial.

6.5 Third phase calculations extended to implement NI hypothesis

In chapter five, the feasibility of applying a MAMS design with long-term outcomes using a superiority hypothesis were evaluated. The third phase of calculations that were applied in chapter five were extended within this chapter to evaluate the feasibility of applying MAMS designs using a NI hypothesis.

6.5.1 Methods

For the DEVA_NI and the QUASAR 2_NI trials, sample sizes for a two arm trial were re-calculated with an absolute NI margin of 3%, the obtained OS results for the control arm, power of 90% and one-sided alpha of 2.5%. The two trials were re-

[1] Data generated for treatment arm two using same survival rate as the control arm, data generated for treatment arm three using the alternative hypothesis.

named with the suffix "_NI" respectively to emphasise that a NI hypothesis has been implemented and the new sample size and number of control arm events can be seen in Table 6.5.

Trial Name	OS (Target HR)	SS (Control SS)	Control events
Deva_NI	5 year 81.8% (1.188)	4264 (2132)	713
QUASAR 2_NI	3 year 89.4% (1.305)	2136 (1068)	301

Table 6.5: Re-calculated sample sizes using OS results with a NI hypothesis for a two arm trial. OS = Overall Survival; HR = Hazard ratio; SS = Sample size.

Thereafter, fictional scenarios where multiple treatments were available that could be no worse than the standard treatment (NI hypothesis) were created and sample sizes calculated for MAMS designs using the 'nstage' command. The stagewise significance and power level as defined by Royston et al (2011) and implemented in the STAMPEDE trial (Table 4.2) were used and it was assumed that all experimental arms would reach the final stage.

Like chapter five, the hypothetical MAMS design for the trials were classed as feasible if the number of control arm patients and events required in the MAMS designs were broadly similar to the two-arm trial. The sample size required for six separate standard two arm trials versus two separate four arm three stage trials versus three separate three arm two stage trials was compared to see if fewer patients were required using the MAMS design if fixing for time.

It was anticipated that more patients would be required using a NI hypothesis compared to the number of patients using a superiority hypothesis (section 5.5) as a smaller treatment difference was used (3% absolute NI margin). Therefore, the number of patients required per month was compared to the estimated maximum number of patients (Table 5.5) but for exploratory purposes. This estimation was calculated based on patients within the UK, however international recruitment could be considered to accommodate extra patients.

Lastly, the trials were assessed taking the work of Schiavone et al (2019) and Hague et al (2019) as mentioned in chapter two into consideration to ensure the practicality aspect of employing a MAMS design (Hague et al. 2019, Schiavone et al. 2019). Based on the total trial length (recruitment and follow-up) of both the DEVA and QUASAR 2

trials, it was anticipated that executing a five arm four stage trial would be infeasible. This was still explored in case the total trial length reduced as seen in chapter five.

6.5.2 Results

The results of the sample size calculations which were used for the third phase can be seen in Table 6.6. Figure 6.6 shows the number of control arm patients and control arm events required for the standard two-arm trial and the MAMS designs for the DEVA_NI and QUASAR 2_NI trials. The number of control arm patients for both these trials were similar for the standard two-arm trial and the three arm two stage MAMS designs. The number of control arm patients increased for both trials for the four arm three stage and five arm four stage because the length of the trial was reduced resulting in more patients being required to obtain the required number of events. The total number of control arm events required for the DEVA_NI and QUASAR 2_NI trial were the same for all trial designs.

Trial	Results	3A2S		4A3S			5A4S			
		Stage		Stage			Stage			
		1	2	1	2	3	1	2	3	4
DEVA_NI	PWER		0.023			0.021				0.021
	Pairwise Power		0.87			0.86				0.85
	Critical HR	1.188	1.070	1.188	1.13	1.07	1.188	1.13	1.107	1.07
	Time	5.9	7.6	5.6	8.0	12.3	4.6	6.6	8.0	9.5
	Sample Size	4995	6760	7043	10087	10120	10505	15005	18236	18240
	CA Patients	1665	2253	1761	2522	2530	2101	3001	3647	3648
	CA Events	183	710	183	364	710	183	364	528	710
QUASAR 2_NI	PWER		0.023			0.021				0.021
	Pairwise Power		0.87			0.86				0.85
	Critical HR	1.305	1.111	1.305	1.208	1.111	1.305	1.208	1.169	1.111
	Time	4.8	11.6	4.6	6.5	9.9	4.3	6.1	7.3	8.7
	Sample Size	2755	3185	3828	5459	5460	5105	7276	8811	9000
	CA Patients	918	1062	957	1365	1365	1021	1455	1762	1800
	CA Events	77	298	77	153	298	77	153	221	298

Table 6.6: Results from the sample size calculation for the three arm two stage (3A2S), four arm three stage (4A3S) and five arm four stage (5A4S) for the DEVA_NI and QUASAR 2_NI trials using the 'nstage' command. PWER = Pairwise Error Rate; HR = Hazard Ratio; CA = Control arm.

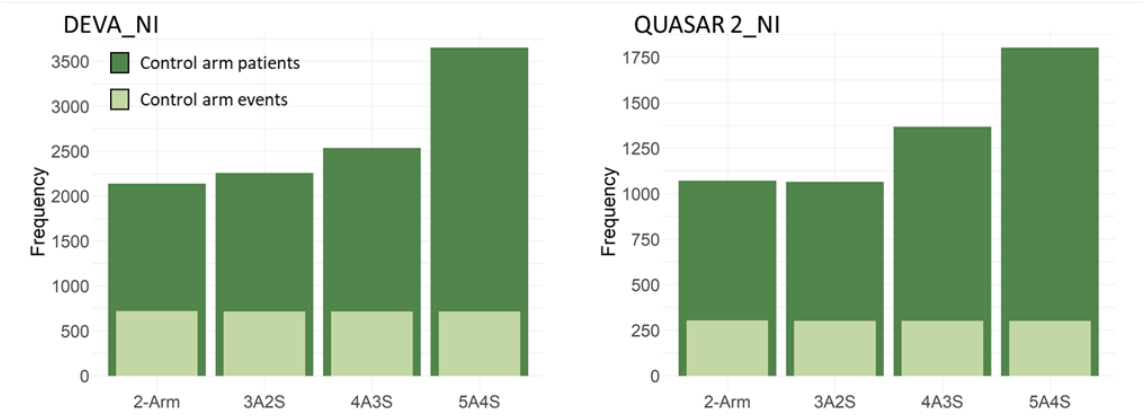

Figure 6.6: Number of control arm patients (dark green) and events (light green) required for DEVA_NI and QUASAR 2_NI 2-arm trial and MAMS trials.

From the sample size calculations it can be seen that if in the situation where six drugs were available for testing for NI against a common control, then performing either three separate three arm two stage or two separate four arm three stage MAMS trials resulted in requiring fewer patients then conducting six separate standard two-arm trials (Figure 6.7). Furthermore, if there was more than one experimental arm to consider, MAMS designs required fewer patients than standard two-arm trials therefore proving that MAMS designs can be more efficient. For example, for the DEVA_NI trial, performing a three arm two stage trial would require 6760 patients compared to 8528 patients required for two separate two-arm trials.

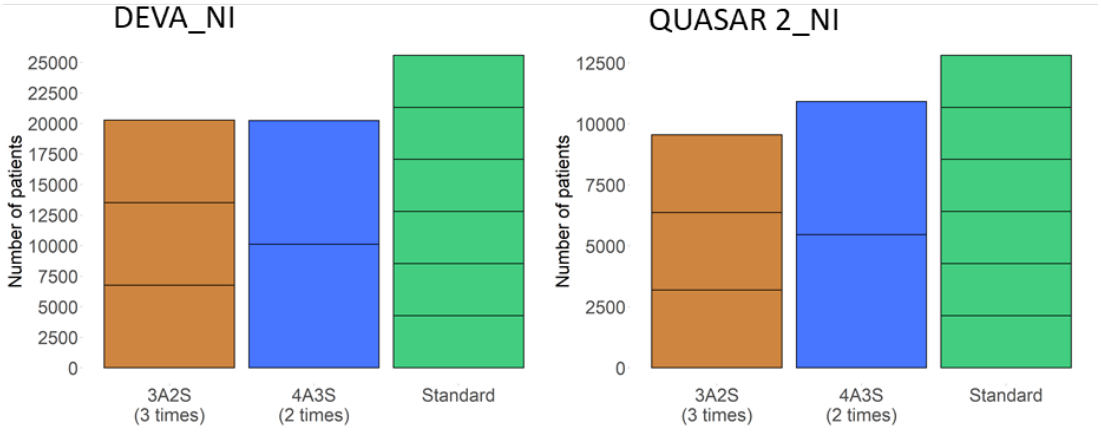

Figure 6.7: Comparing sample sizes for three separate three arm two stage versus two separate four arm two stage versus six standard parallel designs for the DEVA_NI and QUASAR 2_NI trials.
3A2S = Three arm two stage; 4A3S = Four arm three stage

Figure 6.8 shows the number of patients that would be required per month for each of the DEVA_NI and QUASAR 2_NI trial compared to the estimated maximum number of patients that could be recruited. For all the DEVA_NI trials, more patients would be required than the estimated maximum number of patients. The five arm four stage MAMS DEVA_NI trial required 190 patients per month to be recruited to reach the sample size compared to a maximum of 58 patients that could be recruited per month. The three arm two stage and four arm three stage QUASAR 2_NI trials required fewer patients compared to the estimated maximum number of patients.

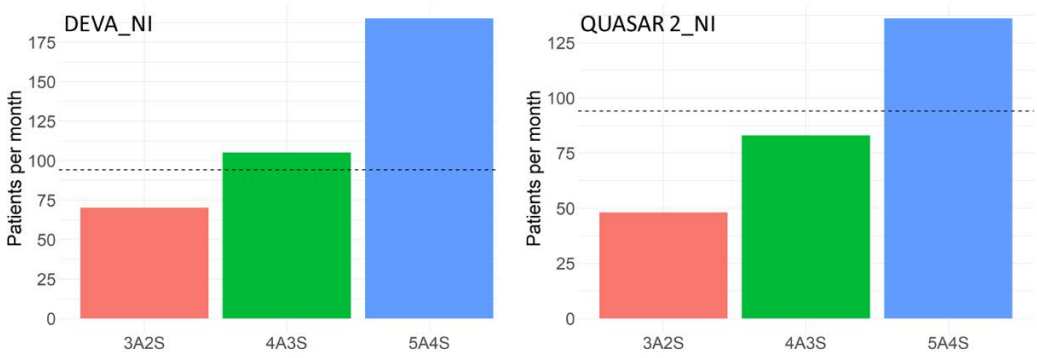

Figure 6.8: Total number of patients required assuming that all experimental arms reach the final stage for a three arm two stage, four arm three stage and five arm four stage MAMS design when applying the parameters obtained from the different breast cancer trials. Absolute line showing the estimated maximum number of patients that could be recruited. 3A2S = Three arm two stage; 4A3S = Four arm three stage; 5A4S = Five arm four stage.

6.6 Discussion

The outcome of the simulation studies showed that the 'nstage' command can be used with a NI hypothesis. The statistical properties were validated using simulations and were similar to the values inputted and produced by the 'nstage' command. The true values for the PWER and pairwise power for scenario two for the QUASAR 2_NI trial were greater than the upper boundary of the 95% CI of the simulated values. This may be due to the slight lack of normality, which was seen in the histogram plots (Figure 6.5), however as the number of simulations tend to infinity, the estimated PWER and pairwise power will tend to the true value. A limitation to these simulations was that only the properties for a three arm two stage MAMS design with the same intermediate and primary outcome were assessed. Hence, further work

could be to perform simulations to validate the statistical properties for when the intermediate and primary outcome differ.

The results from the hypothetical MAMS designs using a NI hypothesis showed that based on the estimated maximum number of patients, MAMS design using up to three experimental arms were feasible. Anything beyond this would require more patients per month compared to the estimated maximum number of patients, which can always be overcome by increasing the recruitment period to allow for a lower rate of patients recruited per month. Additionally, international recruitment could be considered as the estimated maximum number of patients recruited per month was based on figures in the UK only. The results showed that if in the situation where there are six drugs available for testing for NI against a common control then a MAMS design would result in requiring fewer patients compared to conducting six separate two arm trials. For example, if there were six different statin drugs, a MAMS design using a NI hypothesis could be implemented to evaluate these and focus placed on controlling the PWER. However, if the drugs were the same and perhaps assessed different dosing requirements, then emphasis would be placed on controlling the FWER. More investigation into the use of different durations and the need to control the FWER is considered in the next chapter.

One of the rules of the 'nstage' command is that all interim stages must take place prior to the end of the recruitment period (more information provided in chapter two and chapter five), therefore as the number of stages increased and potentially the number of experimental also arms increased, more patients were required to ensure that all interim stages met the required number of control arm events within the pre-specified recruitment period, which led to a decrease in the amount of follow-up required.

When comparing the total sample size of various MAMS trials against repeating two-arm trials, it was clear that MAMS designs required fewer patients because all experimental arms share a common control.

6.7 Conclusion

With the rise of trials implementing a NI hypothesis (section 1.4.3), being able to assess multiple arms across multiple stages instead of executing multiple two-arm trials in this setting can be advantageous and cost effective. In comparison to a superiority setting, NI trials generally require more patients due to smaller margins that are often used. However, the long term benefits such as improving the standard of care by reducing the toxicities, potential administration work as well cost effectiveness can put NI trials in a better limelight with trial stakeholders and patients.

This chapter showed that MAMS designs can be implemented using the 'nstage' command with a NI hypothesis. The understanding developed within this chapter will be used to apply a MAMS design to assess different durations of treatment. Furthermore, MAMS designs may statistically show promise however the operational aspects of running a MAMS design should not be underestimated as seen for the STAMPEDE and FOCUS4 trials. The HERA, FinHer, PHARE, HORG, SOLD and Persephone trials (Piccart-Gebhart et al. 2005, Joensuu et al. 2009, Pivot et al. 2013, Mavroudis et al. 2015, Earl et al. 2019) all assessed the duration of Herceptin in breast cancer patients. The next chapter will assess whether a MAMS design would have been suitable rather than performing these trials separately and what operational aspects would have had to be taken into consideration to successfully execute this hypothetical trial.

7 The application of a MAMS design to test different durations of Herceptin treatment.

7.1 Introduction

In the previous chapters MAMS designs were applied to different cancer sites and it was concluded that these designs are more applicable when there is a high event rate but were still useful for lower event rates such as breast cancer studies. MAMS designs work well due to their efficacy and effectiveness in looking at multiple treatments simultaneously but must have adequate event rates to give power to the comparisons.

The focus of this chapter is on applying a MAMS design with a NI design to assess the optimal duration of Herceptin therapy for patients diagnosed with early breast cancer. The motivation behind this research was due to the Persephone trial which was conducted at Warwick CTU and published primary analysis results in 2019 (Earl et al. 2019). The Persephone trial compared six months of Herceptin therapy against 12 months of Herceptin therapy in patients diagnosed with early breast cancer (Earl et al. 2019). Alongside this trial, there were other international trials being conducted (FinHer trial, SOLD trial, HORG trial and PHARE trial) assessing different durations of Herceptin therapy in the same patient populations. Thus, the aim of this research was to investigate whether a MAMS design would have been more effective instead of many separate trials to assess the optimal Herceptin duration and explore some of the operational aspects needed to execute this hypothetical adaptive trial.

7.2 Existing trials testing the duration of Herceptin therapy

Descriptions of the HERA, FinHer, SOLD, HORG, PHARE and Persephone trials, all of which assessed different durations of Herceptin therapy in patients diagnosed with breast cancer are provided in the following sections. The patients recruited to each of the arms and the duration of Herceptin can be seen in Figure 7.1. A summary of the primary outcome results showing the obtained HR, the 95% CI and the relative margin for each trial can be seen in Figure 7.2. A summary of recruitment and the outcome of the trial can be seen in Table 7.1. These figures and table will be frequently referenced throughout section 7.2.

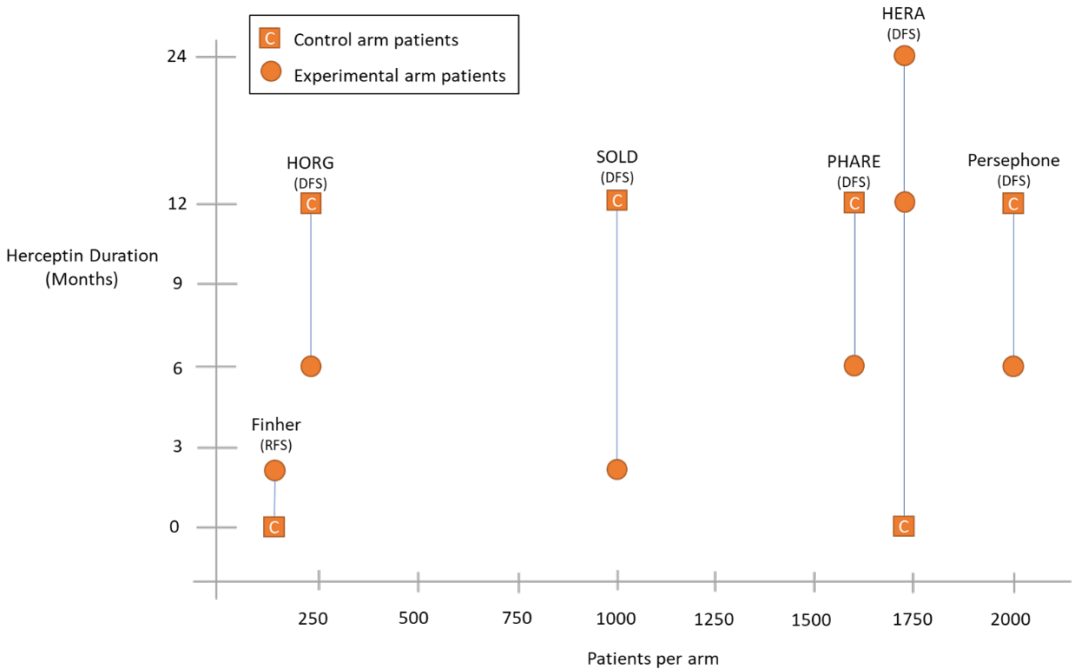

Figure 7.1 Patients recruited for each treatment arm for the HERA, FinHer, SOLD, HORG, PHARE and Persephone trials. RFS = Recurrence-free survival. DFS = Disease-free survival

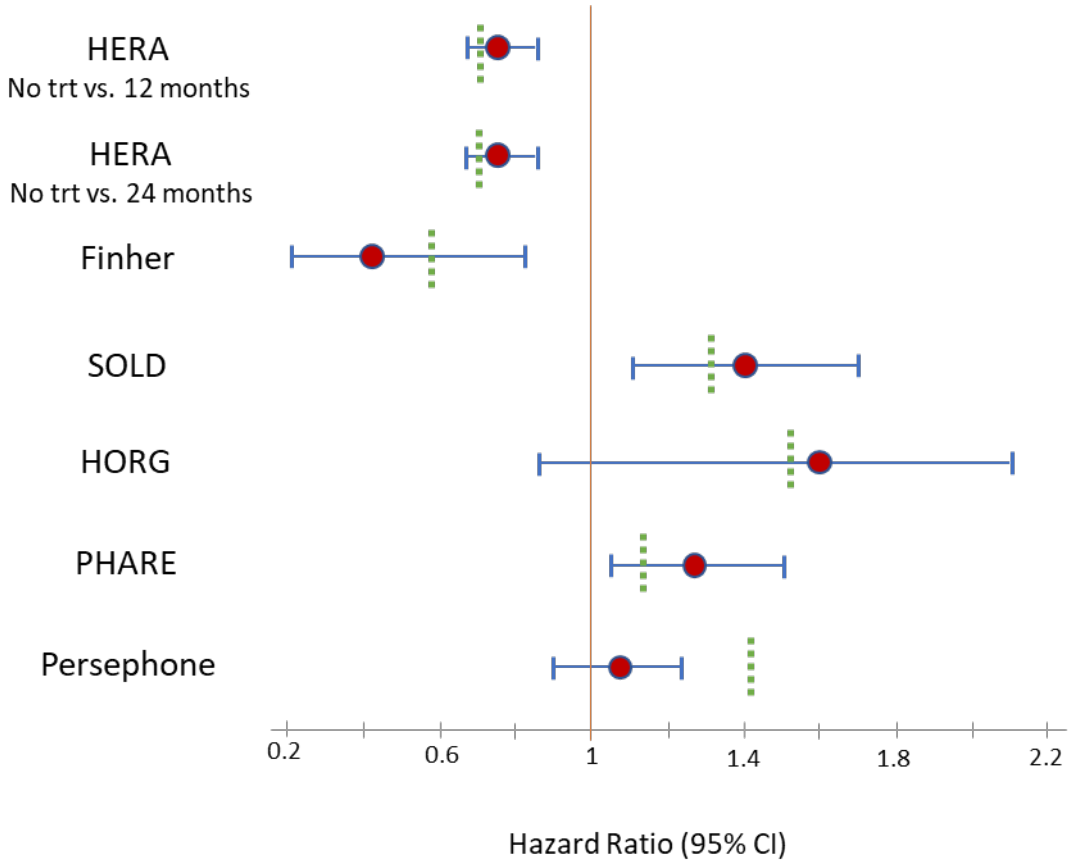

Figure 7.2: Primary outcome results showing the HR and 95% CI obtained and the planned relative margin (shown in green) for the HERA, FinHer, SOLD, HORG, PHARE and Persephone trials. HR = Hazard Ratio; CI = Confidence Interval

Trial information	HERA	FinHer*	SOLD	HORG	PHARE	Persephone
Treatment arms	No trt (control) vs 12 mths vs 24 mths	No trt (control) vs 9 wks	9 wks vs 12 mths (control)	6 mths vs 12 mths (control)	6 mths vs 12 mths (control)	6 mths vs 12 mths (control)
Total number of patients	5099	232	2176	481	3384	4088
Recruitment time (years)	3	3	7	8	4	8
Median follow-up	11 years	36 months	5.2 years	51 months & 47 months	42.5 months	5.4 years
TTE PO	DFS	RFS	DFS	DFS	DFS	DFS
Primary outcome results						
No treatment	63% - 10 yr DFS	77.6% - 3 yr RFS	-	-	-	-
9 weeks	-	89.3% - 3 yr RFS	88% - 5 yr DFS	-	-	-
6 months	-	-	-	93.3% - 3 yr DFS	91.2% - 2 yr DFS	89.4% - 4 year DFS
12 months	69% - 10 yr DFS	-	90.5% - 5 yr DFS	95.7% - 3 yr DFS	93.8% - 2 yr DFS	89.8% - 4 year DFS
24 months	69% - 10 yr DFS	-	-	-	-	-
Hazard Ratio (95% CI)	0.76 (0.68 to 0.86) No trt vs 12 mths	0.42 (0.21 – 0.83) No trt vs 9 wks	1.39 (90% CI: 1.12-1.72) 9 wks vs 12 mths	1.58 (0.86-2.10) 12 mths vs 6 mths	1.28 (1.05-1.56) 6 mths vs 12 mths	1.07 (90% CI: 0.93-1.24) 6 mths vs 12 mths
	0.77 (0.69-0.87) No trt vs 24 mths	-	-	-	-	-
	1.02(0.89-1.17) 12 mths vs 24 mths	-	-	-	-	-

*Only subgroup treated with Herceptin

Table 7.1: Patients recruited, time taken to recruit, median follow-up time, and primary outcome results for the HERA, FinHer, SOLD, HORG, PHARE and Persephone trials. Latest follow-up paper. TTE – Time to event. DFS – Disease-free survival; RFS – Recurrence free survival

7.2.1 HERA trial

The HERA trial, as mentioned in chapter five, was a three-arm phase III RCT comparing Herceptin for one year, Herceptin for two years or observation (control group) for patients with HER2 positive early stage breast cancer (Piccart-Gebhart et al. 2005).

At the time of designing the HERA study, the impact of Herceptin therapy in patients with early breast cancer had not yet been established. Hence, a superiority primary endpoint was used to compare DFS for patients randomised to one year of Herceptin versus no Herceptin and two years of Herceptin versus no Herceptin. A superiority difference of 6.8% was used to calculate the sample size. Pairwise comparisons were made using the unstratified log-rank test with the addition of the Holm-Bonferroni method which ensured that the FWER was 0.05 (Holm 1979). A target sample size of 4482 patients was required to reach a target number of 951 DFS events for all treatment arms to achieve 80% power to detect at least a 6.8% absolute difference in DFS from the observation arm (estimated to be 65% for 5-year DFS) (Figure 7.3).

An additional exploratory analysis plan was implemented to compare the use of one year of Herceptin versus two years of Herceptin and to examine various TTE outcomes such as OS and RFS.

An interim analysis was implemented to compare DFS for patients randomised to one year Herceptin versus observation and two years of Herceptin versus observation, which took place after half of the estimated total number of DFS events were observed. At this interim analysis, the DMC suggested to release the results of the one-year of Herceptin as these results were highly statistically significant and showed great increase in DFS compared to the observation arm.

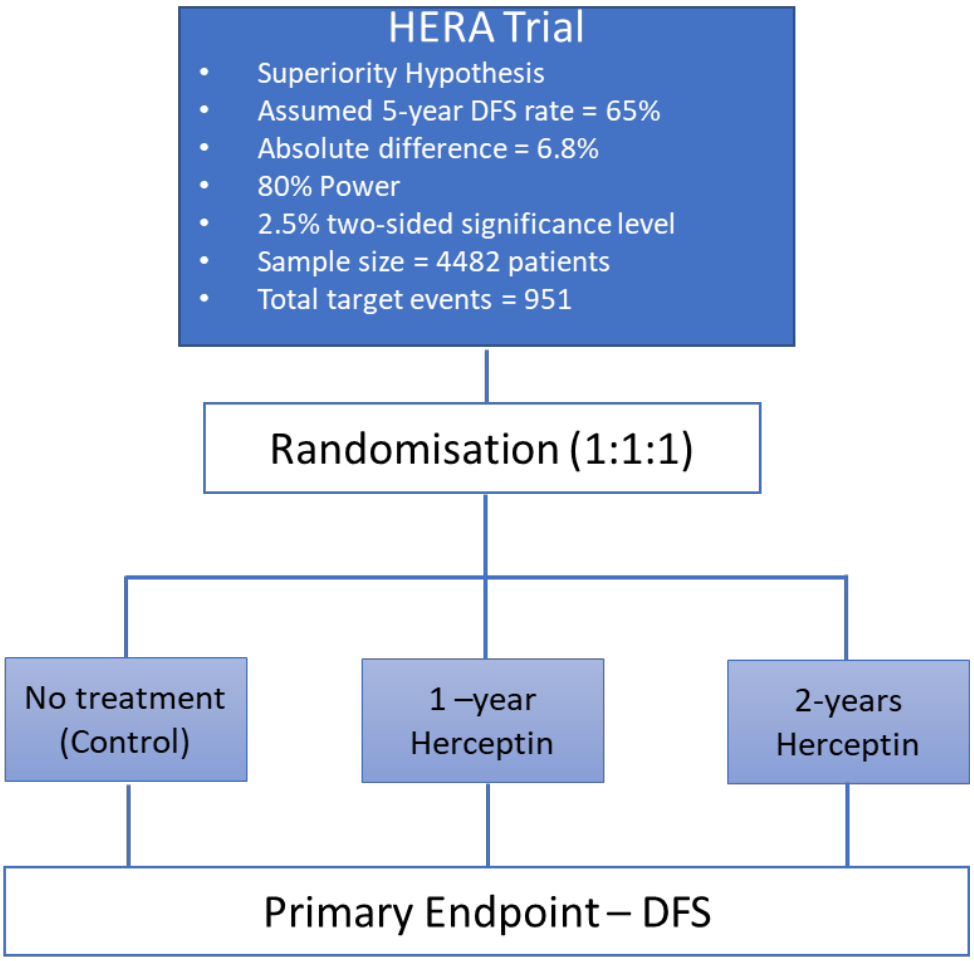

Figure 7.3: Outline of the HERA design parameters. DFS = Disease-free survival

To date, there have been various papers published at different follow-up time points (Piccart-Gebhart et al. 2005, Smith et al. 2007, Goldhirsch et al. 2013, Cameron et al. 2017). The first paper compared patients on the observation arm versus one year Herceptin and was published with at least one-year median follow-up. The results from this paper showed that there was a clinical and statistical significant difference in DFS (HR: 0.54, 95% CI: 0.43 to 0.67, p-value<0.001) for those patients that received Herceptin for one year versus those patients that received no Herceptin (Piccart-Gebhart et al. 2005). This led to a protocol amendment whereby patients in the observation group were invited to receive Herceptin. The latest paper, published in 2017 after patients had reached a median follow-up of 11 years, (Cameron et al. 2017) reported that the 10 year DFS was 69% for both the one year and two year arms suggesting that the study was powered adequately to show that these trial arms were the same or NI, however this was only exploratory analysis (Table 7.1).

The HERA trial proved to be a revolutionary trial, changing practice for women with HER2 positive breast cancer by showing that Herceptin therapy can significantly reduce the risk of breast cancer recurring (Gianni et al. 2011). This trial became the basis of future trials such as the FinHer trial, PHARE trial, Persephone trial and the SOLD trial, all of which assessed shorter durations of Herceptin therapy in women with HER2 positive early stage breast cancer (Persephone 2007, Joensuu et al. 2009, Pivot et al. 2013, Joensuu et al. 2018).

7.2.2 FinHer trial

The FinHer trial compared the use of Docetaxel plus a regimen consisting of fluorouracil, epirubicin and cyclophosphamide (FEC) against Vinorelbine plus FEC but involved a second randomisation for the HER2 positive patients to no Herceptin or nine weeks of Herceptin (Joensuu et al. 2009). This phase III RCT recruited women with axillary-node positive or high-risk node-negative cancer to each of these arms, of which it was estimated that 30% of patients would be HER2 positive. A sample of 300 HER2 positive patients were needed to be randomised to receive either no Herceptin or Herceptin for nine weeks to detect a difference in five year RFS from 50% with no Herceptin to 67% with a power of 80% and 5% two-sided level of significance, which required approximately 1000 patient to be recruited to full trial (Joensuu et al. 2018) (Figure 7.4).

The FinHer trial recruited 232 HER2 positive patients within three years. At the time of analysis, with a median follow-up of three years, there was a statistically significant difference in RFS between the no treatment and nine weeks of Herceptin treatment arms; with a three year RFS of 77.6% and 89.3% respectively (p-value = 0.01) (Table 7.1).

The trial recruited 232 HER2 positive patients out of a required 300 patients, which was a limitation to the study. The trial team concluded that their results indicate that a nine-week period is effective but the optimal duration of Herceptin therapy requires a further RCT (Joensuu et al. 2009). This trial resulted in the creation of the SOLD trial, which is discussed below.

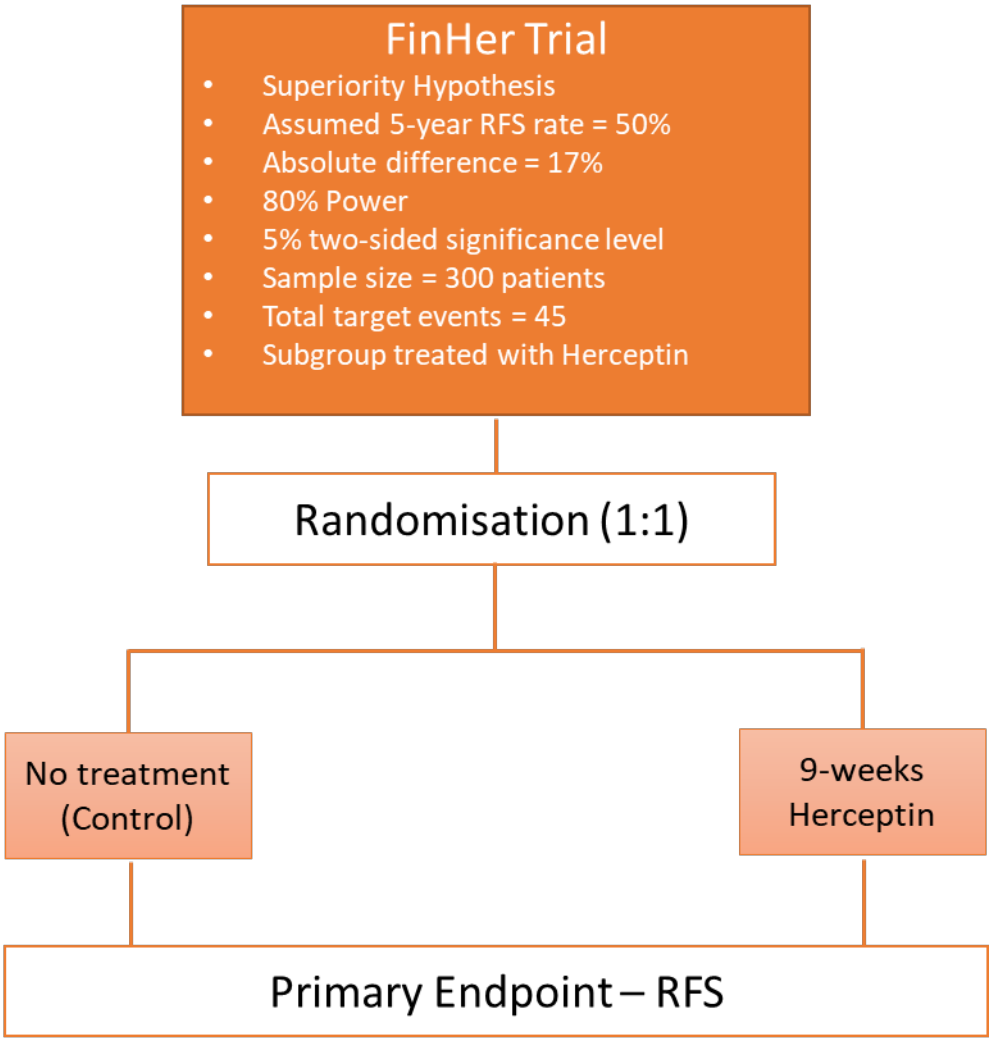

Figure 7.4: Outline of the FinHer design parameters. RFS = Recurrence-free survival

7.2.3 SOLD trial

The SOLD trial was a phase III RCT that compared the use of Herceptin with docetaxel over a nine week period against Herceptin with docetaxel over a 12 month period (current standard) in women with HER2 positive early stage breast cancer (Joensuu et al. 2018).

The initial sample size calculated assumed superiority between the treatment arms. A sample size of approximately 3000 patients with a power of 80% and two-sided significance level of 5% would be required to be recruited over four years to detect an improvement in OS after five years of follow-up from 80% to 84% in the better arm but they did not state which the better arm is.

The sample size calculations for this trial were revised due to two reasons: 1) External data suggesting higher values for five-year DFS and 2) Changing from a superiority assumption to a NI assumption as reducing treatment duration was not likely to reduce the number of breast cancer recurrences. Changing the hypothesis from a superiority to a NI hypothesis is a major change to the design. However, a NI hypothesis would have been more appropriate in the first place and all other worldwide studies looking at the duration of Herceptin were using a NI design, which may have been why the change was agreed. Hence the new sample size calculation required 2168 patients to be recruited in 7.5 years to reach a total of 366 events with a HR set at 1.3 (4% absolute NI margin) and the five-year DFS assumed to be 85% for the 12 month treatment arm as per the updated research with 80% power and a one-sided significance level of 5% (Figure 7.5). The SOLD trial did not consider anything less than a 4% absolute NI margin as it was not considered clinically significant and believed that the estimated 5-year DFS could be higher than 85%. For this reason, sample size calculations were estimated for 5-year DFS rates ranging from 84% to 88%. In re-designing the trial using a NI hypothesis, the trial team also updated the initial assumptions made when designing using a superiority hypothesis as the DFS rates were also updated and the time for recruitment was increased from four years to 7.5 years, as the enrolment rate of patients was slower than expected.

The SOLD trial team recruited 2174 patients and after a median of five years of follow-up, the five-year DFS was found to be 88% for the nine week arm and 90.5% for the 12 month treatment arm with a hazard ratio of 1.39 (90% CI: 1.12-1.72). The trial failed to show that nine weeks of Herceptin was non inferior to one year of Herceptin (Table 7.1).

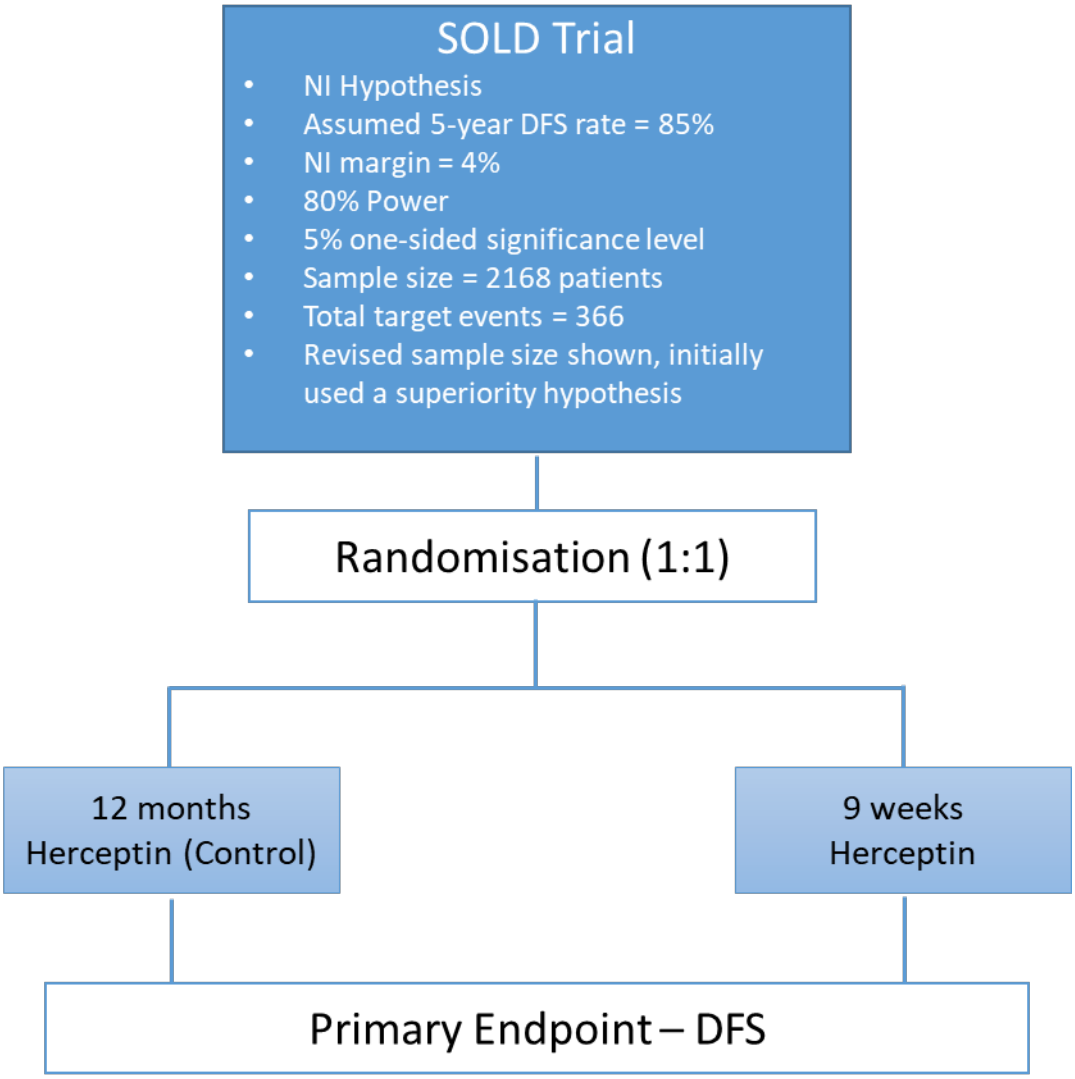

Figure 7.5: Outline of the SOLD design parameters. NI = Non-inferiority; DFS = Disease-free survival

7.2.4 HORG trial

The Hellenic Oncology Research Group (HORG) RCT, based in Greece, compared the use of Herceptin for six months against 12 months for women with HER2 positive early breast cancer (Mavroudis et al. 2015). The sample size calculated for this trial required a total of 478 patients to be accrued within three years to demonstrate NI using an 8% absolute NI margin for three-year DFS assuming an expected DFS in the 12 month group of 85% with 80% power and one-sided significance level of 5% (Figure 7.6). The HORG team reported that "the noninferiority hazard ratio margin of 1.53 was derived from an estimated absolute difference in 3-year DFS of 8%. To reject the null hypothesis and therefore to conclude noninferiority, the upper bound of the 95% confidence interval should be less than this margin" (Mavroudis et al. 2015)..

The trial struggled to recruit the estimated number of patients within three years and hence extended its recruitment to eight years. The median follow-up time of the 481 patients recruited was 51 months for the six month arm and 47 months for the 12 month arm. The three year DFS were 93.3% for six months and 95.7% for 12 months of Herceptin treatment (Table 7.1) with a HR=1.57 (95% CI: 0.86 to 2.10; p-value=0.137) (Mavroudis et al. 2015). The p-value showed that there was no statistically significant difference between the two groups however the HR and therefore the upper bound of the 95% CI for this trial were greater than the HR margin (1.53) and therefore could not conclude the NI of six months of Herceptin treatment. The reporting of the upper 95% CI was a limitation to this study as the team implemented a 5% significance level, they should have used the upper bound of a 90% CI. The study implemented a large NI margin, which reflected the relatively small number of patients required for this study. A larger NI margin may discourage patients to enrol as the NI margin reflects 'how much worse of a treatment' patients are willing to accept for a shorter duration of treatment. Another limitation to the study was that the original assumption for three-year DFS improved from 85% to 95%, therefore there were fewer events than originally predicted. The original critical HR was used to assess NI and therefore there was less power to detect this level as the obtained DFS meant a smaller NI margin was used. Therefore, by using an absolute margin and readjusting the critical HR accordingly, there would be more power as the critical HR is larger as the relative difference will be greater.

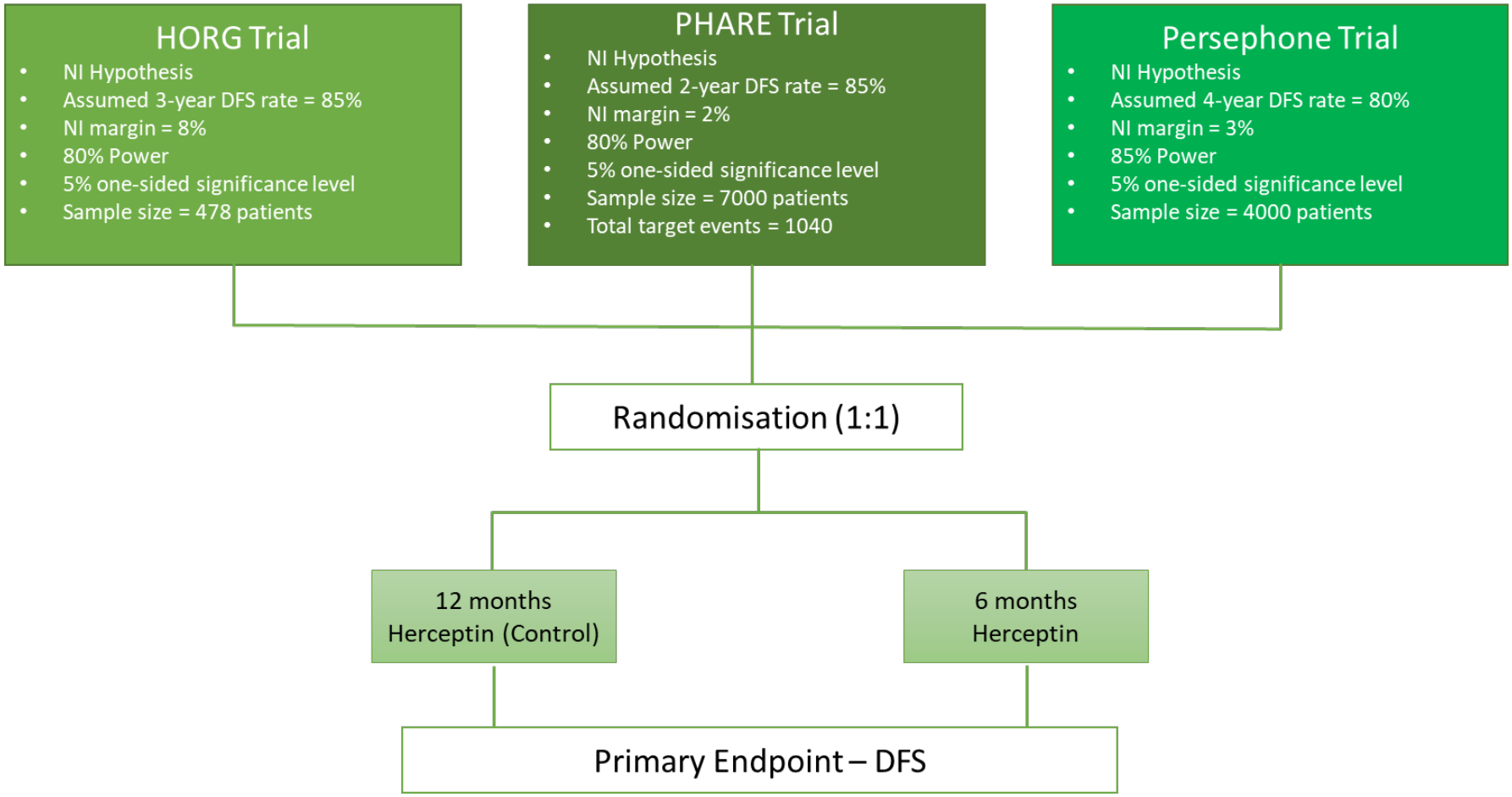

Figure 7.6: Outline of the HORG, PHARE and Persephone design parameters. NI = Non-inferiority; DFS = Disease-free survival

7.2.5 PHARE trial

The PHARE trial was the French phase III RCT similar to Persephone which compared the use of Herceptin for six months against 12 months in women with HER2 positive early breast cancer (Pivot et al. 2013, Pivot et al. 2019).

The initial sample size calculated required a total of 7000 patients to observe 1040 events to demonstrate a 2% absolute NI margin for DFS based on an expected two-year DFS of 85% for 12 months of Herceptin together with a one-sided 5% significance level and 80% power assuming two years recruitment followed by two years follow-up (Figure 7.6). The trial team reported that "a two-sided 95% CI with the upper bound with a HR less than 1.15 would conclude NI as the primary objective of the trial was to assess if six months of Herceptin was no worse than 12 months". As mentioned in chapter one, assessing the upper bound of a 95% CI would correspond to using a 2.5% one-sided significance level whereas the initial sample size was based on a 5% one-sided significance level. Thus instead, the PHARE trial should have concluded NI if the upper bound of the 90% CI was less than 1.15.

The PHARE trial team struggled to obtain the required accrual rate and hence the sample size was changed to a four year recruitment period with an analysis planned at eight years, which would require 3400 patients instead to achieve the same number of events. Annual interim efficacy analyses were performed using the Haybittle-Peto method whereby the treatment would be stopped if there was a statistically significant difference of $p<0.001$ for efficacy (Pocock 2005, Pivot et al. 2013). A stop-go criteria was enforced whereby the trial would be stopped if the HERA study results showed that two years of Herceptin was superior to one year of Herceptin.

The PHARE trial recruited 3384 patients in four years and had a median of 42.5 months of follow-up instead four years follow-up. The original sample size calculated that 1040 DFS events would be required however the actual trial accumulated 384 DFS events as the DFS rate was considerably higher than originally expected, which could impact the overall power of the study as considered in the previous section. The 2-year DFS rate was 93.8% and 91.1% in the six month and 12 month arm

respectively (Table 7.1) giving a HR of 1.28 (95% CI: 1.05 to 1.56; p-value=0.29). After a median of 3.5 years of follow-up, the PHARE trial team reported that they could not show that six months of Herceptin was non-inferior to 12 months of Herceptin as the upper confidence interval of the HR exceeded 1.15 (Pivot et al. 2019).

7.2.6 Persephone trial

The Persephone trial, similar to the PHARE trial, was a RCT comparing the use of Herceptin for six months versus 12 months in a phase III NI trial in HER2 positive patients diagnosed with early breast cancer (Earl et al. 2019). 4000 patients were required to evaluate an absolute NI margin of 3% for DFS for six months Herceptin versus 12 month Herceptin with 85% power and 5% one-sided significance level (Figure 7.6).

The Persephone trial ran parallel to the PHARE trial, which therefore influenced the oversight of the trial as new evidence from PHARE emerged. Three interim analyses were performed in Persephone; two pre-planned and the third was an emergency unplanned analysis triggered by the PHARE trial group publishing their own third interim analysis and concluding that 12 months duration was superior. The Persephone DMC concluded that there was no reason to stop Persephone and encouraged the trial to continue.

The Persephone trial demonstrated that six months of Herceptin treatment is not inferior to 12 months in this cohort of patients (Earl et al. 2019). The Persephone team set the NI margin to be no more than 3% absolute reduction of 4-year DFS. Furthermore, the upper bound of a 90% confidence interval was used to determine whether six months of Herceptin was non-inferior to 12 months of Herceptin treatment. After recruiting 4089 patients with 512 events, the four-year DFS rate was 89.8% in the 12 month arm and 89.4% in the six month arm (Table 7.1). This gave a HR of 1.07 (upper bound of the 90% CI: 1.24) with one-sided p-value of 0.01 for NI, hence demonstrating that six months of Herceptin treatment was not inferior to 12 months (Earl et al. 2019).

7.2.7 Summary of the results for the HERA, FinHer, SOLD, HORG PHARE and Persephone trials

The results of all these trials were used to determine the parameters that would be implemented to explore whether the use of a MAMS design would have been more efficient to test the optimal duration of Herceptin compared to performing separate trials using the 12 months of Herceptin treatment as the control arm and treatment durations less than 12 months as the experimental arm using a NI hypothesis (Table 7.1).

The results of the HERA, FinHer and SOLD trials showed that 12 months of Herceptin was better than no treatment and nine weeks. The HORG, PHARE and Persephone trials all compared six months of Herceptin treatment against 12 months but all had different countries involved. The Persephone trial showed six months to be non-inferior to 12 months. PHARE and HORG trials both failed to demonstrate NI of six months of Herceptin treatment compared to 12 months but it could be argued that the statistical considerations were flawed.

In 2020, three meta-analyses were published comparing the effects of different durations of Herceptin therapy (Deng et al. 2020, Gulia et al. 2020, Yu et al. 2020). Deng et al (2020) performed a meta-analysis across the PHARE, Persephone and HORG trials; all three of these trials compared 12 months of Herceptin against six months. The meta-analysis found no statistically significant difference for DFS (HR: 1.10, 95% CI: 0.99-1.23, p-value = 0.09) and OS (HR = 1.14, 95% CI: 0.99-1.32, p-value = 0.07) between the two treatment durations suggesting that there was no evidence of a difference between six and 12 months of Herceptin. Furthermore, Deng et al (2020) reported more AEs for patients treated with 12 months Herceptin versus six months (Risk ratio = 0.66, 95% CI: 0.56-0.77, p-value < 0.001). Gulia et al (2020) used published data from six trials assessing Herceptin therapy; four of the trials were Persephone, SOLD, PHARE and HORG (Gulia et al. 2020). The meta-analysis compared shorter duration of Herceptin versus 12 months of Herceptin using a relative NI margin of 1.3, calculated by taking the median NI margin across the six trials. The meta-analysis found that shorter use of Herceptin was no worse than using 12 months of Herceptin therapy (HR = 1.14, 95% CI: 1.03-1.25, p-value = 0.004) and less

risk of cardiac events using shorter durations (Relative risk = 0.53, 95%: 0.38-0.74). Yu et al (2020) combined 12 Herceptin based trials including PHARE, HORG, SOLD, Persephone and HERA trials, consisting of approximately 20,000 patients and used 12 months of Herceptin as the standard therapy (Yu et al. 2020). The team found that there was a statistically significant association (p-value = 0.002) between treatment time and recurrence risk. Each month Herceptin treatment was shortened from the 12 month standard would result in an increase in recurrence risk by 5.1%.

The three meta-analyses described here all obtained the data through public platforms, as of December 2020, the Persephone team at Warwick CTU are in the process of obtaining individual patient data for the PHARE, SOLD and HORG trials with the aim of performing a combined meta-analysis.

7.3 MAMS design with NI hypothesis

The trials described in section 7.2 were used to design a hypothetical MAMS trial using a NI hypothesis. The control arm was 12 months of Herceptin treatment and was compared to nine weeks, three months and six months of Herceptin treatment, therefore a four arm three stage MAMS design was implemented. The previous chapter showed the statistical properties of using a NI hypothesis with the 'nstage' command. The example used in the previous chapter was a three arm two stage trial but extending to a four arm three stage trial with a NI hypothesis would also work using the 'nstage' command.

7.3.1 Selection of parameter values for the MAMS design

The intermediate and primary outcome used for the MAMS design was DFS as this is what was used in the majority of the trials. It was assumed that the DFS rate at three years for 12 months of Herceptin was 90%, which was based on the latest DFS rates obtained from the Persephone and PHARE trials (Pivot et al. 2019, Earl et al. 2020) and randomisation would take place prior to any treatment. The total length of the trial was approximately 11 years; seven years for recruitment and four years follow-up. The MAMS designs used the same stagewise power values used in section 5.5, i.e. 95% at stages one and two and 90% at stage three and a one-sided FWER no more than 5%; strong emphasis was placed on controlling the FWER as all the arms of the

trial were implementing the same drug at different durations. Therefore, the stagewise alpha values were set at 0.50 for stage one, 0.25 for stage two and 0.02 for final stage to ensure a FWER no more than 5%. The focus of these MAMS designs was strong control of the FWER however in order to assess the impact of this approach, the MAMS designs were also calculated to have strong control of the PWER with the assumption that all treatments were independent of each other. For the MAMS designs with strong control for the PWER, the stagewise alpha values were set at 0.50 for stage one, 0.25 for stage two and 0.05 for the final stage to ensure a PWER no more than 5%. The MAMS designs with strong control of FWER and the MAMS designs with strong control of PWER were compared in terms of trial duration, number of control arm events and the critical HR. Like in chapters five and six, it was assumed that all experimental arms continued to the final stage and therefore maximise on the number of patients and time required as it is important that trial teams consider the extreme case scenario to present to funders to ensure they capitalise on funding given to support the trial. If trial teams do not consider the extreme case scenario, it may result in an under-budgeted study, which could result in either re-applying for additional funding or the closure of the trial. An outline of the hypothetical four arm three stage with strong control for the FWER and PWER can be seen in Figure 7.7.

A NI hypothesis was implemented as all experimental arms were reduced durations of the Herceptin drug. To consider the impact on the trial design, various NI margins were considered that reflected the superiority differences and NI margins that were implemented in the aforementioned trials (section 7.2). Additionally, the ICH E9 guidelines suggest that the NI margin should be smaller than the difference observed in the superiority trials of the control group (see section 1.4) (European Medicines Agency 1998). The superiority difference used in the HERA trial to compare 12 months of Herceptin against observation only was an absolute difference of 6.8%, therefore the NI absolute margin needed to be less than 6.8%. Hence, NI absolute margins ranging from 2% to 5% were evaluated.

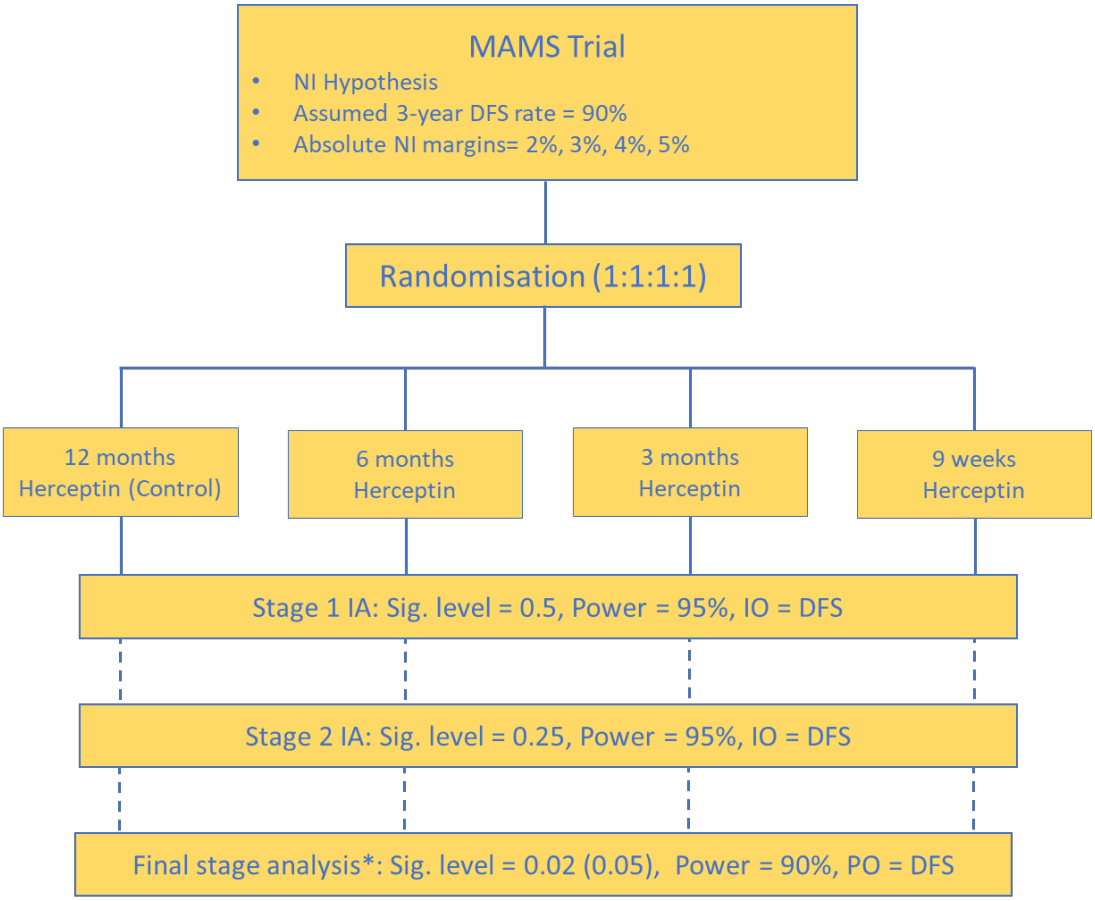

*Figure 7.7: Flow diagram of the hypothetical four arm three stage MAMS trial with strong control for FWER and PWER. *Final Significance level of 0.02 used for FWER and 0.05 for PWER. MAMS = Multi-arm multi-stage; NI = Non-inferiority; DFS = Disease-free survival; IA = Interim analysis; IO = Intermediate outcome; PO = Primary outcome.*

The total number of patients, the number of control arm events, the critical HR and timing of the interim analysis required for each of the different NI margins were recorded and compared. The sample size for each of the MAMS designs with different NI margins were compared against SOLD, HORG, PHARE and Persephone trials if these trials were repeated three times to accommodate three different durations. These trials were used as comparators as they all implemented a NI hypothesis.

7.3.2 Results

Applying a four arm three stage MAMS design with strong control for the FWER with the different NI margins resulted in, as expected, the 2% absolute NI margin requiring the most patients; approximately 2600 control arm patients followed by the design with a 3% margin, which required 1243 control arm patients, 4% margin required 752

control arm patients and then 5% margin that required 516 control arm patients (Figure 7.8). When using a 5% absolute NI margin, approximately 120 control arm events were required at the final stage compared to approximately 174 control arm events required when implementing a 4% NI margin. A 3% absolute NI margin required approximately 289 control arm events at the final stage analysis and more than double the number of control arm events were required at the final stage analysis for a 2% absolute NI margin (approximately 613 control arm events). The output for these calculations using the 'nstage' command can be found in appendix 10.

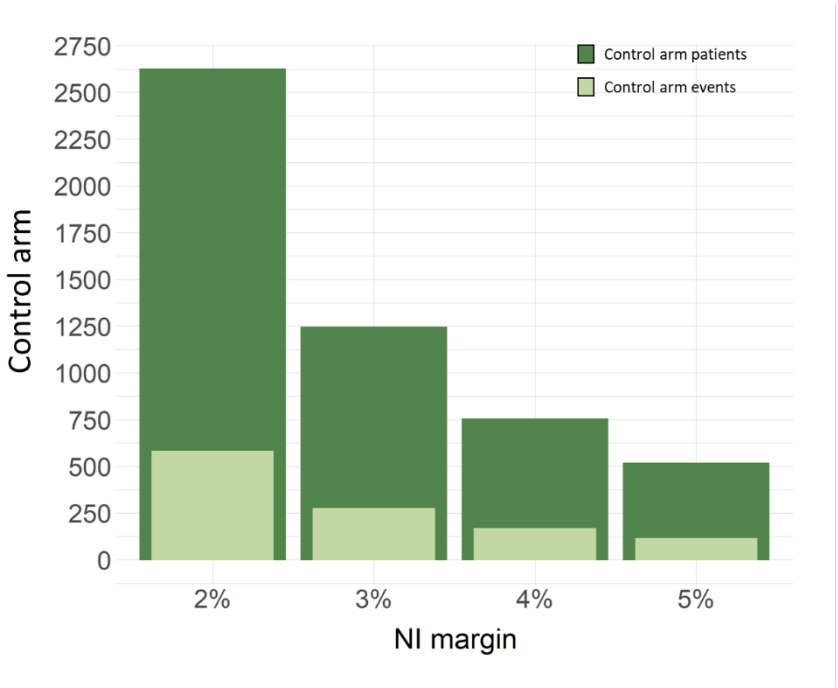

Figure 7.8: Number of control arm patients (dark green) and events (light green) for the MAMS designs with strong control of the FWER using different absolute NI margins. NI = Non-inferiority.

Table 7.2 shows the number of control arm events, the cumulative time of the analysis and the critical HR at each stage for the different NI margin values. If the pairwise comparison between the control arm and experimental arm has a HR greater than the critical HR, then this would result in discontinuing the experimental arm as it would deduce the inferiority of an experimental arm. It can be seen, as expected, that as the NI margin increases, the number of control arm events required

decreases and the critical HR values gets further away from the target HR. Furthermore, the timing of the analysis is consistent for all NI margins.

Stage	Target HR	Absolute NI margin					
		2%			3%		
		No. of Control arm events	Time of Analysis* (Years)	Consider discontinuation if HR (observed) is…	No. of Control arm events	Time of Analysis* (Years)	Consider discontinuation if HR (observed) is…
Stage I	1	≈149	4.9	> 1.210	≈71	4.9	> 1.320
Stage II	1	≈297	7.0	> 1.145	≈140	7.0	> 1.218
Final Stage III	1	≈613	11.1	> 1.076	≈289	11.1	> 1.113
		4%			5%		
Stage I	1	≈43	4.9	> 1.430	≈30	5.0	> 1.540
Stage II	1	≈85	7.0	> 1.289	≈58	7.0	> 1.359
Final Stage III	1	≈174	11.1	> 1.147	≈120	11.1	> 1.181

*Time is cumulative

Table 7.2: The number of control arm events and critical HR at each stage for the different absolute NI margins. NI = Non-inferiority; HR = Hazard ratio.

The sample sizes for each of these different MAMS designs were compared against the number of patients recruited in the SOLD, HORG, PHARE and Persephone trials repeated three times (Figure 7.9). The HORG trial would require the lowest number of patients however this trial did implement an absolute NI margin (8%), larger than the difference used in the HERA trial (HORG took place after the HERA trial results were released). Using a 3% (Persephone absolute NI margin), 4% (SOLD absolute NI margin) or 5% absolute NI margin would require fewer patients for a MAMS design compared to repeatedly performing the SOLD, PHARE or Persephone trial. Implementing a MAMS design with a 2% absolute NI margin would require fewer patients compared to repeatedly performing the Persephone trial however required more patients compared to repeatedly performing the PHARE trial (10500 patients for 2% absolute NI margin vs. 10152 patients for PHARE). The reason for the difference in the number of patients between the two trials could be due to the PHARE trial implementing a power of 80% compared to a higher power used at each of the stages in the MAMS designs (appendix 10).

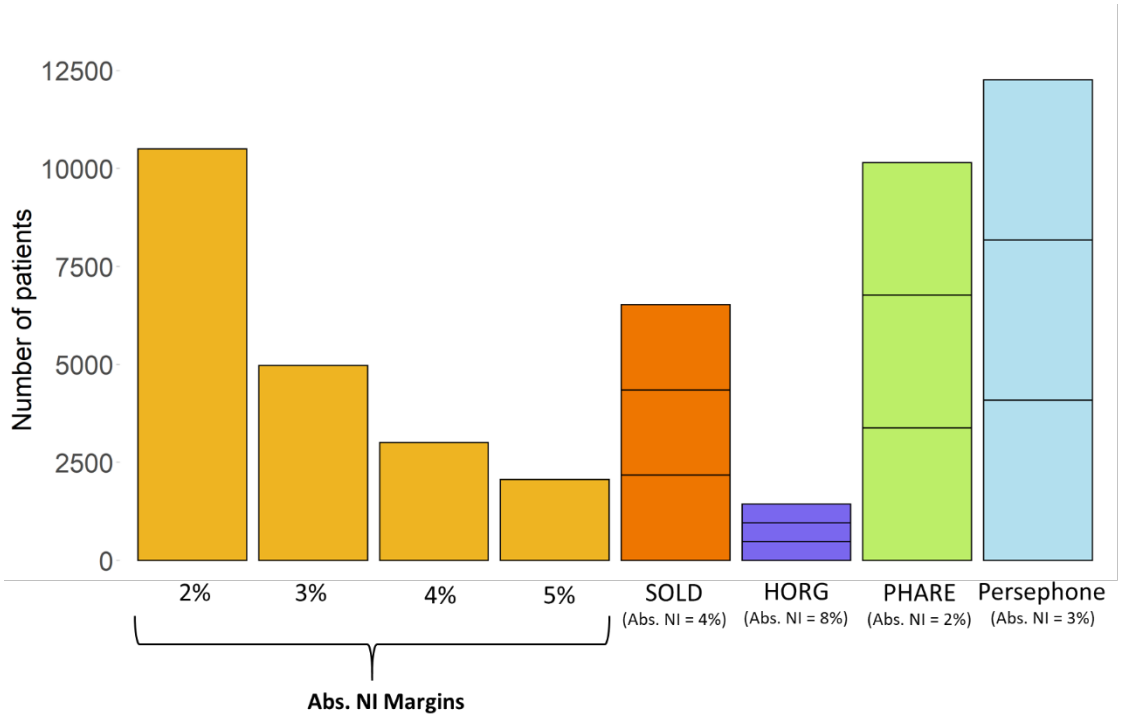

Figure 7.9: Comparing the sample size of the MAMS designs with different absolute NI margins against the number of patients recruited in the SOLD, HORG, PHARE and Persephone trials if these were repeated three times. NI = Non-inferiority.

The MAMS calculations for strong control of the PWER for the different absolute NI margins produced the same results for the interim control arm events, interim critical HRs and timing of the interim stages compared to the MAMS calculations for strong control of the FWER. The differences between the two occurred at the final stage. It can be seen in Table 7.3 that a larger final critical HR was used when controlling for the PWER in comparison to controlling for the FWER and as a result fewer control arm events were required. Therefore, it would take less time to reach these events and hence strong control for the PWER instead of the FWER would result in a trial duration of approximately 9 years in comparison to 11 years required. These results are as expected as a larger final stage alpha value of 0.05 is being used for the PWER MAMS calculations compared to a value of 0.02 when control for FWER.

MAMS designs	PWER	FWER
2% Abs. NI Margin		
Control arm events	≈472	≈613
Consider NI if HR (observed) is…	< 1.087	< 1.076
Time of Analysis (Years)	9.2	11.1
3% Abs. NI Margin		
Control arm events	≈223	≈289
Consider NI if HR (observed) is…	< 1.130	< 1.113
Time of Analysis (Years)	9.2	11.1
4% Abs. NI Margin		
Control arm events	≈134	≈174
Consider NI if HR (observed) is…	< 1.170	< 1.147
Time of Analysis (Years)	9.2	11.1
5% Abs. NI Margin		
Control arm events	≈92	≈120
Consider NI if HR (observed) is…	< 1.208	< 1.181
Time of Analysis (Years)	9.2	11.1

Table 7.3: Comparison of the final number of control arm events, final critical HR, time of the analysis at the final stage for when calculating MAMS designs for the different absolute NI margins with strong control of PWER and strong control of the FWER. MAMS = Multi-arm Multi-stage; PWER = Pairwise type I error rate; FWER = Familywise type I error rate; NI = Non-inferiority; HR = Hazard ratio.

7.3.3 Summary

Hypothetical four arm three stage MAMS designs using different NI margins were calculated and compared between each other and to existing trials that have tested different durations of Herceptin treatment to see if a MAMS design could have been implemented with less patients. The extreme case, where all experimental arms continue to the end of the trial was considered, however if the 'best case' scenario where all but one experimental arm showed inferiority at the first interim stage was considered then using the same number of patients for each NI margin would further reduce the trial length from 11 years to approximately 9.75 years. Especially when assessing different durations, the timing of the interim analysis is essential as there needs to be sufficient time for the treatment arms to differ. This is because there will be points in the trial where patients in different arms may have had the same duration, for example a patient randomised to 12 months of Herceptin may have only had three months of Herceptin when then interim analysis takes place. Therefore, if the interim analysis takes place too early, then there may be some patients that may

not have completed treatment and therefore do not have sufficient follow-up which may lead to incorrectly dropping an experimental arm or taking forward inadequate treatments as all patients are having the same treatment for a period of time and therefore unlikely to see any early differences. The MAMS designs considered in this chapter had the first interim analysis taking place at approximately five years, allowing patients on the control arm that had the longest duration of Herceptin therapy (12 months) to complete treatment and have sufficient follow-up.

Our calculations showed that these trials could have been executed using fewer patients with a MAMS design however it would be dependent on the choice of the NI margin which would require clinical input. Additionally, a snapshot of these designs were calculated however these could vary further by using different DFS rates, varying the allocation ratio, varying the number of interim stages etc.

The next section will look at some of the operational aspects that must be considered to execute these designs.

7.4 Operational considerations for MAMS trials

From a statistical standpoint, executing a MAMS trial may seem feasible however this is conditional upon the competence of where the trial will be hosted. Trial teams must take this into account whilst planning the trial and must take into consideration that other trials running simultaneously at the host site may impact operational efficiency.

Prior to conducting a MAMS trial, it is important that there is sufficient funding as MAMS trials may require extra funds especially when implementing a NI hypothesis where smaller margins are used compared to trials with a superiority hypothesis and therefore more patients are required. It is recommended that trial teams should consider the maximum number of patients and time required to ensure maximum funding is obtained and thus plan for all experimental arms continuing to the final phase. Additionally, trial teams could outline in the funding application that experimental arms may be added and by combining these additional arms into one trial can be considerably less in terms of cost then creating another standard two-arm RCT (Schiavone et al. 2019).

Upon writing the protocol, it is important to write it according to the SPIRIT guidelines (Chan et al. 2013). Include in the protocol the option to expand the trial to new experimental arms if new treatments warrant investigation during the course of the trial. Ensure the protocol is version controlled and any changes are clearly outlined to ensure transparency to all parties involved.

Data collection is a crucial component of any clinical trial. CRFs are used to record data for a trial and can be recorded in different ways. The method of obtaining data can directly impact the trial in terms of efficiency and cost. Implementing paper CRFs can be time consuming as site staff are required to fill these out and then trial team members would have to input this data onto a database. Time would be further consumed if there are difficulties in understanding what has been written on the CRFs. Trial teams can consider remote data entry performed by site staff, this ensures that data is legible, and processes can be put into place to try and ensure data is correctly entered. For example, if site staff input a patients' height as five meters, then the system would highlight this incorrect value. Innovative patient-centric methods to collect data where possible can be put into place. A mobile phone application linked to hospital records could be used to capture data. Implementing a smart application can reduce the number of patient and carer visits to sites and the smart application could be used to remind patients to enter data as well as direct reminders for appointments. However, creating an application may be costly, out of budget and may not be popular with patients who are not familiar with the technology.

Every clinical trial has its own designed database to reflect the CRFs requirements and to store all the data. The development of these systems can be timely and if the system is created in-house where there are a few specialists to create these systems then thorough planning is required to ensure this database is ready on-time. Using online cloud-based platforms to develop an appropriate database and capture patient data could be an alternative as these platforms allow for sufficient and safe data storage however these online platforms can be costly.

As discussed in chapter two, it is important to outline all trial processes before, during and after a trial to encourage efficiency. For example, at an interim analysis where

experimental arms may not continue, trial teams can plan for different situations to ensure that they are ready to go based upon the outcome of the interim analysis.

Independent committees such as DMCs and trial management committees guide the decisions made during the trial. Trial teams can consider employing larger committees to ensure that sufficient quorate is met.

The COVID19 pandemic has encouraged trialists to think of innovative methods to run clinical trials. One way could be to decentralise trials by utilising technology so trials can be conducted virtually which is beneficial as it allows home video consultations and minimises hospital visits and these can continue amidst a global pandemic. The FDA have also encouraged trial stakeholders to look at decentralised trials and utilise mobile technology to assist in conducting clinical trials (Shapiro et al. 2019, US Food and Drug Administration 2019).

Promoting efficient working practices is essential for all trials but especially for larger and more complicated MAMS design to ensure that costs are kept to a minimum.

7.5 Discussion

A MAMS design was considered to assess the duration of Herceptin therapy in women with early breast cancer to understand whether implementing a MAMS design would have been more efficient and effective compared to conducting multiple two arm trials to compare different durations.

Twelve months of Herceptin therapy became standard of care for women with early breast cancer after the execution of the HERA trial. The HERA trial became the basis on which many Herceptin trials were moulded with the aim of assessing the optimal duration of Herceptin. Six trials relating to the duration of Herceptin therapy in women with early breast cancer were assessed to help to form a MAMS based trial. It was vital that a NI hypothesis was used as the MAMS trial aimed to show a shorter duration could be no worse than the current standard of 12 months.

The PHARE trial specified a one-sided test with a 5% significance level and therefore should have used a 90% CI. The Persephone trial also specified a 5% one-sided significance level and hence it correctly uses the 90% CI limit for the HR to assess NI.

Additionally, a 90% CI was used within the SOLD trial (Joensuu et al. 2018). Therefore, if the PHARE trial implemented a 90% CI and a 2% absolute margin, i.e. recalculated the critical HR given their obtained control arm rate, as in the Persephone trial, then the results of PHARE would have demonstrated NI. Sample sizes for MAMS trials with different NI margins were calculated and compared. The bigger the NI margin, the fewer number of patients would be required. If the margin was too big then this would reflect on the robustness of the conclusions deemed from the trial, the clinical impact and patients may not favour a large NI margin. Implementing a 2% absolute NI margin would provide strong evidence to the conclusions obtained from the MAMS trial however the sample size required for this was over double the sample size required compared to using a 3% NI margin. It is important to note that different assumptions were made for all the original trials, for example the Persephone trial used 85% power and the PHARE trial used 80% power. These differences have not been fully taken into consideration when comparing these trials to the MAMS designs in Figure 7.9. However, it is envisaged that these discrepancies do not make a significant difference to the overall message that is being portrayed i.e. MAMS designs can be more efficient.

Emphasis was placed on controlling the FWER as this controls the probability of erroneously concluding that any duration is more effective than the control treatment as all the trial arms implemented the same drug at different durations (European Medicines Agency 2002). The trial team can make a priori assumption that longer durations will not be dropped earlier than shorter durations and therefore provide an ordering preference of dropping experimental arms, in this case a step-down procedure can be used to control the FWER (Bauer 1997). If an experimental arm were to stop during an interim analysis, the 'nstage' command calculates the FWER using the Dunnett probability which assumes that all experimental arms will continue to the final stage therefore the spend of alpha is conservative in that the FWER is smaller than the nominal level. Posch et al (2005) provided guidance on how to reduce this conservatism by allowing only for the arms which were actually included at each stage (Posch et al. 2005). There still remains some debate over whether control of the FWER is required when the arms are different drugs rather

than different durations of the same drug as evaluated in this chapter (Wason et al. 2014, Bretz et al. 2020, Parker et al. 2020).

The decision to drop or continue a trial arm at the interim stage of a MAMS clinical trial is often the role of the DMC, based on the pre-planned interim stopping rules provided by the trial team. Quartagno et al (2018) propose 'duration designs' whereby the duration-response curve is modelled to identify appropriate treatment durations, this method can be incorporated into a MAMS design to aid the DMC in decision making. For example, information could be gathered from the duration-response curve to estimate an acceptable duration which would have a specified loss of efficiency. This information could also be combined with toxicity or cost-effectiveness to ensure that the DMC have an adequate amount of information to assist in making key decisions. The choice of design cannot be solely decided by a statistician hence various scenarios must be reviewed when designing adaptive trials to paint a clear picture for all those involved with the decision of which scenario to implement.

Some of the operational aspects to execute an adaptive trial have been discussed in this chapter and in chapter two. The lessons learnt from the STAMPEDE and FOCUS4 trials demonstrate that to encourage the smooth running of a trial, proper processes and thorough planning is a necessity. Trial teams that wish to execute large trials are encouraged to refer to the operational papers published by the STAMPEDE and FOCUS4 team (Hague et al. 2019, Schiavone et al. 2019).

7.6 Conclusion

The application of a MAMS designs with a NI hypothesis from a statistical viewpoint seems feasible however the input of other trial stakeholders including clinicians, patients and trial team members is crucial to understand the clinical and operational feasibility of implementing large adaptive trials.

8 Discussion, future work and conclusions

8.1 Discussion and Key thesis findings

Phase III RCTs are the gold standard way to test a new intervention or procedure in people with the aim to find out if these treatments are safe, have any side effects, are better than the currently used treatment or have any effect on the quality of life of patients (Cancer Research UK 2018). The use of adaptive designs has greatly aided and improved the way trials are executed and conducted in all phases by making them more efficient. These innovative approaches in conducting trials are becoming more appreciated and known by people within the medical profession and fit very well in the oncology setting where there are often many drugs to consider at the same time (Kairalla et al. 2012).

The first chapter laid the foundations of the research by providing an introduction into RCTs, survival analysis, adaptive designs and briefly highlighting the differences between Bayesian and frequentist philosophies. Hypothesis testing options were also addressed and it was shown how a superiority hypothesis and a NI hypothesis can be interchanged, which was necessary in order to support the research reported in subsequent chapters throughout the thesis.

Chapter two explored the commonly considered adaptive design methods, e.g. sample size re-estimation, adaptive randomisation, group sequential designs, and their application within phase III cancer trials. This thesis has primarily focused on the MAMS framework by Royston et al (2011), hence chapter two details the methodology of this adaptive method and provides an example of how to apply the 'nstage' command in Stata that can be used to apply this framework. The draft guidance on adaptive designs given by the FDA in 2010 was originally followed (US Food and Drug Administration 2010) but modified to reflect the amended FDA guidelines for adaptive designs (US Food and Drug Administration 2018) that were released during the PhD. The latest set of guidelines are more compact and efficient, but the use of certain key words is different, i.e. the 2010 guidelines used 'Type I error rate' whereas the updated guidelines use 'Type I error probability'. These updated guidelines emphasise the vital role simulations play in planning adaptive

designs compared to the previous guidelines which referred to the use of simulations as 'not fully understood'. As mentioned in chapter two, the FDA originally categorised the different adaptive designs into well understood and less well understood methods; these have now been removed from the updated version which reflects the FDAs familiarity with these designs. Therefore, it is evident from the different adaptive methods researched that there will be continuous evolution of adaptive design methodology with more novel methods being developed and considered to assist the clinical trial community at all phases of clinical trials.

It was clear from the literature review different adaptive design methods are available however it wasn't clear how much these adaptive designs had been applied in phase III RCTs in Oncology. Hence, this led to performing a review of the current literature to recognise the application of adaptive design methodology in RCTs in cancer and how these RCTs were being reported. A literature search was conducted reviewing the use of adaptive design methods within cancer trials in 2015. Within one year, 54 papers that used some form of adaptive design method within cancer trials were found. Previous literature reviews (Bauer et al. 2006, Stevely et al. 2015, Hatfield et al. 2016) investigated the methods and reporting of adaptive design, however the review in chapter three was unique and relevant to this PhD as the aim was to only concentrate on the application of adaptive designs methods within phase II, phase II/III and phase III RCTs in cancer. The conclusions drawn from this literature review suggested that adaptive design methods are regularly implemented but reporting of these methods were poor. This reaffirmed the conclusions made by previous literature reviews who also investigated the reporting of adaptive designs (Bauer et al. 2006, Stevely et al. 2015, Hatfield et al. 2016).

Due to the poor reporting of adaptive designs highlighted from the literature review, an extension to the current CONSORT Statement guidelines was proposed to improve the reporting of adaptive designs in the future. The literature review reported within chapter three was published during the course of this PhD (Mistry et al. 2017a). A consensus driven CONSORT Statement extension for RCTs using adaptive design was later published (Dimairo et al. 2018) and a workshop was held by the ACE group in 2019 whereby the finalised guidelines were disseminated (appendix 5). The

differences and similarities between the two guidelines have been compared in chapter three.

The conclusions drawn from chapter three led to the scoping exercise of the cancer trials run within Warwick CTU to assess their use of adaptive methods as it was evident from chapter three that adaptive design methods were being implemented but not being reported as adaptive. The scoping exercise in chapter four showed that the cancer trials within Warwick CTU regularly made adaptations without classing them as adaptive designs, thus supporting the need for better guidelines to report adaptive designs. This led to the assessment of exemplar trials in chapter four where trials that had successfully implemented adaptive designs, one of which was the STAMPEDE trial which used a MAMS design to compare different treatments for patients diagnosed with prostate cancer which has a poor prognosis and therefore the events occur quickly. Additionally, the MAMS framework by Royston et al (2011) was used in the FOCUS4 and CompARE trials.

MAMS designs are becoming increasingly popular due to the ability and efficiency of assessing multiple experimental arms within one trial to answer many questions rather than conducting separating trials and the subsequent cost saving implications. MAMS designs have received great support from patients and clinicians (Parmar et al. 2014). It was felt important to further delve into the application of these MAMS designs within the field of Oncology and its use in trials with differing prognosis. Therefore, the trials that implemented a MAMS design in chapter four aided in the selection of key parameters in subsequent chapters.

Chapter five found that applying a MAMS designs with a superiority hypothesis to three cancer sites (breast, colon and lung cancer) would be feasible and fewer patients are required compared to conducting several two-arm trials. The reason for selecting these specific cancer sites was because the five-year OS rate could range from 16% in lung cancer to 85% in breast cancer and hence would give a good overview of the application of these designs. For poor, moderate and good survival outcomes it was concluded that the outcome of a MAMS design can be influenced by various factors and how the factors influence the sample size and duration of a

MAMS design are listed below and these learnings were taken forward in subsequent chapters:

- Primary and intermediate outcome: Using a time-to-event outcome that provides with sufficient events that clinicians are comfortable with, for example in chapter five, implementing OS in the HERA_OS trial resulted in only 20 events required to conclude superiority. Therefore, clinicians and patients may not be convinced that conclusions can be made with few events;
- Number of arms at each stage: It is recommended that trial teams consider a range of scenarios. When applying for funding, show the most extreme case to make sure that there are enough resources available when applying for money from funders. The definition of 'extreme' can be subjective as per the trial teams objectives however in this thesis 'extreme' is defined as all experimental arms continuing to the final stage as this would require the most patients and therefore the most cost to funding;
- Significance level (α): Suitable choice for alpha as the value at each stage can influence the overall alpha value and the timing of the interim analyses. Implementing the values specified by Royston et al (2011) and the STAMPEDE are a suitable benchmark however these can vary depending on the requirements of the trial.

A key novel aspect of this PhD was presented in chapter six where it was shown how to implement a NI hypothesis using the 'nstage' command. Simulations were performed that validated the statistical properties of using the 'nstage' command with a NI hypothesis. Furthermore, the calculations applied in chapter five were extended to implement MAMS designs with two of these trials using a NI hypothesis. The calculations showed that MAMS designs with a NI hypothesis can be

implemented and required fewer patients in comparison to performing separate two-arm trials. However, it is recommended that the MAMS designs with a NI hypothesis are not applied where there is a low survival rate with small NI margins e.g. in lung cancer as it is felt that within this patient population that only large clinically relevant differences would be appropriate.

The assessment of the Persephone trial in chapter four, the application of a MAMS design to early breast cancer in chapter five and implementing a NI hypothesis in chapter six led to the development of a trial looking at Herceptin treatment duration in chapter seven. The use of Herceptin therapy in women with early breast cancer was first assessed in the HERA trial. Thereafter, the Persephone trial and many other trials were created to assess the optimal duration of Herceptin therapy.

Various trials that have evaluated the different durations of Herceptin treatment in patients diagnosed in early breast cancer were assessed to see if applying a MAMS design with a NI hypothesis would have been more efficient than conducting the trials separately and explore the operational aspects required to execute these designs. The results of these investigations have shown that MAMS designs can be considered for cancers with good survival rates, whereas they have been only previously implemented for cancer with poor survival rates. Thus, extending the use of these MAMS designs to more situations.

There was a difference in the assumed rate of DFS used for the Persephone trial. The sample size calculation used 80% four-year DFS rate compared to the observed rate in the final analysis of 89.8% four-year DFS rate. The substantial change in the DFS rate reflects the progress made in the management of the cancer disease from screening procedures or early diagnosis to treatment and care provided to patients diagnosed with the disease (Quaresma et al. 2015). The UK took active precautions to reduce the number of deaths due to Cancer by creating 'The NHS Cancer plan' (Department of Health 2000). Over time this plan has developed and continues to pledge to reduce the number of cancer deaths (Burki 2019). The plan published in January 2019 (Burki 2019) aims to ensure that any patient that may be at the risk of cancer will receive a diagnosis within 28 days of being referred. Hence, there has

been active and prioritised research in the field of Oncology which explains the improvement in the DFS rates over time.

In chapters five and six, the treatment arms in the MAMS designs were different drugs and therefore considered independent whereas in chapter seven the investigations involved different durations of the same drug. This added an additional complexity to the decision making process. Thus emphasis was placed on controlling the FWER as this controls the probability of erroneously concluding that any duration is more effective than the control treatment. However, MAMS designs with strong control of the PWER were also calculated and contrasted to the MAMS designs with strong control of the FWER. Strong control of the PWER would result in requiring a larger final critical HR and hence fewer control arm events in comparison to controlling for the FWER due to a larger final stage alpha value of 0.05 for the PWER MAMS calculation compared to a value of 0.02 when controlling for the FWER.

In 2018, a methodology paper investigating the use of NI for optimising treatment duration was published (Quartagno et al. 2018). The authors of this paper proposed a multi-arm randomised trial of different treatment durations and modelled the duration-response curve to identify appropriate treatment durations. The aim of this paper was similar to the aims of this PhD, exploring the use of a NI hypothesis in a trial of multiple arms to assess treatment duration. However, this PhD investigates the optimal treatment duration within Oncology trials using multiple arms as well as multiple stages allowing the dropping of experimental arms at each stage due to lack of benefit or futility. The modelling of the duration-response curve could be implemented within this MAMS framework to aid decision making by estimating the duration that allows for a specified loss of efficiency. The approach by Quartagno et al (2018) considered models looking at treatment duration, these models could be extended further to incorporate key prognostic factors. Therefore, allow clinicians to prescribe different durations for specified subgroups and hence a more personalised medicine approach.

8.2 Limitations of this thesis and future work

The aim of this thesis has been primarily focused on the statistical elements of applying adaptive designs, in particular MAMS designs with a superiority and NI hypothesis within the field of Oncology. Future work could entail implementing a MAMS design with the vital input from other trial stakeholders. This would prove valuable in raising the robustness that these MAMS design can provide. Obtaining valuable guidance from experienced trial stakeholders and patients could support the inclusion for a set of guidelines for various stages of a MAMS design and address the following:

- Setting the key parameters for the MAMS design such as the hazard ratio, the treatment difference or the number of years required to recruit patients;
- Seeking guidance from anyone that may internally/externally specialise in adaptive designs to assist with the write up of the protocol or statistical analysis plan;
- Speaking to teams that have employed adaptive designs to ensure efficiency in all operational aspects. For example, what process could be put into place to allow sufficient turn-around time for data management, how to engage site staff and keep them motivated, how to conduct training to get new staff members up to speed etc.
- Allowing sufficient time between any interim analyses for other tasks such as up to date data entry, data validation, data queries etc.;
- Consider whether a pilot study is required to assess the feasibility of performing a MAMS design;
- Is there sufficient funding to support the MAMS design? What is the potential for obtaining extra funding to support the inclusion of additional experimental arms?

It may be difficult to apply a MAMS design to rare cancers such as renal cancer, where the diagnosis of patients is low or where there is a very good prognosis cancer with high survival rates hence very low event rates. Furthermore, MAMS designs are only valuable if there are multiple treatments to assess. Also trial teams that do not have the operational capacity should not employ MAMS designs.

The 'nstage' command offers a wide range of different options that could be applied to MAMS designs. The latest update has been released for the 'nstage' command whereby a selection of efficacy stopping boundaries can be implemented (Blenkinsop et al. 2019). The scope of this work could be further extended in various ways, few of which have been mentioned in previous chapters and some suggested below:

- Continuing with a NI design and assess adding different treatment durations whilst a trial is on-going;
- Implementing an adaptive randomisation procedure within a MAMS design whereby more patients are allocated to treatments that are performing better;
- Sample size re-estimation based on observed data at an interim analysis;
- Extend the 'nstage' command to consider non-proportional hazards;
- Implement the MAMS design to other cancer sites or non-cancer areas to assess treatment duration. For example, these designs could be applied to the treatment of bacterial infections to understand the optimal duration for anti-biotics treatment.

8.3 Conclusions

It is evident that since the start of this PhD in January 2016, adaptive designs have come a long way with more streamlined guidelines released by the FDA, a consensus driven CONSORT Statement extension for reporting of adaptive designs and more literature related to adaptive design published. It is apparent from these achievements that the use and demand for adaptive designs is continually increasing.

Additionally, adaptive designs have proved to be extremely valuable in the investigation of new treatments and vaccines for COVID-19 (Noor et al. 2020, Stallard et al. 2020).

This thesis demonstrated with time to event outcomes in the field of oncology that the use of MAMS designs can be more efficient than the gold standard parallel two-arm RCT and be applied in disease areas with good survival outcomes, i.e. in early breast cancer where the five-year disease free survival rates can be around 90%. Furthermore, these MAMS designs can be extended to implement a NI hypothesis, and not just comparing treatments in the superiority setting. These MAMS designs can be applied to other therapeutic areas like cardiovascular disease where TTE outcomes such as time to first occurrence of myocardial infarction are implemented. A similar version of the 'nstage' command is called 'nstagebin', which applies MAMS designs with binary intermediate and primary outcomes and are analysed using the absolute difference in proportions (Bratton et al. 2013). These designs can be applied to various disease areas with endpoints such as whether a patient is cured or not, if a patient had a major cardiac event etc. This research has shown the value of applying adaptive designs but urges trial stakeholders and funders to consider implementing adaptive designs where possible. These designs will greatly aid all those involved in clinical trials by ensuring that trials are ethical (fewer patients required and quicker set-up times), efficient (one control arm for several treatment comparisons and quicker set-up times) and effective (variety of treatments considered so trial is never outdated by new treatments). It has been emphasised in this thesis that trial teams must be fully aware of the operationally aspects required for these designs and not to underestimate the resources required to run them. It is anticipated that with more support and trust given to these types of designs, adaptive designs may eventually become the new 'gold standard' for phase III clinical trials or at least be considered when designing a new RCT.

9 Appendices

Appendix 1: The VIETNARMS trial

There are very limited trials that have focused on decreasing the length of treatment of hepatitis C (currently treatment last between 8-12 weeks) especially in cohorts of patients with genotype six. The VIETNARMS trial team have proposed a large trial using two NI hypotheses followed by a superiority hypothesis in which patients are factorially randomised to 14 treatment arms and stratified by genotype six vs. non-genotype six (McCabe et al. 2020). The trial implements a binary primary endpoint where patients have either a sustained virologic response or treatment failure. The term 'factorially randomised' has been used as patients are randomised to either one of two drug regiments given for 12 weeks (regimen comparison), control or three treatment decreasing strategies (strategy comparison) and of those randomised to treatment decreasing strategies, they will also be randomised to adjunctive ribavirin or no adjunctive ribavirin (RBV comparison) (Figure 9.1).

The sample size calculations assumed that the true cure rate was 90% in each group (null hypothesis) compared to an unacceptably low cure rate of 70% (alternative). A sample size of 39 patients would be required within each treatment arm to exclude the cure rate being lower than 70% with 90% power and a one-sided significance level of 5%. A total of 1092 patients will be required as there are two strata multiplied by the 14 treatment arms multiple by 39 patients in each stratified treatment arm. If any one of the groups are stopped at an interim analysis, then future patients will be assigned to any of the remaining groups to ensure that maximum information is gain rather than to minimising the sample size.

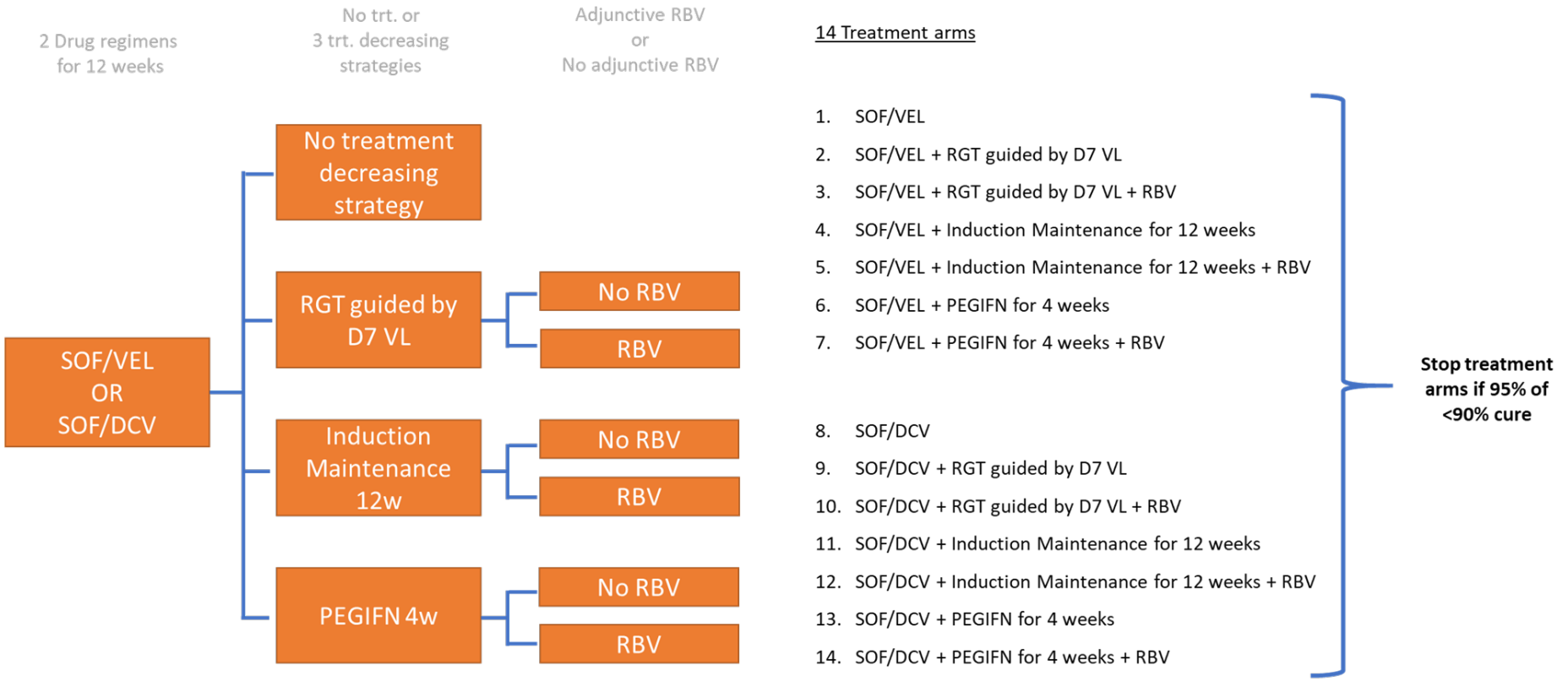

Figure 9.1: Patients are factorially randomly assigned to two drug regimens, three treatment reducing strategies or control and either adjunctive ribavirin or no ribavirin. SOF = Sofosbuvir; VEL = Velpatasvir; DCV = Daclatasvir; RGT = Response guided therapy; D7 VL = Day 7 Viral load; PEGIFN = Pegylated interferon; RBV = Ribavirin.

Based on the initial sample size calculations for all treatment arms and assuming an overall cure rate of 95%, 1092 patients are sufficient to demonstrate 5% NI between drug regimens with 97% power and 10% NI between the strategy comparison with 96% power both with a one-sided significance value of 5%. Superiority comparisons will take place between ribavirin and the regimen comparisons and the strategy comparison conditional that the above two comparisons meet the NI margin. The sample size will allow for an absolute difference of 5% to be detected with 90% power for the drug regimens or ribavirin comparison and an absolute difference of 7% to be detected with 80% power for the strategy comparison both with two-sided significance of 5%. The trial team selected a 5% absolute NI margin for the regimens comparison based on clinical judgement and its previous use in other trials. A larger NI margin was used for the strategy comparison because for the same cost significantly more patients can be treated.

The hypothesis, treatment, allocation ratio and sample size used for each comparison can be seen in Table 9.1.

Comparison	Hypothesis	Treatment	Allocation ratio	Sample size
Regimen comparison	5% NI	1. SOF/VEL	1:1	546
		2. SOF/DCV		546
Strategy comparison	10% NI	1. No strategy (control)	1:2:2:2	156
		2. RGT guided by D7 VL		312
		3.Induction Maintenance		312
		4. PEGIFN 4W		312
RBV comparison*	5% Superiority difference - Regimen/RBV comparison > 7% superiority difference - Strategy/RBV comparison	1. No RBV	1:1	468
		2. RBV		468

*Excludes patients that have been allocated no strategy control group

Table 9.1: Hypothesis, treatment, allocation ratio and sample size used for each comparison for the VIETNARM trial. SOF = Sofosbuvir; VEL = Velpatasvir; DCV = Daclatasvir; RGT = Response guided therapy; D7 VL = Day 7 Viral load; PEGIFN = Pegylated interferon; RBV = Ribavirin

This trial has implemented a Bayesian monitoring procedure to allow the stopping of inferior arms early and hence successive patients to be randomised to the remaining arms if there is 95% posterior probability of less than a 90% cure in any of the treatment arms. This approach has been implemented because there is limited

knowledge about what effect the interventions will have as it will allow for flexibility in deciding when to implement an interim analyses compared to frequentist methods such as the MAMS design in which the timing and number of interim analyses are pre-specified (Royston et al. 2011). The trial implemented four interim analyses, the timing of these were determined by assessing projected recruitment schedules to ensure sufficient patients were accrued and applying the probability of stopping the groups over different cure rates using a monitoring beta (α, β) prior, where α and β are shape parameters and $\alpha = 4.5$ and $\beta = 0.5$. This prior beta (4.5, 0.5) was selected with a mean of 0.9, variance of 0.015 and a 34% probability that a cure rate of less than 90%; a low precision prior was selected as it would allow the data to be more influential in the posterior distribution.

The trial will be assessed by an independent DMC and they will make the final decision as to whether a group should be stopped using the pre-specified stopping guidelines as well as different sensitivity analyses such as different prior values to ensure the correct decision is made.

Appendix 2: Published literature review

Mistry et al. BMC Medical Research Methodology (2017) 17:108
DOI 10.1186/s12874-017-0393-6

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

A literature review of applied adaptive design methodology within the field of oncology in randomised controlled trials and a proposed extension to the CONSORT guidelines

Pankaj Mistry*, Janet A Dunn and Andrea Marshall

Abstract

Background: The application of adaptive design methodology within a clinical trial setting is becoming increasingly popular. However the application of these methods within trials is not being reported as adaptive designs hence making it more difficult to capture the emerging use of these designs. Within this review, we aim to understand how adaptive design methodology is being reported, whether these methods are explicitly stated as an 'adaptive design' or if it has to be inferred and to identify whether these methods are applied prospectively or concurrently.

Methods: Three databases; Embase, Ovid and PubMed were chosen to conduct the literature search. The inclusion criteria for the review were phase II, phase III and phase II/III randomised controlled trials within the field of Oncology that published trial results in 2015. A variety of search terms related to adaptive designs were used.

Results: A total of 734 results were identified, after screening 54 were eligible. Adaptive designs were more commonly applied in phase III confirmatory trials. The majority of the papers performed an interim analysis, which included some sort of stopping criteria. Additionally only two papers explicitly stated the term 'adaptive design' and therefore for most of the papers, it had to be inferred that adaptive methods was applied. Sixty-five applications of adaptive design methods were applied, from which the most common method was an adaptation using group sequential methods.

Conclusions: This review indicated that the reporting of adaptive design methodology within clinical trials needs improving. The proposed extension to the current CONSORT 2010 guidelines could help capture adaptive design methods. Furthermore provide an essential aid to those involved with clinical trials.

Keywords: Adaptive, Adaptive design, Review, Clinical trials, Cancer, Interim analysis

* Correspondence: Pankaj.mistry@warwick.ac.uk
Warwick Clinical Trials Unit, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

Background

In recent years, there has been a rise of interest in adaptive design methodology [1, 2]. The Food and Drug Administration (FDA) define an adaptive design as "a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data from subjects in the study" [3]. Conversely Chow and Chang (2008) have broadened this definition by classifying adaptive designs as any modifications made prospectively, concurrently or retrospectively during the conduct of a trial.

The implementation of adaptive designs can have potential benefits in clinical trials over other study designs such as parallel design, crossover etc. [4]. They can be more efficient, more cost effective, the likelihood of success increases, and there is an improved understanding of treatment success [1, 4, 5], however these benefits are only possible if the validity and integrity of the proposed study isn't undermined [1]. Furthermore adaptive design methods may appeal more to clinical investigators due to the flexibility and prospect of making changes based on data at an interim stage of the trial.

Previous reviews into the application of adaptive design methodology have already been conducted. For example, Bauer (2006) conducted a review to investigate papers that were published between the years 1989 and 2004. The purpose of their review was to explore the impact of adaptive design methodology in medicine, particularly those designs based on the combination test or conditional error function, and to see if these methods were applied and presented appropriately [6]. A review conducted by Stevely (2015) assessed the standards of phase III group sequential randomised controlled trials (RCT) against the CONSORT 2010 checklist [7]. This review looked at papers published between 1st January 2001 and 23rd September 2014. Eligible papers were found in 11 different therapeutic fields, of which the majority (76%) were found in the field of oncology. This review concluded that there were issues with the reporting of group sequential trials and hence suggest an extension to the CONSORT checklist to help solve this problem. A recently published review by Hatfield (2016) examines the state of phase II, II/III and III adaptive design trials registered between the years 2000 and 2014. The ClinicalTrials.gov website and National Institute for Health Research register were used to collate registered trials with adaptive designs. Of all the registered trials with adaptive designs, the review found that adaptive designs are most often used in the field of oncology. The review did not successfully capture all trials with adaptive designs, and hence suggested that clinical trial registers should dedicate sections for adaptive designs and encourage the use of the term 'adaptive design' in either the title, summary or design sections of the register [8].

The aforementioned reviews have looked into specific methods related to adaptive design methodology or attempted to find all adaptive designs registered on Clinical trial registries. It is unclear the current extent of applications of adaptive design methodology specifically in oncology trials and whether reporting has improved over time. Therefore this literature review specifically attempts to capture papers published in 2015 that are using adaptive design methodology within phase II, II/III or III RCT in the field of Oncology. This review aims to understand how adaptive design methodology is currently being reported within journals, whether these methods are being applied prospectively or concurrently, what methods are commonly applied and to understand whether the use of adaptive design methodology is being explicitly stated or whether it has to be inferred. Furthermore this review will suggest extensions to the CONSORT that can be implemented to aid better reporting and to maximise the capture of adaptive design methodology.

Methods

Literature search

The Embase, Ovid and PubMed databases were chosen to conduct the literature search into the application of adaptive design methodology. The review was constrained to phase II, phase III or phase II/III RCT's with patients diagnosed with cancer that presented primary outcome trial results and were published in 2015. Eligible papers should indicate the use of adaptive design methodology, papers should be full text publications in the English language and accessible. Duplicate records were excluded based on the title, authors, abstract and year of publication. All required data were extracted and recorded on an excel spreadsheet.

The definition that will be used to identify the application of adaptive design methodology will be any potential modifications made to the trial/statistical procedure that is either prospective, ad-hoc or retrospective [1, 9].

A free text search was conducted to capture phase II, II/III, or IIII Cancer RCTs, the following keywords were used: "phase II(2)", "phase III(3)", "phase II(2)/III(3)", "Oncology", "Cancer", "Neoplasm", "Carcinoma", "randomis(z)ed controlled trial(s)", "randomis(z)ed clinical trial(s)", "trial", "controlled clinical trial(s)". To capture as many results related to adaptive design methodology within these trials, the following keywords were used alongside the Boolean operator "OR": "adaptive design(s)", "flexible design(s)", "group sequential", "adaptive randomis(z)ation", "sample size re-estimation", "sample size adjustment", "sample size modification", "MAMS", "multi(−)arm multi(−)stage", "multi(ple) arm", "multi(ple)

stage", "interim analysi(e)s", "adaptive seamless", "biomarker adaptive", "adaptive clinical trial(s)", "two-stage adaptive", "multiple adaptive", "adaptive enrichment", "dose escalation", "dose selection", "drop the loser", "pick the winner" and "treatment switch(ing)".

Data extraction

An excel spreadsheet was used to record the following data:

- Standard demographics such as first author, title, name of the trial;
- The journal that the paper was published in;
- The funder or sponsor of the study;
- The phase of the trial;
- The type of cancer being reported;
- The nature of the primary outcome;
- The number of trial arms;
- The type of intervention being implemented;
- The number of any planned interim analyses;
- The stage of the trial being reported, i.e. interim or final analysis;
- Any planned stopping criteria and reason;
- Whether the trial was terminated early and the reason for early stopping;
- The initial planned sample size and the reported sample size;
- The adaptive designs methodology used;
- The number of adaptive designs that were applied;
- Whether the adaptive design was predetermined or concurrent;
- Whether the use of adaptive designs was explicitly stated;
- The trial identifier if it had been registered on clinical trial websites.

The papers identified by the literature search were all reviewed and information extracted and recorded in the aforementioned excel spreadsheet. Data that could not be found in the paper was researched by using the trial identifier or trial name to find out the relevant information. If no further information was available then the data was classified as missing. One person (PM) extracted the information from the papers and any papers that needed further clarification were checked and validated by two reviewers as a form of quality control (AM and JD).

Results

A total of 8288 records were identified related to RCTs in the field of Oncology that were published in 2015 across the Ovid, Embase and PubMed databases. Of which 734 records were identified using the key search terms mentioned above that were related to adaptive designs within the phase II, phase II/III and phase III trial setting.. After the removal of duplicates 464 records were screened, of which 368 records were excluded due to the following reasons: published abstract only (n = 263), not related to the literature review (n = 66), either methodology or review papers hence would not contain results (n = 33) or the full reports were inaccessible (n = 6). The remaining 96 full text records were further assessed for eligibility, of which 42 were excluded for the following reasons: not a RCT (n = 29), not cancer related (n = 3), not the analysis of the primary outcome (n = 9) or no information provided (n = 1). This left a total of 54 records for analysis (Fig. 1) (see Additional file 1 for the full list of titles for each paper).

Of the 54 papers reviewed, 38 papers (70%) were phase III confirmatory trials, 12 out of 54 papers (22%) were phase II trials, 2 (4%) were phase IIb trials and 2 (4%) were phase II/III trials. The number of arms in a trial ranged from two to five arms, 46 (85%) papers were two arm trials of which 35 of these were phase III trials, 10 of the two arm trials were phase II trials, there was 1 four arm phase III trial and 1 five arm phase II/III trial (Table 1).

The literature review identified adaptive trials being reported in 21 different journals (Appendix 1). Of these the Journal of Clinical Oncology published the highest number of papers with 12 out of 54 papers (22%), followed by the Lancet Oncology with 11 papers (21%), then the New England Journal of Medicine with 6 papers (11%) and the European Journal of Cancer publishing 3 papers (6%). The remaining 17 journals published either one or two papers (Appendix 1).

There were 45 out of 54 (81%) papers that had a time to event outcome as its primary outcome (Appendix 2). Furthermore 49 out of 54 (92%) were drug related trials (Appendix 3).

Of the 54 papers reviewed, 33 (61%) were published based on results during an interim analysis, the results of the remaining papers were based on either final analysis (20/54) or subgroup analysis (1/54). Of the 33 papers published based on results during an interim analysis, 26 papers resulted in the trial stopping early, of which all 26 had a pre-planned stopping criteria that stopped early due to either safety/efficacy/futility i.e. group sequential methods (Table 2). The remaining 7 papers stated to continue with the trial as planned. The majority of the papers (48/54) had a pre-planned interim analysis with 34 out of the 48 specifying one interim analysis during the course of the trial, 9 specifying two interim analyses, 3 specifying three interim analyses and the remaining 2 papers conducting interim analyses annually during the course of the trial.

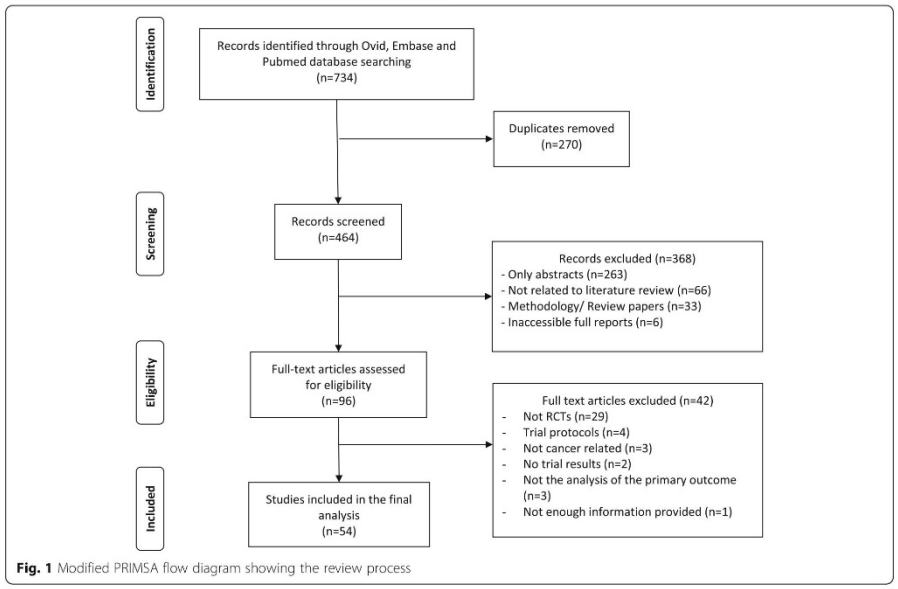

Fig. 1 Modified PRIMSA flow diagram showing the review process

Majority of trials applied a single adaptive design method (44/54 papers, 80%), 9 out of 54 papers (19%) applied two methods and 1 paper (2%) applied three methods. In total there were 65 applications of adaptive design methods, of which the most reported method was adaptations using group sequential methods with 50 out of 65 applications, followed by dose modifications (8/65), sample size re-estimation (4/65), adaptive randomisation (1/65),) change in primary endpoint (1/65) and change in patient eligibility (1/65).

Table 3 shows the different variables extracted split by the adaptive method applied. All papers that applied group sequential methods had incorporated a planned stopping criteria. Furthermore 35/50 of these group sequential methods were in a phase III setting. All 4 papers that implemented sample size re-estimation methods resulted in the trial stopping early. Additionally 3 out of 4 of these papers had both pre-determined and ad-hoc applications of adaptive design methods. Conversely all papers that applied dose modification methods had pre-determined the use of adaptive design methods.

Of the 54 papers reviewed, 49 (91%) had predetermined the use of adaptive design methods, 1 paper had applied ad-hoc adaptive design methods and 4 papers had predetermined and ad-hoc use of adaptive design methods. However only 2 out of 54 papers (4%) had explicitly used the phrase 'adaptive design'

Table 1 Two way table of phase of a trial and number of trial arms it has

Trial phase	Number of arms				Total
	2	3	4	5	
Phase II	10	2	0	0	12
Phase IIb	1	1	0	0	2
Phase II/III	0	1	0	1	2
Phase III	35	2	1	0	38
Total	46	6	1	1	54

Table 2 Two way table of whether the trial stopped early compared to if the trial had a planned stopping criteria

Trial stopped early	Pre-planned stopping criteria			Total
	No	Yes	Unknown	
No	5	22	1	28
Yes	0	26	0	26
Total	5	48	1	54

Table 3 The data extracted split by the adaptive method applied

Data extracted	Adaptive method applied					
	Group sequential methods (n = 50)	Dose modifications (n = 8)	Sample size re-estimation (n = 4)	Adaptive randomisation (n = 1)	Change in primary endpoint (n = 1)	Change in patient eligibility (n = 1)
Trial phase						
II	13	2	0	1	0	0
II/III	2	1	1	0	0	0
III	35	5	3	0	1	1
Number of arms						
2	42	7	2	1	1	1
3	6	1	0	0	0	0
4	1	0	1	0	0	0
5	1	0	1	0	0	0
Stage of reporting						
Interim analysis	33	2	2	0	1	1
Subgroup analysis	0	1	0	0	0	0
Final analysis	17	5	2	1	0	0
Planning of adaptive design method						
Pre-determined	45	8	1	1	0	1
Ad-hoc	1	0	0	0	0	0
Both	4	0	3	0	1	0
Explicitly stated Adaptive design						
Yes	1	1	1	1	0	0
No	49	7	3	0	1	1
Planned stopping criteria						
Yes	50	5	4	0	1	1
No	0	3	0	1	0	0
Trial stopped early						
Yes	26	1	4	0	1	0
No	24	7	0	1	0	1

to explain that adaptive methodology was applied, the remaining papers indicated the use of adaptive design methods. Of the papers that explicitly stated the use of adaptive design methodology, one paper had implemented a Bayesian adaptive response design which included adaptive randomisation and dose modification. The other paper applied multiple adaptive methods which included the use of group sequential methods, sample size re-estimation, and dose modification.

Discussion

This literature review aimed to understand the reporting of adaptive design methodology in RCTs in particular within the field of oncology. This review has highlighted that the reporting of these methods needs improving, which confirms the outcome of other reviews. Stevely et al. reported that there are issues related to the reporting of group sequential trials and suggested a consort extension to alleviate the issues related to this [7]. Hatfield et al. emphasises the need for improving the way adaptive designs are reported and suggests a modification to the CONSORT statement [8]. Additionally Bauer and Einfalt suggest that the presentation of adaptive design methodology needs to be developed [6].

The classifications of the adaptive design methods were based upon those mentioned by the FDA. One form of adaptive design methods are those using group sequential methods, these methods have been used extensively for a number of years hence the robustness of these methods have qualified them to be known as well understood methods by the regulators [3, 10]. This design employs stopping boundaries at regular interims to assist in decision making with

Table 4 Proposed extensions to the current CONSORT diagram [15]

Section/Topic	Item No	Standard Checklist item	Extension for adaptive designs
Title and abstract			
	1a	Identification as a randomised trial in the title	Identification as an adaptive randomised trial if it is an adaptive design
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts [16, 17])	Include the term 'adaptive design' or ' adaptive methods'
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	Rational for implementing an adaptive design
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	Define what adaptive design/ adaptive method will be applied
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	Any changes during the trial should be reported as an adaptive method.
Participants	4a	Eligibility criteria for participants	Any changes in eligibility during the trial, should be classed as an adaptive design or adaptive method.
	4b	Settings and locations where the data were collected	
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	
	6b	Any changes to trial outcomes after the trial commenced, with reasons	Any changes during the trial are classed as an adaptive method and should be mentioned.
Sample size	7a	How sample size was determined	Any changes to sample size or power during trial classed as an adaptive design or adaptive method and should be mentioned.
	7b	When applicable, explanation of any interim analyses and stopping guidelines	Explain why the interim analysis will be taking place, if potential pre-planned adaptations during interim analysis taking place then these should be mentioned in the methods as well (3b). Include details of any planned stopping boundaries for either the trial or dropping any of the intervention arms.
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	Details if adaptive randomisation has been implemented.
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	

Table 4 Proposed extensions to the current CONSORT diagram [15] *(Continued)*

	11b	If relevant, description of the similarity of interventions	
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	Details of how the adaptive design or the adaptive methods were applied Details of how the statistical methods were evaluated before implementation i.e. through the use of simulations?
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	Ensure any adaptations are shown on this diagram, such as dropping of arms, treatment switching.
	13b	For each group, losses and exclusions after randomisation, together with reasons	
Recruitment	14a	Dates defining the periods of recruitment and follow-up	
	14b	Why the trial ended or was stopped	Any changes to recruitment during trial classed as an adaptive method, should be mentioned.
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms [18])	
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	If ad-hoc adaptive methods were implemented, at what point was it decided to implement this and why.
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	
Registration	23	Registration number and name of trial registry	
Protocol	24	Where the full trial protocol can be accessed, if available	First and last protocol, with a list of amendments made.
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	

regards to the trial or treatment. Many methods such as Simon's two-stage design, O'Brien and Fleming design, multi-arm multi-stage designs can be included within the umbrella of group sequential designs [11–14]. Hence any papers that employed the aforementioned designs or applied methods whereby trial/treatment related decisions could be made during an interim analysis were classed as group sequential methods. It was found from the literature review that group sequential methods were used within 50 out of 65 applications.

The review found that majority of the papers (53/54) applied adaptive methods that were prospectively planned hence supporting the definition given by the FDA. The list of search terms helped in capturing applications of adaptive design methodology however only two papers explicitly stated the term 'adaptive design' and it was inferred from the remaining papers that adaptive design methodology was used.

The reporting of these studies has not improved and hence this review supports the need for a set of guidelines of how adaptive designs should be reported. A proposed extension to the current CONSORT 2010 guidelines has been made (Table 4). This extension should be used for any trials that fit the definition of adaptive designs as used for this literature review, i.e. for trials with any modifications made to the trial/statistical procedure that was either prospective, ad-hoc or retrospective. The justification for attempting to create an extension to the CONSORT guidelines is to ensure that as many adaptive design based trials are captured. A crucial question that needs to be answered is 'at what point exactly a trial become classed as adaptive?' Table 4 can assist that decision for those involved in the running of a trial.

Pragmatically this extended CONSORT would greatly assist in efficient capturing of published papers related to adaptive design methodology. Furthermore extending the CONSORT 2010 guidelines would encourage greater capture of predetermined adaptations which would greatly benefit all those involved in clinical trials ensuring explicit and thorough reporting of the adaptive nature of RCTs and the methodology used. It will allow full transparency of all adaptations carried out during the trial.

Conclusions

It can be concluded that the reporting of adaptive design methodology within RCT's is inadequate and requires improvement. To assist in the capture of adaptive design methods the proposed extension to the CONSORT 2010 guidelines can be implemented; this will also prove to be a crucial aid to all members involved with clinical trials.

Appendix 1

Table 5 List of journals from which the papers were published in

Journal	Frequency
Annals of Oncology	2
BMC Research Notes	1
Breast Cancer Research & Treatment	1
British Journal of Haematology	2
Cancer	2
Cancer Management and Research	1
Clinical Cancer Research	1
Clinical Lung Cancer	2
Drug design, development & therapy	1
European Urology	1
European Journal of Cancer	3
Gynaecologic Oncology	2
Head & Neck	1
Journal of American Medical Association	1
Journal of Clinical Oncology	12
Journal of Urology	1
Lancet Oncology	11
Molecular Cancer Therapeutics	1
New England Journal of Medicine	6
Leukemia	1
Radiotherapy & Oncology	1
Total	54

Appendix 2

Table 6 List of primary outcomes of the papers

Primary outcome	Frequency
Bowel movement/Flushing episodes	1
Cytogenic Response	2
Disease Control Rate	1
Optimal dose	1
Overall pain response	1
Response to treatment	1
Success/Failure	2
Time to event outcome	45
Total	54

Appendix 3

Table 7 List of categorised interventions of the papers

Intervention	Frequency
Adding acid	1
Drug	49
Surgery/Chemotherapy/Radiotherapy	3
Vaccine	1
Total	54

Additional file

Additional file 1: Dose modification. (XLSX 14 kb) (XLSX 14 kb)

Abbreviations
FDA: Food and Drug Administration; RCT: Randomised controlled trials; SAP: Statistical analysis plan

Acknowledgements
Not applicable.

Funding
Not applicable.

Availability of data and materials
The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Authors' contributions
PM conducted the literature search, analysed the findings and contributed majorly in writing the manuscript. AM and JD contributed to the conception and design of the study, performed quality checks on the findings from the literature review, participated in the interpretation of the data and the revision of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 20 March 2017 Accepted: 10 July 2017
Published online: 18 July 2017

References
1. Chow S-C, Chang M. Adaptive design methods in clinical trials-a review. Orphanet J Rare Dis. 2008;3(11):169–90.
2. Kairalla JA, et al. Adaptive trial designs: a review of barriers and opportunities. Trials. 2012;13(1):145.
3. Food US, Administration D. Adaptive design clinical trials for drugs and biologics. Biotechnol Law Rep. 2010;29(2):173.
4. FDA, Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics, U.S.D.o.H.a.H. Services, Editor. 2010.
5. CHMP, CHMP reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design, 2007. 2012.
6. Bauer P, Einfalt J. Application of adaptive designs–a review. Biom J. 2006; 48(4):493–506.
7. Stevely A, et al. An Investigation of the Shortcomings of the CONSORT 2010 Statement for the reporting of group sequential randomised controlled trials: a methodological systematic review. PLoS One. 2015;10(11):e0141104.
8. Hatfield I, et al. Adaptive designs undertaken in clinical research: a review of registered clinical trials. Trials. 2016;17(1):1.
9. Bretz F, et al. Adaptive designs for confirmatory clinical trials. Stat Med. 2009;28(8):1181–217.
10. Todd S. A 25-year review of sequential methodology in clinical studies. Stat Med. 2007;26(2):237–52.
11. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;35(3):549–56.
12. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989;10(1):1–10.
13. Royston P, Parmar MK, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. Stat Med. 2003;22(14):2239–56.
14. Medical Research Council Clinical Trials Unit. MAMS Design. Available from: http://www.ctu.mrc.ac.uk/our_research/research_types/methodology1/design/mams_design/. Accessed 23 Feb 2017.
15. CONSORT. Consolidated Standards of Reporting Trials. 2010. Available from: http://www.consort-statement.org/. Accessed 10 Apr 2017.
16. Hopewell S, et al. CONSORT for reporting randomised trials in journal and conference abstracts. Lancet. 2008;371(9609):281–3.
17. Hopewell S, et al. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. PLoS Med. 2008;5(1):e20.
18. Ioannidis JP, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med. 2004;141(10):781–8.

Appendix 3: List of search terms used for literature review

A free text search was conducted using the following key search terms to maximise the capture of phase II, II/III or III RCTs in the field of Oncology:

- “phase II”
- “phase 2”
- “phase III”
- “phase 3”
- “phase II/III”
- “phase 2/3”
- “Oncology”
- “Cancer”
- “Neoplasm”
- “Carcinoma”
- “randomised controlled trial”
- “randomised controlled trials”
- “randomized controlled trial”
- “randomized controlled trials”
- “randomised clinical trial”
- “randomised clinical trials”
- “randomized clinical trial”
- “randomized clinical trials”
- “trial”
- “controlled clinical trial”
- “controlled clinical trials”

The following key search terms were used together with the Boolean operator "OR" to maximise the capture of results related to adaptive design methods:

- "adaptive design"
- "adaptive designs"
- "flexible design"
- "flexible designs"
- "group sequential"
- "adaptive randomisation"
- "adaptive randomization"
- "sample size re-estimation"
- "sample size adjustment"
- "sample size modification"
- "MAMS"
- "multi arm multi stage"
- "multi-arm multi-stage
- "multi arm"
- "multiple arm"
- "multi stage"
- "multi-stage"
- "interim analysis"
- "interim analyses"
- "adaptive seamless"
- "biomarker adaptive"
- "adaptive clinical trial"

- “adaptive clinical trials”
- “two-stage adaptive”
- “multiple adaptive”
- “adaptive enrichment”
- “dose escalation”
- “dose selection”
- “drop the loser”
- “pick the winner”
- “treatment switch”
- “treatment switching”

Appendix 4: List of 54 published papers used in the literature review

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
1	5-fluorouracil, leucovorin, and oxaliplatin (mFOLFOX6) plus sunitinib or bevacizumab as first-line treatment for metastatic colorectal cancer: A randomized Phase IIb study	Hecht, J.R	NCT00609622	2b	Group sequential methods	PFS	Drug	Interim analyses	Yes	Yes	Lack of efficacy
2	A Multicenter, Phase II, Randomized, Noncomparative Clinical Trial of Radiation and Temozolomide with or without Vandetanib in Newly Diagnosed Glioblastoma Patients	Lee, E.Q	NCT00441142	2	Group sequential methods	OS	Drug	Interim analyses	Yes	Yes	Futility
3	A phase IIb multicentre study comparing the efficacy of trabectedin to doxorubicin in patients with advanced or metastatic untreated soft tissue sarcoma: the TRUSTS trial	Bui-Nguyen, B.	NCT01189253	2b	Group sequential methods	PFS	Drug	Interim analyses	Yes	Yes	Lack of superiority
4	A phase III randomized trial of adding topical nitroglycerin to first-line chemotherapy for advanced nonsmall-cell lung cancer: the Australasian lung cancer trials group NITRO trial	Davidson, A.	ACTRN12608000588392	3	Group sequential methods	PFS	Drug	Interim analyses	Yes	Yes	no demonstrable effect of nitroglycerin
5	A phase III study of radiation therapy (RT) and O6-benzylguanine, (O6-BG) plus BCNU versus RT and BCNU alone and methylation status in newly-diagnosed glioblastoma (GBM) and gliosarcoma: Southwest Oncology Group (SWOG) Study S0001	Blumenthal, D.T	NCT00017147	3	Group sequential methods	OS	Drug	Interim analyses	Yes	Yes	Futility
6	A randomized phase II study of the MEK1/MEK2 inhibitor trametinib (GSK1120212) compared with docetaxel in KRAS-mutant advanced non-small-cell lung cancer (NSCLC)	Blumenschein Jr, G.R	NCT01362296	2	Group sequential methods	PFS	Drug	Interim analyses	Yes	Yes	Futility

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
7	A randomized trial comparing concurrent chemoradiotherapy with single-agent cisplatin versus cisplatin plus gemcitabine in patients with advanced cervical cancer: An Asian Gynecologic Oncology Group study	Wang, C.C	NCT00842660	3	Group sequential methods	OS/PFS	Drug	Interim analyses	Yes	Yes	Slow recruitment
8	A randomized, controlled trial of oral propranolol in infantile hemangioma	Leaute-Labreze, C.	NCT01056341	Phase 2/3	Group sequential methods Adaptations to Sample Size	Success/Failure	Drug	Final analyses	Yes	No	N/A
9	A randomized, open-label phase II trial of volasertib as monotherapy and in combination with standard-dose pemetrexed compared with pemetrexed monotherapy in second-line treatment for non-small-cell lung cancer	Ellis, P.M.	NCT00824408	2	Group sequential methods	PFS	Drug	Final analyses	Yes	No	N/A
10	Abiraterone acetate plus prednisone versus placebo plus prednisone in chemotherapy-naive men with metastatic castration-resistant prostate cancer (COU-AA-302): final overall survival analysis of a randomised, double-blind, placebo-controlled phase 3 study	Ryan, C.J	NCT00887198	3	Group sequential methods	OS/PFS	Drug	Final analyses	Yes	No	N/A
11	Adjuvant sorafenib for hepatocellular carcinoma after resection or ablation (STORM): A phase 3, randomised, double-blind, placebo-controlled trial	Bruix, J.	NCT00692770	3	Group sequential methods Adaptations to Treatment Arm Selection	RFS	Drug	Final analyses	Yes	No	N/A

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
12	BCR-ABL1 mutation development during first-line treatment with dasatinib or imatinib for chronic myeloid leukemia in chronic phase	Hughes, T.P.	NCT00481247	3	Adaptations to Treatment Arm Selection	Cytogenic Response	Drug	Final analyses	No	No	N/A
13	Bendamustine, thalidomide and dexamethasone combination therapy for relapsed/refractory myeloma patients: results of the MUKone randomized dose selection trial	Schey, S.	ISRCTN90889843	2	Group sequential methods	Optimal dose	Drug	Interim analyses	Yes	Yes	Futility
14	Biological 18[F]-FDG-PET image-guided dose painting by numbers for painful uncomplicated bone metastases: A 3-arm randomized phase II trial	Berwouts, D.	NCT01429493	Phase 2/3	Group sequential methods Adaptations to Treatment Arm Selection	Overall pain response	Radiotherapy	Interim analyses	Yes	No	N/A
15	Bosutinib versus imatinib in newly diagnosed chronic-phase chronic myeloid leukaemia: Results from the 24-month follow-up of the BELA trial	Brummendorf, T.H.	NCT00574873	3	Adaptations to Treatment Arm Selection	Cytogenic Response	Drug	Final analyses	Yes	No	N/A
16	Cabozantinib versus Everolimus in Advanced Renal-Cell Carcinoma	Choueiri, T.K.	NCT01865747	3	Group sequential methods	PFS	Drug	Final analyses	Yes	No	N/A
17	Capecitabine combined with docetaxel versus vinorelbine followed by capecitabine maintenance medication for first-line treatment of patients with advanced breast cancer: Phase 3 randomized trial	Want, J.	NCT01126138	3	Group sequential methods Adaptations to Treatment Arm Selection	PFS	Drug	Interim analyses	Yes	Yes	Met primary endpoint

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
18	Carfilzomib and dexamethasone versus bortezomib and dexamethasone for patients with relapsed or refractory multiple myeloma (ENDEAVOR): a randomised, phase 3, open-label, multicentre study	Dimopoulos, M. A	NCT01568866	3	Group sequential methods	PFS	Drug	Final analyses	Yes	No	N/A
19	Efficacy and Safety of Abiraterone Acetate in Elderly (>=75 years) Chemotherapy-Naive patients with Metastatic Castration-Resistant Prostate Cancer	Demetri, G. D.	NCT00887198	3	Group sequential methods	OS	Drug	Final analyses	Yes	No	N/A
20	Efficacy and Safety of Trabectedin or Dacarbazine for Metastatic Liposarcoma or Leiomyosarcoma After Failure of Conventional Chemotherapy: Results of a Phase III Randomized Multicenter Clinical Trial	Lonial, S.	NCT01343277	3	Group sequential methods	OS/PFS	Elotuzumab Therapy	Final analyses	Yes	No	N/A
21	Elotuzumab Therapy for Relapsed or Refractory Multiple Myeloma	Yao, J. C.	NCT01239797	3	Group sequential methods	PFS	Drug	Final analyses	Yes	No	N/A
22	Everolimus for the treatment of advanced, non-functional neuroendocrine tumours of the lung or gastrointestinal tract (RADIANT-4): a randomised, placebo-controlled, phase 3 study	Fogarty, G.B	NCT01524783	3	Group sequential methods	OS	Surgery/RT	Interim analyses	Yes	No	N/A
23	First interim analysis of a randomised trial of whole brain radiotherapy in melanoma brain metastases confirms high data quality	Zaman, K.	NCT01503827	2	Group sequential methods	Disease Control Rate	Drug	Interim analyses	Yes	Yes	Safety/Efficacy

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
24	Fulvestrant with or without selumetinib, a MEK 1/2 inhibitor, in breast cancer progressing after aromatase inhibitor therapy: a multicentre randomised placebo-controlled double-blind phase II trial, SAKK 21/08	Manfredi, S.	NCT01160718.	2	Group sequential methods	Success/Failure	Drug	Interim analyses	Yes	Yes	Futility
25	High-Dose FOLFIRI plus Bevacizumab in the Treatment of Metastatic Colorectal Cancer Patients with Two Different UGT1A1 Genotypes: FFCD 0504 Study	Chanan-Khan, A.	NCT00628810.	3	Group sequential methods	PFS	Drug	Interim analyses	Yes	Yes	Met primary endpoint
26	Ibrutinib combined with bendamustine and rituximab compared with placebo, bendamustine, and rituximab for previously treated chronic lymphocytic leukaemia or small lymphocytic lymphoma (HELIOS): a randomised, double-blind, phase 3 study	Saad, F.	NCT01611090.	3	Group sequential methods Adaptations to Patient Population	PFS	Drug	Interim analyses	Yes	No	N/A
27	Impact of bone-targeted therapies in chemotherapy-naive metastatic castration-resistant prostate cancer patients treated with abiraterone acetate: post hoc analysis of study COU-AA-302	Robert, C.	NCT00887198	3	Group sequential methods	OS	Drug	Interim analyses	Yes	Yes	Efficacy
28	Improved overall survival in melanoma with combined dabrafenib and trametinib	Tiseo, M.	NCT01597908	3	Group sequential methods	OS	Drug	Interim analyses	Yes	No	N/A
29	Italian multicenter phase III randomized study of cisplatin-etoposide with or without bevacizumab as first-line treatment in extensive stage small cell lung cancer: treatment rationale and protocol design of the GOIRC-AIFA FARM6PMFJM trial	O'Brien, M. E.		3	Group sequential methods	OS	Drug	Interim analysis	Yes	Yes	Efficacy

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
30	Maintenance pazopanib versus placebo in Non-Small Cell Lung Cancer patients non-progressive after first line chemotherapy: A double blind randomised phase III study of the lung cancer group, EORTC 08092 (EudraCT: 2010-018566-23, NCT01208064)	Stupp, R.	NCT01208064	3	Group sequential methods	PFS	Drug	Interim analysis	Yes	Yes	Met primary endpoint
31	Maintenance Therapy With Tumor-Treating Fields Plus Temozolomide vs Temozolomide Alone for Glioblastoma: A Randomized Clinical Trial	Kordes, S.	NCT00916409	2	Group sequential methods	OS	Drug	Final analysis	Yes	No	N/A
32	Metformin in patients with advanced pancreatic cancer: A double-blind, randomised, placebo-controlled phase 2 trial	Bielack, S.S.	NCT01210911	3	Group sequential methods Adaptations to Treatment Arm Selection	EFS	Drug	Final analysis	Yes	No	N/A
33	Methotrexate, doxorubicin, and cisplatin (MAP) plus maintenance pegylated interferon alfa-2b versus MAP alone in patients with resectable high-grade osteosarcoma and good histologic response to preoperative MAP: First results of the EURAMOS-1 good response randomized controlled trial	Blay, J.Y.	NCT00134030	3	Group sequential methods Adaptations to Sample Size	PFS	Drug	Interim analysis	Yes	Yes	Futility
34	Nilotinib versus imatinib as first-line therapy for patients with unresectable or metastatic gastrointestinal stromal tumours (ENESTg1): a randomised phase 3 trial	Weber, J.S.	NCT00785785	3	Group sequential methods	OS	Drug	Final analysis	Yes	No	N/A

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
35	Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial	van Oers, M. H.	NCT01721746	3	Group sequential methods	PFS	Drug	Interim analysis	Yes	No	N/A
36	Ofatumumab maintenance versus observation in relapsed chronic lymphocytic leukaemia (PROLONG): an open-label, multicentre, randomised phase 3 study	Turner, N.C.	NCT00802737	3	Group sequential methods	PFS	Drug	Final analysis	Yes	No	N/A
37	Palbociclib in Hormone-Receptor-Positive Advanced Breast Cancer	Ribas, A.	NCT01942135	2	Group sequential methods	PFS	Drug	Interim analysis	Yes	No	N/A
38	Pembrolizumab versus investigator-choice chemotherapy for ipilimumab-refractory melanoma (KEYNOTE-002): a randomised, controlled, phase 2 trial	Swain, S.M.	NCT01704287	3	Group sequential methods	PFS	Drug	Final analysis	Yes	No	N/A
39	Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer	McMeekin, S.	NCT00567190	3	Group sequential methods	OS	Drug	Interim analysis	Yes	Yes	Futility
40	Phase III randomized trial of second-line ixabepilone versus paclitaxel or doxorubicin in women with advanced endometrial cancer	Wolin, E.M.	NCT00883116	3	Group sequential methods	Bowel movement/Flushing episodes	Drug	Interim analysis	Yes	Yes	Safety/Efficacy/Futility
41	Phase III study of pasireotide long-acting release in patients with metastatic neuroendocrine tumors and carcinoid symptoms refractory to available somatostatin analogues		NCT00690430	3	Group sequential methods	OS	Surgery/CT	Interim analysis	Yes	Yes	Slow recruitment

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
42	Phase III Study of Surgery Versus Definitive Concurrent Chemoradiotherapy Boost in Patients With Resectable Stage IIIA(N2) and Selected IIIB Non-Small-Cell Lung Cancer After Induction Chemotherapy and Concurrent Chemoradiotherapy (ESPATUE)	Luck, H.J.		3	Group sequential methods	PFS		Interim analysis	Yes	Yes	Safety/Futility
43	Phase III study on efficacy of taxanes plus bevacizumab with or without capecitabine as first-line chemotherapy in metastatic breast cancer	Oki, E.	NCT01200212	3	Group sequential methods	Time to neuropathy	Drug	Interim analysis	Yes	Yes	Safety
44	Preventive effect of Goshajinkigan on peripheral neurotoxicity of FOLFOX therapy (GENIUS trial): a placebo-controlled, double-blind, randomized phase III study	Oza, A.M.		2	Group sequential methods	PFS	Drug	Final analysis	Yes	No	N/A
45	Randomized Phase II Trial of Ridaforolimus in Advanced Endometrial Carcinoma	Rodriguez, C.P.	NCT00739830	3	Group sequential methods	RFS	Drug	Interim analysis	Yes	Yes	Futility
46	Randomized phase III study of 2 cisplatin-based chemoradiation regimens in locally advanced head and neck squamous cell carcinoma: impact of changing disease epidemiology on contemporary trial design	Abe, T.		3	Group sequential methods	OS	Drug	Interim analysis	Yes	Yes	Safety/Efficacy/Futility
47	Randomized phase III trial comparing weekly docetaxel plus cisplatin versus docetaxel monotherapy every 3 weeks in elderly patients with advanced non-small-cell lung cancer: the intergroup trial JCOG0803/WJOG4307L	Rugo, H.S.		3	Group sequential methods	PFS	Drug	Interim analysis	Yes	Yes	Futility

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
48	Randomized Phase III Trial of Paclitaxel Once Per Week Compared With Nanoparticle Albumin-Bound Nab-Paclitaxel Once Per Week or Ixabepilone With Bevacizumab As First-Line Chemotherapy for Locally Recurrent or Metastatic Breast Cancer: CALGB 40502/NCCTG N063H (Alliance)	Komatsu, Y	NCT00785291	3	Group sequential methods	PFS	Drug	Subgroup analysis	No	No	N/A
49	Regorafenib for advanced gastrointestinal stromal tumors following imatinib and sunitinib treatment: a subgroup analysis evaluating Japanese patients in the phase III GRID trial	Issa, J.P.	NCT01271712	2	Adaptations to Treatment Arm Selection	Response to treatment	Adding acid	Final analysis	No	No	N/A
50	Results of phase 2 randomized study of low-dose decitabine with or without valproic acid in patients with myelodysplastic syndrome and acute myelogenous leukemia	Le, D.T.	NCT0l305499	2	Adaptation to Patient Allocation Adaptations to Treatment Arm Selection	OS	Vaccine	Final analysis	Yes	No	N/A
51	Safety and survival with GVAX pancreas prime and Listeria Monocytogenes-expressing mesothelin (CRS-207) boost vaccines for metastatic pancreatic cancer	Budd, G.T.	NCT01417000	3	Group sequential methods	DFS	Drug	Third Interim	Yes	No	N/A
52	SWOG S0221: a phase III trial comparing chemotherapy schedules in high-risk early-stage breast cancer	Finn, R.S.	NCT00070564	2	Group sequential methods Adaptations to Sample Size	PFS	Drug	Final analysis	Yes	No	N/A

No.	Title	First Author	Trial identifier	Phase	Adaptive methods applied	Primary Outcome	Type of intervention	Stage of reporting	Planned stopping criteria	Trial stopped early	Reason for early stopping
53	The cyclin-dependent kinase 4/6 inhibitor palbociclib in combination with letrozole versus letrozole alone as first-line treatment of oestrogen receptor-positive, HER2-negative, advanced breast cancer (PALOMA-1/TRIO-18): a randomised phase 2 study	Casali, P.G.	NCT00721409	3	Group sequential methods	OS/RFS	Drug	Interim Analysis	Yes	No	N/A
54	Time to Definitive Failure to the First Tyrosine Kinase Inhibitor in Localized GI Stromal Tumors Treated With Imatinib As an Adjuvant: A European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group Intergroup Randomized Trial i	Smith, R.M.	NCT00103168	3	Group sequential methods Adaptations to Sample Size Adaptations to Endpoint Selection	OS/PFS	Drug	Interim analyses	Yes	No	N/A

Appendix 5: Dissemination of Adaptive Design CONSORT Extension

From: Sarah Gonzalez <s.k.gonzalez@sheffield.ac.uk>
Sent: 07 June 2019 15:57
To: Sarah Gonzalez <s.k.gonzalez@sheffield.ac.uk>
Subject: Re: Adaptive Designs CONSORT Extension Dissemination Workshop Agenda:12 June 2019

Dear all,

Please find the papers for the Adaptive Designs CONSORT Extension Dissemination Workshop next week.

1) Background information on adaptive designs; why use them, how to run and report them: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-018-1017-7
2) Development process of the ACE guidance: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-018-1196-2
3) ACE checklist including the abstract (Table 1 and 2) - attached
4) ACE guideline in open peer-review - attached https://www.researchsquare.com/article/8c909820-f69b-4a54-b805-e952a4f71373/v1
5) Group exercises 1, 2 and 3 - attached

We would recommend attendees read 3) and 4) before the day.

Paper copies of 3) 4) and 5) will be available on the day.
Please let me know if you have any questions at all.

Kind regards,
Sarah

On Thu, 30 May 2019 at 15:10, Sarah Gonzalez <s.k.gonzalez@sheffield.ac.uk> wrote:

> Good afternoon,
>
> We are delighted that you are able to join us on Wednesday 12th June, 2019 at **Room 7, National Council for Voluntary Organisations NCVO, Society Building, 8 All Saints Street, London, N1 9RL**
>
> Refreshments will be available at 9:15 and the meeting will begin at 10:00, finishing at 16:00.
>
> For your information, please find attached the agenda for the day and a map of the venue. The venue is a 10 minute walk from London Kings Cross/St Pancras Underground Station.
>
> Please get in touch in you have any questions.
>
> We are looking forward to meeting with you and wish you a safe journey to London.
>
> With kind regards,
>
> The ACE Team

Appendix 6: List of all trials obtained from ClinicalTrials.gov (Extraction date: 28 March 2020)

No.	Title	Phase(s)	Enrolment	URL
Breast cancer trials				
1	Platinum and Polyadenosine 5'Diphosphoribose Polymerisation (PARP) Inhibitor for Neoadjuvant Treatment of Triple Negative Breast Cancer (TNBC) and/or Germline BRCA (gBRCA) Positive Breast Cancer	Phase 2|Phase 3	527	https://ClinicalTrials.gov/show/NCT03150576
2	Pre-operative Immunotherapy Combination Strategies in Breast Cancer	Phase 2	160	https://ClinicalTrials.gov/show/NCT03395899
3	A Pre-operative Window Study of Letrozole Plus PR Agonist (Megestrol Acetate) Versus Letrozole Alone in Post-menopausal Patients With ER-positive Breast Cancer	Phase 2	189	https://ClinicalTrials.gov/show/NCT03306472
4	ROS1 Targeting With Crizotinib in Advanced E-cadherin Negative, ER Positive Lobular Breast Cancer or Diffuse Gastric Cancer Study	Phase 2	58	https://ClinicalTrials.gov/show/NCT03620643
5	A Study Evaluating the Efficacy and Safety of GDC-0077 + Palbociclib + Fulvestrant vs Placebo + Palbociclib + Fulvestrant in Patients With PIK3CA-Mutant, Hormone Receptor-Positive, Her2-Negative, Locally Advanced or Metastatic Breast Cancer	Phase 2|Phase 3	400	https://ClinicalTrials.gov/show/NCT04191499
6	A Study of Ipatasertib in Combination With Paclitaxel as a Treatment for Participants With PIK3CA/AKT1/PTEN-Altered, Locally Advanced or Metastatic, Triple-Negative Breast Cancer or Hormone Receptor-Positive, HER2-Negative Breast Cancer	Phase 2|Phase 3	450	https://ClinicalTrials.gov/show/NCT03337724
7	DS-8201a in Pre-treated HER2 Breast Cancer That Cannot be Surgically Removed or Has Spread [DESTINY-Breast02]	Phase 3	600	https://ClinicalTrials.gov/show/NCT03523585
8	DS-8201a Versus T-DM1 for Human Epidermal Growth Factor Receptor 2 (HER2)-Positive, Unresectable and/or Metastatic Breast Cancer Previously Treated With Trastuzumab and Taxane [DESTINY-Breast03]	Phase 3	500	https://ClinicalTrials.gov/show/NCT03529110

No.	Title	Phase(s)	Enrolment	URL
9	Pilot Study of Cabazitaxel and Paclitaxel in HER2 Negative Breast Cancer	Phase 2	160	https://ClinicalTrials.gov/show/NCT03048942
10	The UK Plasma Based Molecular Profiling of Advanced Breast Cancer to Inform Therapeutic CHoices (plasmaMATCH) Trial	Phase 2	1150	https://ClinicalTrials.gov/show/NCT03182634
11	Phase 1 / 2 Study of SAR439859 Single Agent and in Combination With Palbociclib in Postmenopausal Women With Estrogen Receptor Positive Advanced Breast Cancer (AMEERA-1)	Phase 1|Phase 2	259	https://ClinicalTrials.gov/show/NCT03284957
12	Capivasertib+Fulvestrant vs Placebo+Fulvestrant as Treatment for Locally Advanced (Inoperable) or Metastatic HR+/HER2â^' Breast Cancer	Phase 3	834	https://ClinicalTrials.gov/show/NCT04305496
13	Study of Nivolumab Versus Placebo in Participants With High-Risk Breast Cancer	Phase 3	1200	https://ClinicalTrials.gov/show/NCT04109066
14	Phase 3 Trial of Elacestrant vs. Standard of Care for the Treatment of Patients With ER+/HER2- Advanced Breast Cancer	Phase 3	466	https://ClinicalTrials.gov/show/NCT03778931
15	A Trial to Evaluate Efficacy and Safety of Ribociclib With Endocrine Therapy as Adjuvant Treatment in Patients With HR+/HER2- Early Breast Cancer	Phase 3	4000	https://ClinicalTrials.gov/show/NCT03701334
16	Trastuzumab Deruxtecan (DS-8201a) Versus Investigator's Choice for HER2-low Breast Cancer That Has Spread or Cannot be Surgically Removed [DESTINY-Breast04]	Phase 3	540	https://ClinicalTrials.gov/show/NCT03734029
17	Study Assessing the Efficacy and Safety of Alpelisib Plus Fulvestrant or Letrozole, Based on Prior Endocrine Therapy, in Patients With PIK3CA Mutation With Advanced Breast Cancer Who Have Progressed on or After Prior Treatments	Phase 2	340	https://ClinicalTrials.gov/show/NCT03056755

No.	Title	Phase(s)	Enrolment	URL
18	Study of Pembrolizumab (MK-3475) Versus Placebo in Combination With Neoadjuvant Chemotherapy & Adjuvant Endocrine Therapy in the Treatment of Early-Stage Estrogen Receptor-Positive, Human Epidermal Growth Factor Receptor 2-Negative (ER+/HER2-) Breast Cancer (MK-3475-756/KEYNOTE-756)	Phase 3	1140	https://ClinicalTrials.gov/show/NCT03725059
19	A Study Of Ipatasertib in Combination With Atezolizumab and Paclitaxel as a Treatment for Participants With Locally Advanced or Metastatic Triple-Negative Breast Cancer.	Phase 3	1155	https://ClinicalTrials.gov/show/NCT04177108
20	MEN1611 With Trastuzumab (+/- Fulvestrant) in Metastatic Breast Cancer	Phase 1	48	https://ClinicalTrials.gov/show/NCT03767335
21	PHOENIX DDR/Anti-PD-L1 Trial: A Pre-surgical Window of Opportunity and Post-surgical Adjuvant Biomarker Study of DNA Damage Response Inhibition and/or Anti-PD-L1 Immunotherapy in Patients With Neoadjuvant Chemotherapy Resistant Residual Triple Negative Breast Cancer	Phase 2	81	https://ClinicalTrials.gov/show/NCT03740893
22	SYD985 vs. Physician's Choice in Participants With HER2-positive Locally Advanced or Metastatic Breast Cancer	Phase 3	345	https://ClinicalTrials.gov/show/NCT03262935
23	Study of AZD9833 Alone or in Combination With Palbociclib in Women With Advanced Breast Cancer	Phase 1	182	https://ClinicalTrials.gov/show/NCT03616587
24	A Study of Novel Anti-cancer Agents in Patients With Metastatic Triple Negative Breast Cancer	Phase 1|Phase 2	110	https://ClinicalTrials.gov/show/NCT03742102
25	A Study Comparing Atezolizumab (Anti PD-L1 Antibody) In Combination With Adjuvant Anthracycline/Taxane-Based Chemotherapy Versus Chemotherapy Alone In Patients With Operable Triple-Negative Breast Cancer	Phase 3	2300	https://ClinicalTrials.gov/show/NCT03498716
26	HR+/HER2- Advanced Breast Cancer and Endocrine Resistance	Phase 2	196	https://ClinicalTrials.gov/show/NCT03322215

No.	Title	Phase(s)	Enrolment	URL
27	A Study of Atezolizumab and Paclitaxel Versus Placebo and Paclitaxel in Participants With Previously Untreated Locally Advanced or Metastatic Triple Negative Breast Cancer (TNBC)	Phase 3	600	https://ClinicalTrials.gov/show/NCT03125902
28	A Study of GDC-9545 in Postmenopausal Women With Stage I-III Operable, Estrogen Receptor-Positive Breast Cancer	Phase 1	75	https://ClinicalTrials.gov/show/NCT03916744
29	A Study of Ipatasertib Plus Palbociclib and Fulvestrant Versus Placebo Plus Palbociclib and Fulvestrant in Hormone Receptor Positive and HER2 Negative Locally Advanced Unresectable or Metastatic Breast Cancer	Phase 3	370	https://ClinicalTrials.gov/show/NCT04060862
30	A Study to Evaluate the Safety and Efficacy of Ipatasertib in Combination With Atezolizumab and Paclitaxel or Nab-Paclitaxel in Participants With Locally Advanced or Metastatic Triple-Negative Breast Cancer	Phase 1	202	https://ClinicalTrials.gov/show/NCT03800836
31	A Study of GDC-9545 Alone or in Combination With Palbociclib and/or Luteinizing Hormone-Releasing Hormone (LHRH) Agonist in Locally Advanced or Metastatic Estrogen Receptor-Positive Breast Cancer	Phase 1	220	https://ClinicalTrials.gov/show/NCT03332797
32	G1T38, a CDK 4/6 Inhibitor, in Combination With Fulvestrant in Hormone Receptor-Positive, HER2-Negative Locally Advanced or Metastatic Breast Cancer	Phase 1|Phase 2	102	https://ClinicalTrials.gov/show/NCT02983071
33	A Study Evaluating the Efficacy and Safety of Multiple Immunotherapy-Based Treatment Combinations in Patients With Metastatic or Inoperable Locally Advanced Triple-Negative Breast Cancer	Phase 1|Phase 2	280	https://ClinicalTrials.gov/show/NCT03424005
34	A Study of the Efficacy and Safety of Atezolizumab Plus Chemotherapy for Patients With Early Relapsing Recurrent Triple-Negative Breast Cancer	Phase 3	572	https://ClinicalTrials.gov/show/NCT03371017
35	Adjuvant Palbociclib in Elderly Patients With Breast Cancer	Phase 2	366	https://ClinicalTrials.gov/show/NCT03609047

No.	Title	Phase(s)	Enrolment	URL
36	To Evaluate the Safety, Tolerability, and Pharmacokinetics of GDC-0077 Single Agent in Participants With Solid Tumors and in Combination With Endocrine and Targeted Therapies in Participants With Breast Cancer	Phase 1	104	https://ClinicalTrials.gov/show/NCT03006172
37	Study of IMMU-132 in HR+/HER2- MBC (TROPICS-02)	Phase 3	400	https://ClinicalTrials.gov/show/NCT03901339
38	Fulvestrant +/- Vandetanib in Advanced Aromatase Inhibitor Resistant Breast Cancer	Phase 2	160	https://ClinicalTrials.gov/show/NCT02530411
39	Pivotal Study in HER2 Negative, Locally Recurrent or Metastatic Breast Cancer	Phase 3	384	https://ClinicalTrials.gov/show/NCT03786094
40	To Assess Safety and Efficacy of Agents Targeting DNA Damage Repair With Olaparib Versus Olaparib Monotherapy.	Phase 2	350	https://ClinicalTrials.gov/show/NCT03330847
41	The XENERAâ„¢ 1 Study Tests Xentuzumab in Combination With Everolimus and Exemestane in Women With Hormone Receptor Positive and HER2-negative Breast Cancer That Has Spread	Phase 2	100	https://ClinicalTrials.gov/show/NCT03659136
42	PALbociclib Rechallenge in horMone Receptor-posItive/HER2-Negative Advanced Breast Cancer (PALMIRA)	Phase 2	198	https://ClinicalTrials.gov/show/NCT03809988
43	Capivasertib+Paclitaxel as First Line Treatment for Patients With Locally Advanced or Metastatic TNBC	Phase 3	800	https://ClinicalTrials.gov/show/NCT03997123
44	Study of Olaparib Plus Pembrolizumab Versus Chemotherapy Plus Pembrolizumab After Induction With First-Line Chemotherapy Plus Pembrolizumab in Triple Negative Breast Cancer (TNBC) (MK-7339-009/KEYLYNK-009)	Phase 2|Phase 3	932	https://ClinicalTrials.gov/show/NCT04191135
45	Ascending Doses of Ceralasertib in Combination With Chemotherapy and/or Novel Anti Cancer Agents	Phase 1|Phase 2	322	https://ClinicalTrials.gov/show/NCT02264678
46	A Study of PDR001 in Combination With LCL161, Everolimus or Panobinostat	Phase 1	315	https://ClinicalTrials.gov/show/NCT02890069

No.	Title	Phase(s)	Enrolment	URL
47	A Study Of PF-05212384 In Combination With Other Anti-Tumor Agents and in Combination With Cisplatin in Patients With Triple Negative Breast Cancer in an Expansion Arm (TNBC)	Phase 1	124	https://ClinicalTrials.gov/show/NCT01920061
48	Cambridge Brain Mets Trial 1	Phase 1|Phase 2	70	https://ClinicalTrials.gov/show/NCT02768337
49	GEN1029 (HexaBodyÂ®-DR5/DR5) Safety Trial in Patients With Malignant Solid Tumors	Phase 1|Phase 2	520	https://ClinicalTrials.gov/show/NCT03576131
50	Trastuzumab Deruxtecan With Nivolumab in Advanced Breast and Urothelial Cancer	Phase 1	99	https://ClinicalTrials.gov/show/NCT03523572
51	Clinical Trial to Evaluate the Safety and Effectiveness of GDC-0032 When Given Alongside Tamoxifen	Phase 1|Phase 2	290	https://ClinicalTrials.gov/show/NCT02285179
52	MRx0518 in Patients With Solid Tumours Waiting Surgical Removal of the Tumour	Phase 1	120	https://ClinicalTrials.gov/show/NCT03934827
53	A Study Evaluating Safety, Pharmacokinetics, and Therapeutic Activity of RO6874281 as a Single Agent (Part A) or in Combination With Trastuzumab or Cetuximab (Part B or C)	Phase 1	205	https://ClinicalTrials.gov/show/NCT02627274
54	A Study of RO7198457 as a Single Agent and in Combination With Atezolizumab in Participants With Locally Advanced or Metastatic Tumors	Phase 1	770	https://ClinicalTrials.gov/show/NCT03289962
55	GB1275 Monotherapy and in Combination With an Anti-PD1 Antibody in Patients With Specified Advanced Solid Tumors or in Combination With Standard of Care in Patients With Metastatic Pancreatic Adenocarcinoma	Phase 1|Phase 2	202	https://ClinicalTrials.gov/show/NCT04060342
56	Safety and Efficacy of KY1044 and Atezolizumab in Advanced Cancer	Phase 1|Phase 2	412	https://ClinicalTrials.gov/show/NCT03829501
57	A CR-UK Phase I Trial of LY3143921	Phase 1	68	https://ClinicalTrials.gov/show/NCT03096054
58	Add-Aspirin: A Trial Assessing the Effects of Aspirin on Disease Recurrence and Survival After Primary Therapy in Common Non Metastatic Solid Tumours	Phase 3	11000	https://ClinicalTrials.gov/show/NCT02804815

No.	Title	Phase(s)	Enrolment	URL
59	Study of Cabozantinib in Combination With Atezolizumab to Subjects With Locally Advanced or Metastatic Solid Tumors	Phase 1|Phase 2	1732	https://ClinicalTrials.gov/show/NCT03170960
60	A Phase I/Ib Study of NZV930 Alone and in Combination With PDR001 and /or NIR178 in Patients With Advanced Malignancies.	Phase 1	344	https://ClinicalTrials.gov/show/NCT03549000
61	A Dose Escalation and Cohort Expansion Study of NKTR-214 in Combination With Nivolumab and Other Anti-Cancer Therapies in Patients With Select Advanced Solid Tumors (PIVOT-02)	Phase 1|Phase 2	780	https://ClinicalTrials.gov/show/NCT02983045
62	Javelin Parp Medley: Avelumab Plus Talazoparib In Locally Advanced Or Metastatic Solid Tumors	Phase 2	242	https://ClinicalTrials.gov/show/NCT03330405
63	Basket Study of Entrectinib (RXDX-101) for the Treatment of Patients With Solid Tumors Harboring NTRK 1/2/3 (Trk A/B/C), ROS1, or ALK Gene Rearrangements (Fusions)	Phase 2	300	https://ClinicalTrials.gov/show/NCT02568267
64	PROCLAIM-CX-2009: A Trial to Find Safe and Active Doses of an Investigational Drug CX-2009 for Patients With Selected Solid Tumors	Phase 1|Phase 2	150	https://ClinicalTrials.gov/show/NCT03149549
65	Phase 1b Multi-indication Study of Anetumab Ravtansine in Mesothelin Expressing Advanced Solid Tumors	Phase 1	348	https://ClinicalTrials.gov/show/NCT03102320
66	Phase 1/2 Study of LOXO-292 in Patients With Advanced Solid Tumors, RET Fusion-Positive Solid Tumors, and Medullary Thyroid Cancer	Phase 1|Phase 2	970	https://ClinicalTrials.gov/show/NCT03157128
67	Study to Assess Safety, Tolerability and Clinical Activity of BGB-290 in Combination With Temozolomide (TMZ) in Participants With Locally Advanced or Metastatic Solid Tumors	Phase 1|Phase 2	150	https://ClinicalTrials.gov/show/NCT03150810
68	A Study to Test the Effect of the Drug Larotrectinib in Adults and Children With NTRK-fusion Positive Solid Tumors	Phase 2	203	https://ClinicalTrials.gov/show/NCT02576431
69	A Study Of Avelumab In Combination With Other Cancer Immunotherapies In Advanced Malignancies (JAVELIN Medley)	Phase 2	620	https://ClinicalTrials.gov/show/NCT02554812

No.	Title	Phase(s)	Enrolment	URL
70	Study Of Entrectinib (Rxdx-101) in Children and Adolescents With Locally Advanced Or Metastatic Solid Or Primary CNS Tumors And/Or Who Have No Satisfactory Treatment Options	Phase 1|Phase 2	68	https://ClinicalTrials.gov/show/NCT02650401
71	Study of Intratumorally Administered Stimulator of Interferon Genes (STING) Agonist E7766 in Participants With Advanced Solid Tumors or Lymphomas - INSTAL-101	Phase 1	120	https://ClinicalTrials.gov/show/NCT04144140
72	A Study of the CD73 Inhibitor LY3475070 Alone or in Combination With Pembrolizumab in Participants With Advanced Cancer	Phase 1	120	https://ClinicalTrials.gov/show/NCT04148937
Colon cancer trials				
1	Study of Intrahepatic Arterial Infusion of TG6002 in Combination With 5-FC in Patients With Metastatic Colorectal Cancer	Phase 1|Phase 2	75	https://ClinicalTrials.gov/show/NCT04194034
2	iSCORE: Immunotherapy Sequencing in COlon and REctal Cancer	Phase 2	25	https://ClinicalTrials.gov/show/NCT03867799
3	Acoustic Cluster Therapy (ACT) With Chemotherapy in Metastatic Liver Metastases of Gastrointestinal Origin	Phase 1	37	https://ClinicalTrials.gov/show/NCT04021277
4	A Safety Study of NUC-3373 in Combination With Standard Agents Used in Colorectal Cancer Treatment	Phase 1	62	https://ClinicalTrials.gov/show/NCT03428958
5	alloSHRINK - Standard cHemotherapy Regimen and Immunotherapy With Allogeneic NKG2D-based CYAD-101 Chimeric Antigen Receptor T-cells	Phase 1	36	https://ClinicalTrials.gov/show/NCT03692429
6	A Safety and Effectiveness Study of Pre-operative Artesunate in Stage II/III Colorectal Cancer	Phase 2	200	https://ClinicalTrials.gov/show/NCT02633098
7	A Phase II Trial Assessing Nivolumab in Strong Class II Expressing Microsatellite Stable Colorectal Cancer	Phase 2	36	https://ClinicalTrials.gov/show/NCT03981146
8	A Study Evaluating the Efficacy and Safety of Multiple Immunotherapy-Based Treatment Combinations in Patients With Metastatic Colorectal Cancer (Morpheus-CRC)	Phase 1|Phase 2	326	https://ClinicalTrials.gov/show/NCT03555149

No.	Title	Phase(s)	Enrolment	URL
9	Phase III Study in First-line Treatment of Patients With Metastatic Colorectal Cancer Who Are Not Candidate for Intensive Therapy.	Phase 3	854	https://ClinicalTrials.gov/show/NCT03869892
10	A Study of Nivolumab, Nivolumab Plus Ipilimumab, or Investigator's Choice Chemotherapy for the Treatment of Patients With Deficient Mismatch Repair (dMMR)/Microsatellite Instability High (MSI-H) Metastatic Colorectal Cancer (mCRC)	Phase 3	494	https://ClinicalTrials.gov/show/NCT04008030
11	GEN1042 Safety Trial in Subjects With Malignant Solid Tumors	Phase 1|Phase 2	126	https://ClinicalTrials.gov/show/NCT04083599
12	A Study of PDR001 in Combination With LCL161, Everolimus or Panobinostat	Phase 1	315	https://ClinicalTrials.gov/show/NCT02890069
13	Gevokizumab With Standard of Care Anti-cancer Therapies for Metastatic Colorectal, Gastroesophageal, and Renal Cancers	Phase 1	172	https://ClinicalTrials.gov/show/NCT03798626
14	EPA for Metastasis Trial 2	Phase 3	448	https://ClinicalTrials.gov/show/NCT03428477
15	PIPAC for the Treatment of Colorectal Peritoneal Metastases	Phase 2	30	https://ClinicalTrials.gov/show/NCT03868228
16	A CR-UK Phase I Trial of LY3143921	Phase 1	68	https://ClinicalTrials.gov/show/NCT03096054
17	Add-Aspirin: A Trial Assessing the Effects of Aspirin on Disease Recurrence and Survival After Primary Therapy in Common Non Metastatic Solid Tumours	Phase 3	11000	https://ClinicalTrials.gov/show/NCT02804815
18	A Dose Escalation and Cohort Expansion Study of NKTR-214 in Combination With Nivolumab and Other Anti-Cancer Therapies in Patients With Select Advanced Solid Tumors (PIVOT-02)	Phase 1|Phase 2	780	https://ClinicalTrials.gov/show/NCT02983045
19	GEN1029 (HexaBodyÂ®-DR5/DR5) Safety Trial in Patients With Malignant Solid Tumors	Phase 1|Phase 2	520	https://ClinicalTrials.gov/show/NCT03576131
20	A Phase 1/2 Study of INCB001158 in Combination With Chemotherapy in Subjects With Solid Tumors	Phase 1|Phase 2	249	https://ClinicalTrials.gov/show/NCT03314935
21	Basket Study of Entrectinib (RXDX-101) for the Treatment of Patients With Solid Tumors Harboring NTRK 1/2/3 (Trk A/B/C), ROS1, or ALK Gene Rearrangements (Fusions)	Phase 2	300	https://ClinicalTrials.gov/show/NCT02568267

No.	Title	Phase(s)	Enrolment	URL
22	A Study of RO7198457 as a Single Agent and in Combination With Atezolizumab in Participants With Locally Advanced or Metastatic Tumors	Phase 1	770	https://ClinicalTrials.gov/show/NCT03289962
23	GB1275 Monotherapy and in Combination With an Anti-PD1 Antibody in Patients With Specified Advanced Solid Tumors or in Combination With Standard of Care in Patients With Metastatic Pancreatic Adenocarcinoma	Phase 1|Phase 2	202	https://ClinicalTrials.gov/show/NCT04060342
24	Study of Cabozantinib in Combination With Atezolizumab to Subjects With Locally Advanced or Metastatic Solid Tumors	Phase 1|Phase 2	1732	https://ClinicalTrials.gov/show/NCT03170960
25	A Phase I/Ib Study of NZV930 Alone and in Combination With PDR001 and /or NIR178 in Patients With Advanced Malignancies.	Phase 1	344	https://ClinicalTrials.gov/show/NCT03549000
26	A Study to Test the Effect of the Drug Larotrectinib in Adults and Children With NTRK-fusion Positive Solid Tumors	Phase 2	203	https://ClinicalTrials.gov/show/NCT02576431
27	Efficacy and Safety Study of Tisotumab Vedotin for Patients With Solid Tumors	Phase 2	250	https://ClinicalTrials.gov/show/NCT03485209
28	A First-in-Humans Dose Finding Study for an Aryl Hydrocarbon Receptor Inhibitor (AhRi) in Patients With Advanced Cancer	Phase 1	114	https://ClinicalTrials.gov/show/NCT04069026
29	Study of Intratumorally Administered Stimulator of Interferon Genes (STING) Agonist E7766 in Participants With Advanced Solid Tumors or Lymphomas - INSTAL-101	Phase 1	120	https://ClinicalTrials.gov/show/NCT04144140
30	Study of Pembrolizumab (MK-3475) in Participants With Advanced Solid Tumors (MK-3475-158/KEYNOTE-158)	Phase 2	1395	https://ClinicalTrials.gov/show/NCT02628067
Lung cancer trials				
1	SBRT With Immunotherapy in Early Stage Non-small Cell Lung Cancer: Tolerability and Lung Effects	Phase 1|Phase 2	31	https://ClinicalTrials.gov/show/NCT03383302
2	AST-VAC2 Vaccine in Patients With Non-small Cell Lung Cancer	Phase 1	48	https://ClinicalTrials.gov/show/NCT03371485
3	Hyperpolarized Xenon Gas MR Imaging in NSCLC Radiotherapy	Phase 2	50	https://ClinicalTrials.gov/show/NCT02151604

No.	Title	Phase(s)	Enrolment	URL
4	Targeted Stem Cells Expressing TRAIL as a Therapy for Lung Cancer	Phase 1|Phase 2	46	https://ClinicalTrials.gov/show/NCT03298763
5	Deciphering Antitumour Response and Resistance With INtratumour Heterogeneity	Phase 2	119	https://ClinicalTrials.gov/show/NCT02314481
6	STUDY 15 - Comparing Gemcitabine/Carboplatin and Hydroxychloroquine Versus Carboplatin/Etoposide Therapy Alone in Small Cell Lung Cancer (SCLC)	Phase 2	112	https://ClinicalTrials.gov/show/NCT02722369
7	ATL001 in Patients With Advanced Unresectable or Metastatic NSCLC	Phase 1|Phase 2	50	https://ClinicalTrials.gov/show/NCT04032847
8	A Study of Atezolizumab Plus Carboplatin and Etoposide With or Without Tiragolumab in Patients With Untreated Extensive-Stage Small Cell Lung Cancer	Phase 3	400	https://ClinicalTrials.gov/show/NCT04256421
9	A Study of Osimertinib With or Without Chemotherapy as 1st Line Treatment in Patients With Mutated Epidermal Growth Factor Receptor Non-Small Cell Lung Cancer (FLAURA2)	Phase 3	586	https://ClinicalTrials.gov/show/NCT04035486
10	A Study to Determine Safety of Durvalumab After Sequential Chemo Radiation in Patients With Unresectable Stage III Non-Small Cell Lung Cancer	Phase 2	150	https://ClinicalTrials.gov/show/NCT03693300
11	Efficacy and Safety Study of Stereotactic Body Radiotherapy (SBRT) With or Without Pembrolizumab (MK-3475) in Adults With Medically Inoperable Stage I or IIA Non-Small Cell Lung Cancer (NSCLC) (MK-3475-867/KEYNOTE-867)	Phase 3	530	https://ClinicalTrials.gov/show/NCT03924869
12	Brief Title: Study of Efficacy and Safety of Canakinumab as Adjuvant Therapy in Adult Subjects With Stages AJCC/UICC v. 8 II-IIIA and IIIB (T>5cm N2) Completely Resected Non-small Cell Lung Cancer Acronym: CANOPY-A	Phase 3	1500	https://ClinicalTrials.gov/show/NCT03447769

No.	Title	Phase(s)	Enrolment	URL
13	Study of Pemetrexed + Platinum Chemotherapy With or Without Pembrolizumab (MK-3475) in Adults With Tyrosine Kinase Inhibitor-(TKI)-Resistant Epidermal Growth Factor Receptor- (EGFR)-Mutated Metastatic Non-squamous Non-small Cell Lung Cancer (NSCLC) (MK-3475-789/KEYNOTE-789)	Phase 3	480	https://ClinicalTrials.gov/show/NCT03515837
14	An Investigational Immuno-therapy Trial of Nivolumab, or Nivolumab Plus Ipilimumab, or Nivolumab Plus Platinum-doublet Chemotherapy, Compared to Platinum Doublet Chemotherapy in Patients With Stage IV Non-Small Cell Lung Cancer (NSCLC)	Phase 3	2220	https://ClinicalTrials.gov/show/NCT02477826
15	A Study of Nivolumab and Ipilimumab in Untreated Patients With Stage 3 Non-small Cell Lung Cancer (NSCLC) That is Unable or Not Planned to be Removed by Surgery	Phase 3	1400	https://ClinicalTrials.gov/show/NCT04026412
16	Safety and Efficacy Study of Pemetrexed + Platinum Chemotherapy + Pembrolizumab (MK-3475) With or Without Lenvatinib (MK-7902/E7080) as First-line Intervention in Adults With Metastatic Nonsquamous Non-small Cell Lung Cancer (MK-7902-006/E7080-G000-315/LEAP-006)	Phase 3	726	https://ClinicalTrials.gov/show/NCT03829319
17	Efficacy and Safety of Pembrolizumab (MK-3475) With Platinum Doublet Chemotherapy as Neoadjuvant/Adjuvant Therapy for Participants With Resectable Stage II, IIIA, and Resectable IIIB (T3-4N2) Non-small Cell Lung Cancer (MK-3475-671/KEYNOTE-671)	Phase 3	786	https://ClinicalTrials.gov/show/NCT03425643
18	A Study of LY3295668 Erbumine in Participants With Extensive-stage Small-Cell Lung Cancer	Phase 1|Phase 2	64	https://ClinicalTrials.gov/show/NCT03898791
19	Efficacy and Safety of Pembrolizumab (MK-3475) With Lenvatinib (E7080/MK-7902) vs. Docetaxel in Participants With Metastatic Non-Small Cell Lung Cancer (NSCLC) and Progressive Disease (PD) After Platinum Doublet Chemotherapy and Immunotherapy (MK-7902-008/E7080-G000-316/LEAP-008)	Phase 3	405	https://ClinicalTrials.gov/show/NCT03976375

No.	Title	Phase(s)	Enrolment	URL
20	Study of JNJ-61186372, a Human Bispecific EGFR and cMet Antibody, in Participants With Advanced Non-Small Cell Lung Cancer	Phase 1	460	https://ClinicalTrials.gov/show/NCT02609776
21	A Study of Biomarker-Directed, Pembrolizumab (MK-3475) Based Combination Therapy for Advanced Non-Small Cell Lung Cancer (MK-3475-495/KEYNOTE-495)	Phase 2	318	https://ClinicalTrials.gov/show/NCT03516981
22	A Study to Investigate the Pharmacokinetics, Efficacy, and Safety of Atezolizumab Subcutaneous in Patients With Stage IV Non-Small Cell Lung Cancer (IMscin001)	Phase 1|Phase 2	260	https://ClinicalTrials.gov/show/NCT03735121
23	National Lung Matrix Trial: Multi-drug Phase II Trial in Non-Small Cell Lung Cancer	Phase 2	569	https://ClinicalTrials.gov/show/NCT02664935
24	Study of Telisotuzumab Vedotin (ABBV-399) in Subjects With Previously Treated c-Met+ Non-Small Cell Lung Cancer	Phase 2	310	https://ClinicalTrials.gov/show/NCT03539536
25	Study of OSE2101 Versus Standard Treatment as 2nd or 3rd Line in HLA-A2 Positive Patients With Advanced NSCLC After Failure of Immune Checkpoint Inhibitor	Phase 3	363	https://ClinicalTrials.gov/show/NCT02654587
26	A Study of Selpercatinib (LY3527723) in Participants With Advanced or Metastatic RET Fusion-Positive Non-Small Cell Lung Cancer	Phase 3	400	https://ClinicalTrials.gov/show/NCT04194944
27	A Trial of Pembrolizumab in Combination With Chemotherapy and Radiotherapy in Stage III NSCLC (KEYNOTE-799, MK-3475-799).	Phase 2	216	https://ClinicalTrials.gov/show/NCT03631784
28	Phase 1/2 Study of the Highly-selective RET Inhibitor, Pralsetinib (BLU-667), in Patients With Thyroid Cancer, Non-Small Cell Lung Cancer, and Other Advanced Solid Tumors	Phase 1|Phase 2	527	https://ClinicalTrials.gov/show/NCT03037385
29	Study of Pembrolizumab With Maintenance Olaparib or Maintenance Pemetrexed in First-line (1L) Metastatic Nonsquamous Non-Small-Cell Lung Cancer (NSCLC) (MK-7339-006, KEYLYNK-006)	Phase 3	792	https://ClinicalTrials.gov/show/NCT03976323
30	BT1718 in Patients With Advanced Solid Tumours.	Phase 1|Phase 2	130	https://ClinicalTrials.gov/show/NCT03486730

No.	Title	Phase(s)	Enrolment	URL
31	Durvalumab vs Placebo Following Stereotactic Body Radiation Therapy in Early Stage Unresected Non-small Cell Lung Cancer Patients	Phase 3	706	https://ClinicalTrials.gov/show/NCT03833154
32	First-in-human Study of S-588210 (S-488210+S-488211)	Phase 1	10	https://ClinicalTrials.gov/show/NCT04316689
33	Clinical Study of Oral cMET Inhibitor INC280 in Adult Patients With EGFR Wild-type Advanced Non-small Cell Lung Cancer	Phase 2	373	https://ClinicalTrials.gov/show/NCT02414139
34	Deciphering Afatinib Response and Resistance With INtratumour Heterogeneity	Phase 2	48	https://ClinicalTrials.gov/show/NCT02183883
35	Safety And Efficacy Study Of Avelumab Plus Chemotherapy With Or Without Other Anti-Cancer Immunotherapy Agents In Patients With Advanced Malignancies	Phase 2	80	https://ClinicalTrials.gov/show/NCT03317496
36	Study to Test the Safety and How Radium-223 Dichloride an Alpha Particle-emitting Radioactive Agent Works in Combination With Pembrolizumab an Immune Checkpoint Inhibitor in Patients With Stage IV Non-small Cell Lung Cancer With Bone Metastases	Phase 1|Phase 2	164	https://ClinicalTrials.gov/show/NCT03996473
37	Study Evaluating Safety, Tolerability and PK of AMG 757 in Adults With Small Cell Lung Cancer	Phase 1	162	https://ClinicalTrials.gov/show/NCT03319940
38	A Study Comparing Adjuvant Alectinib Versus Adjuvant Platinum-Based Chemotherapy in Patients With ALK Positive Non-Small Cell Lung Cancer	Phase 3	255	https://ClinicalTrials.gov/show/NCT03456076
39	A Study of Neoadjuvant Atezolizumab Plus Chemotherapy Versus Placebo Plus Chemotherapy in Patients With Resectable Stage II, IIIA, or Select IIIB Non-Small Cell Lung Cancer (IMpower030)	Phase 3	374	https://ClinicalTrials.gov/show/NCT03456063

No.	Title	Phase(s)	Enrolment	URL
40	A Study of Neoadjuvant Chemotherapy Plus Nivolumab Versus Neoadjuvant Chemotherapy Plus Placebo, Followed by Surgical Removal and Adjuvant Treatment With Nivolumab or Placebo for Participants With Surgically Removable Early Stage Non-small Cell Lung Cancer	Phase 3	452	https://ClinicalTrials.gov/show/NCT04025879
41	A Study of Pembrolizumab (MK-3475) With or Without Maintenance Olaparib in First-line Metastatic Squamous Non-small Cell Lung Cancer (NSCLC, MK-7339-008/KEYLYNK-008)	Phase 3	735	https://ClinicalTrials.gov/show/NCT03976362
42	A Study Of Multiple Immunotherapy-Based Treatment Combinations In Participants With Metastatic Non-Small Cell Lung Cancer (Morpheus- Non-Small Cell Lung Cancer)	Phase 1|Phase 2	305	https://ClinicalTrials.gov/show/NCT03337698
43	GEN1042 Safety Trial in Subjects With Malignant Solid Tumors	Phase 1|Phase 2	126	https://ClinicalTrials.gov/show/NCT04083599
44	Cambridge Brain Mets Trial 1	Phase 1|Phase 2	70	https://ClinicalTrials.gov/show/NCT02768337
45	A Study Of Lorlatinib Versus Crizotinib In First Line Treatment Of Patients With ALK-Positive NSCLC	Phase 3	280	https://ClinicalTrials.gov/show/NCT03052608
46	MRx0518 in Patients With Solid Tumours Waiting Surgical Removal of the Tumour	Phase 1	120	https://ClinicalTrials.gov/show/NCT03934827
47	A Study of PDR001 in Combination With LCL161, Everolimus or Panobinostat	Phase 1	315	https://ClinicalTrials.gov/show/NCT02890069
48	A Study to Determine Safety, Tolerability, Pharmacokinetics, and Recommended Phase 2 Dose (RP2D) of Intravenous ABBV-184 in Adult Participants With Previously Treated Cancers	Phase 1	112	https://ClinicalTrials.gov/show/NCT04272203
49	Bemcentinib (BGB324) in Combination With Pembrolizumab in Patients With Advanced NSCLC	Phase 2	77	https://ClinicalTrials.gov/show/NCT03184571

No.	Title	Phase(s)	Enrolment	URL
50	Combination Study With Soluble LAG-3 Fusion Protein Eftilagimod Alpha (IMP321) and Pembrolizumab in Patients With Previously Untreated Unresectable or Metastatic NSCLC, or Recurrent PD-X Refractory NSCLC or With Recurrent or Metastatic HNSCC	Phase 2	109	https://ClinicalTrials.gov/show/NCT03625323
51	Nintedanib as Switch Maintenance Treatment of Pleural Malignant Mesothelioma	Phase 2	116	https://ClinicalTrials.gov/show/NCT02863055
52	Stereotactic Body Radiotherapy for the Treatment of OPD	Phase 2|Phase 3	110	https://ClinicalTrials.gov/show/NCT03256981
53	Study of Autologous Tumor Infiltrating Lymphocytes in Patients With Solid Tumors	Phase 2	75	https://ClinicalTrials.gov/show/NCT03645928
54	Phase 1/2 Study of LOXO-292 in Patients With Advanced Solid Tumors, RET Fusion-Positive Solid Tumors, and Medullary Thyroid Cancer	Phase 1|Phase 2	970	https://ClinicalTrials.gov/show/NCT03157128
55	A Dose Escalation and Cohort Expansion Study of NKTR-214 in Combination With Nivolumab and Other Anti-Cancer Therapies in Patients With Select Advanced Solid Tumors (PIVOT-02)	Phase 1|Phase 2	780	https://ClinicalTrials.gov/show/NCT02983045
56	This Study Tests the New Medicine BI 754111 Alone or in Combination With Another New Substance BI 754091 in Patients With Advanced Cancer. The Study Tests Different Doses to Find the Best Dose for Continuous Treatment.	Phase 1	215	https://ClinicalTrials.gov/show/NCT03156114
57	A CR-UK Phase I Trial of LY3143921	Phase 1	68	https://ClinicalTrials.gov/show/NCT03096054
58	Basket Study of Entrectinib (RXDX-101) for the Treatment of Patients With Solid Tumors Harboring NTRK 1/2/3 (Trk A/B/C), ROS1, or ALK Gene Rearrangements (Fusions)	Phase 2	300	https://ClinicalTrials.gov/show/NCT02568267
59	GEN1029 (HexaBodyÂ®-DR5/DR5) Safety Trial in Patients With Malignant Solid Tumors	Phase 1|Phase 2	520	https://ClinicalTrials.gov/show/NCT03576131
60	A Trial of BXCL701 and Pembrolizumab in Patients With Small Cell Neuroendocrine Prostate Cancer	Phase 1|Phase 2	40	https://ClinicalTrials.gov/show/NCT03910660

No.	Title	Phase(s)	Enrolment	URL
61	Enapotamab Vedotin (HuMax-AXL-ADC) Safety Study in Patients With Solid Tumors	Phase 1|Phase 2	374	https://ClinicalTrials.gov/show/NCT02988817
62	PROCLAIM-CX-2009: A Trial to Find Safe and Active Doses of an Investigational Drug CX-2009 for Patients With Selected Solid Tumors	Phase 1|Phase 2	150	https://ClinicalTrials.gov/show/NCT03149549
63	A Study of RO7198457 as a Single Agent and in Combination With Atezolizumab in Participants With Locally Advanced or Metastatic Tumors	Phase 1	770	https://ClinicalTrials.gov/show/NCT03289962
64	Safety and Efficacy of KY1044 and Atezolizumab in Advanced Cancer	Phase 1|Phase 2	412	https://ClinicalTrials.gov/show/NCT03829501
65	Study of FAK (Defactinib) and PD-1 (Pembrolizumab) Inhibition in Advanced Solid Malignancies (FAK-PD1)	Phase 1|Phase 2	59	https://ClinicalTrials.gov/show/NCT02758587
66	Study of Cabozantinib in Combination With Atezolizumab to Subjects With Locally Advanced or Metastatic Solid Tumors	Phase 1|Phase 2	1732	https://ClinicalTrials.gov/show/NCT03170960
67	A Phase I/Ib Study of NZV930 Alone and in Combination With PDR001 and /or NIR178 in Patients With Advanced Malignancies.	Phase 1	344	https://ClinicalTrials.gov/show/NCT03549000
68	Preoperative Paravertebral Block in Cancer Surgery of the Lung	Phase 2|Phase 3	100	https://ClinicalTrials.gov/show/NCT04209868
69	A Study of Repotrectinib (TPX-0005) in Patients With Advanced Solid Tumors Harboring ALK, ROS1, or NTRK1-3 Rearrangements	Phase 1|Phase 2	450	https://ClinicalTrials.gov/show/NCT03093116
70	A Study Of Avelumab In Combination With Other Cancer Immunotherapies In Advanced Malignancies (JAVELIN Medley)	Phase 2	620	https://ClinicalTrials.gov/show/NCT02554812
71	Phase 1b Multi-indication Study of Anetumab Ravtansine in Mesothelin Expressing Advanced Solid Tumors	Phase 1	348	https://ClinicalTrials.gov/show/NCT03102320
72	Study to Assess Safety, Tolerability and Clinical Activity of BGB-290 in Combination With Temozolomide (TMZ) in Participants With Locally Advanced or Metastatic Solid Tumors	Phase 1|Phase 2	150	https://ClinicalTrials.gov/show/NCT03150810
73	A Study to Test the Effect of the Drug Larotrectinib in Adults and Children With NTRK-fusion Positive Solid Tumors	Phase 2	203	https://ClinicalTrials.gov/show/NCT02576431

No.	Title	Phase(s)	Enrolment	URL
74	Durvalumab Long-Term Safety and Efficacy Study	Phase 4	600	https://ClinicalTrials.gov/show/NCT04078152
75	Study of TSR-042, an Anti-programmed Cell Death-1 Receptor (PD-1) Monoclonal Antibody, in Participants With Advanced Solid Tumors	Phase 1	740	https://ClinicalTrials.gov/show/NCT02715284
76	Efficacy and Safety Study of Tisotumab Vedotin for Patients With Solid Tumors	Phase 2	250	https://ClinicalTrials.gov/show/NCT03485209
77	Study of Pembrolizumab (MK-3475) in Participants With Advanced Solid Tumors (MK-3475-158/KEYNOTE-158)	Phase 2	1395	https://ClinicalTrials.gov/show/NCT02628067
78	A First-in-Humans Dose Finding Study for an Aryl Hydrocarbon Receptor Inhibitor (AhRi) in Patients With Advanced Cancer	Phase 1	114	https://ClinicalTrials.gov/show/NCT04069026
79	A Study of the CD73 Inhibitor LY3475070 Alone or in Combination With Pembrolizumab in Participants With Advanced Cancer	Phase 1	120	https://ClinicalTrials.gov/show/NCT04148937
80	A Safety, Tolerability and PK Study of DCC-2618 in Patients With Advanced Malignancies	Phase 1	320	https://ClinicalTrials.gov/show/NCT02571036

Appendix 7: Output of all calculations using 'nstage' command with NI hypothesis

DEVA_NI Scenario 1

```
nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 3) t(5 5) s(0.818 0.818) tstop(8)
```

```
n-stage trial design                          version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------
Sample size for a 3-arm 2-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 17.3 time units

Operating characteristics
```

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.5000	0.950	1.188	1.000	**1.188**	**5.912**	**5.912**
2	0.0250	0.900	1.188	1.000	**1.070**	**7.615**	**13.527**

Pairwise Error Rate	**0.0228**	Pairwise Power	**0.8702**
Familywise Error Rate (SE)	**0.0420** (**0.0004**)		

```
Note: patient accrual stopped at time  8.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ----------Stage 1----------
           Overall  Control   Exper.
Arms             3        1        2
Acc. rate      845      282      563
Patients*     4995     1665     3330
Events**       551      183      368
           ----------Stage 2----------
           Overall  Control   Exper.
Arms             3        1        2
Acc. rate      845      282      563
Patients*     6760     2253     4507
Events**      2132      710     1422
           ------------------------------
```

```
nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 2) t(5 5) s(0.818 0.818) tstop(8)

n-stage trial design                        version 4.0.1, 2 Nov 2018
─────────────────────────────────────────────────────────────────────────
Sample size for a 3-arm 2-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
─────────────────────────────────────────────────────────────────────────

Note: I outcome and D outcome are identical
Median survival time: 17.3 time units

Operating characteristics
```

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.5000	0.950	1.188	1.000	**1.188**	**5.912**	**5.912**
2	0.0250	0.900	1.188	1.000	**1.070**	**6.674**	**12.586**

Pairwise Error Rate	**0.0228**	Pairwise Power	**0.8702**
Familywise Error Rate (SE)	**0.0413** (**0.0004**)		

```
Note: patient accrual stopped at time  8.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
```

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	845	282	563
Patients*	**4995**	**1665**	**3330**
Events**	**551**	**183**	**368**
	Stage 2		
	Overall	Control	Exper.
Arms	2	1	1
Acc. rate	845	423	423
Patients*	**6760**	**2547**	**4213**
Events**	**1421**	**710**	**711**

DEVA_NI Scenario 2

```
nstage, nstage(2) alpha(0.5 0.05) omega(0.9 0.9) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 3) t(5 5) s(0.818 0.818) tstop(8)
```

```
n-stage trial design                        version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 3-arm 2-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 17.3 time units

Operating characteristics
```

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.5000	0.901	1.188	1.000	**1.188**	**4.564**	**4.564**
2	0.0500	0.900	1.188	1.000	**1.078**	**6.935**	**11.499**

Pairwise Error Rate	**0.0419**	Pairwise Power	**0.8296**
Familywise Error Rate (SE)	**0.0753 (0.0005)**		

```
Note: patient accrual stopped at time  8.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
```

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	845	282	563
Patients*	**3857**	**1286**	**2571**
Events**	**335**	**111**	**224**
	Stage 2		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	845	282	563
Patients*	**6760**	**2253**	**4507**
Events**	**1739**	**579**	**1160**

```
nstage, nstage(2) alpha(0.5 0.05) omega(0.9 0.9) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 2) t(5 5) s(0.818 0.818) tstop(8)
```

n-stage trial design version 4.0.1, 2 Nov 2018

Sample size for a 3-arm 2-stage trial with time-to-event outcome based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al. (2019) Clinical Trials 16(2)

Note: I outcome and D outcome are identical
Median survival time: **17.3** time units

Operating characteristics

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.5000	0.901	1.188	1.000	**1.188**	**4.564**	**4.564**
2	0.0500	0.900	1.188	1.000	**1.078**	**5.860**	**10.425**

Pairwise Error Rate	**0.0419**	Pairwise Power	**0.8296**
Familywise Error Rate (SE)	**0.0743 (0.0005)**		

Note: patient accrual stopped at time 8.000
* All alphas are one-sided
** Length (duration of each stage) is expressed in periods and assumes survival times are exponentially distributed. Time is expressed in cumulative periods.

Sample size and number of events

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	845	282	563
Patients*	**3857**	**1286**	**2571**
Events**	**335**	**111**	**224**
	Stage 2		
	Overall	Control	Exper.
Arms	2	1	1
Acc. rate	845	423	423
Patients*	**6760**	**2737**	**4023**
Events**	**1159**	**579**	**580**

DEVA_NI Scenario 3

```
nstage, nstage(2) alpha(0.25 0.05) omega(0.9 0.85) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 3) t(5 5) s(0.818 0.818) tstop(8)
```

```
n-stage trial design                          version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------

Sample size for a 3-arm 2-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 17.3 time units

Operating characteristics
-------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
-------------------------------------------------------------------------
1           0.2500    0.900    1.188    1.000    1.119    7.086    7.086
2           0.0500    0.850    1.188    1.000    1.069    3.067   10.154
-------------------------------------------------------------------------
Pairwise Error Rate          0.0436              Pairwise Power   0.8130
Familywise Error Rate (SE)   0.0778 (0.0005)
-------------------------------------------------------------------------
Note: patient accrual stopped at time  8.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
                ----------Stage 1----------
            Overall  Control   Exper.
Arms              3        1        2
Acc. rate       845      282      563
Patients*      5988     1996     3992
Events**        777      259      518
                ----------Stage 2----------
            Overall  Control   Exper.
Arms              3        1        2
Acc. rate       845      282      563
Patients*      6760     2253     4507
Events**       1460      486      974
               ------------------------
```

```
nstage, nstage(2) alpha(0.25 0.05) omega(0.9 0.85) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 2) t(5 5) s(0.818 0.818) tstop(8)
```

n-stage trial design version 4.0.1, 2 Nov 2018

Sample size for a 3-arm 2-stage trial with time-to-event outcome based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al. (2019) Clinical Trials 16(2)

Note: I outcome and D outcome are identical
Median survival time: **17.3** time units

Operating characteristics

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.2500	0.900	1.188	1.000	**1.119**	**7.086**	**7.086**
2	0.0500	0.850	1.188	1.000	**1.069**	**2.898**	**9.985**

Pairwise Error Rate	**0.0436**	Pairwise Power	**0.8130**
Familywise Error Rate (SE)	**0.0784** (**0.0005**)		

Note: patient accrual stopped at time 8.000
* All alphas are one-sided
** Length (duration of each stage) is expressed in periods and assumes survival times are exponentially distributed. Time is expressed in cumulative periods.

Sample size and number of events

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	845	282	563
Patients*	**5988**	**1996**	**3992**
Events**	**777**	**259**	**518**
	Stage 2		
	Overall	Control	Exper.
Arms	2	1	1
Acc. rate	845	423	423
Patients*	**6760**	**2382**	**4378**
Events**	**973**	**486**	**487**

QUASAR 2_NI Scenario 1

```
nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1.3046243 1.3046243)
hr1(1 1) accrue(579 579) arms(3 3) t(3 3) s(0.894 0.894) tstop(5.5)
```

```
n-stage trial design                          version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------
Sample size for a 3-arm 2-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 18.6 time units

Operating characteristics
```

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.5000	0.951	1.305	1.000	**1.305**	**4.759**	**4.759**
2	0.0250	0.901	1.305	1.000	**1.111**	**6.861**	**11.620**

Pairwise Error Rate	**0.0228**	Pairwise Power	**0.8713**
Familywise Error Rate (SE)	**0.0419 (0.0004)**		

```
Note: patient accrual stopped at time  5.500
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
```

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	579	193	386
Patients*	**2755**	**918**	**1837**
Events**	**233**	**77**	**156**
	Stage 2		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	579	193	386
Patients*	**3185**	**1062**	**2123**
Events**	**896**	**298**	**598**

```
nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1.3046243 1.3046243)
hr1(1 1) accrue(579 579) arms(3 3) t(3 2) s(0.894 0.894) tstop(5.5)
```

n-stage trial design version 4.0.1, 2 Nov 2018

Sample size for a 3-arm 2-stage trial with time-to-event outcome based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al. (2019) Clinical Trials 16(2)

Note: I outcome and D outcome are identical
Median survival time: **18.6** time units

Operating characteristics

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.5000	0.951	1.305	1.000	**1.305**	**4.759**	**4.759**
2	0.0250	0.901	1.305	1.000	**1.111**	**6.363**	**11.122**

Pairwise Error Rate **0.0228** Pairwise Power **0.8713**
Familywise Error Rate (SE) **0.0422** (**0.0004**)

Note: patient accrual stopped at time 5.500
* All alphas are one-sided
** Length (duration of each stage) is expressed in periods and assumes survival times are exponentially distributed. Time is expressed in cumulative periods.

Sample size and number of events

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	579	193	386
Patients*	**2755**	**918**	**1837**
Events**	**233**	**77**	**156**
	Stage 2		
	Overall	Control	Exper.
Arms	2	1	1
Acc. rate	579	290	290
Patients*	**3184**	**1133**	**2051**
Events**	**597**	**298**	**299**

QUASAR 2_NI Scenario 2

```
nstage, nstage(2) alpha(0.5 0.05) omega(0.9 0.9) hr0(1.3046243 1.3046243)
hr1(1 1) accrue(410 410) arms(3 3) t(3 3) s(0.894 0.894) tstop(6.5)
```

n-stage trial design version 4.0.1, 2 Nov 2018

Sample size for a 3-arm 2-stage trial with time-to-event outcome based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al. (2019) Clinical Trials 16(2)

Note: I outcome and D outcome are identical
Median survival time: **18.6** time units

Operating characteristics

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.5000	0.902	1.305	1.000	**1.305**	**4.409**	**4.409**
2	0.0500	0.901	1.305	1.000	**1.124**	**7.463**	**11.872**

Pairwise Error Rate	**0.0420**	Pairwise Power	**0.8316**
Familywise Error Rate (SE)	**0.0750** (**0.0005**)		

Note: patient accrual stopped at time 6.500
* All alphas are one-sided
** Length (duration of each stage) is expressed in periods and assumes survival times are exponentially distributed. Time is expressed in cumulative periods.

Sample size and number of events

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	410	137	273
Patients*	**1808**	**603**	**1205**
Events**	**143**	**47**	**96**
	Stage 2		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	410	137	273
Patients*	**2665**	**888**	**1777**
Events**	**731**	**243**	**488**

```
nstage, nstage(2) alpha(0.5 0.05) omega(0.9 0.9) hr0(1.3046243 1.3046243)
hr1(1 1) accrue(410 410) arms(3 3) t(3 2) s(0.894 0.894) tstop(6.5)

n-stage trial design                        version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------

Sample size for a 3-arm 2-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 18.6 time units

Operating characteristics
-------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
-------------------------------------------------------------------------
1           0.5000    0.902    1.305    1.000    1.305    4.409    4.409
2           0.0500    0.901    1.305    1.000    1.124    6.409   10.818
-------------------------------------------------------------------------
Pairwise Error Rate          0.0420              Pairwise Power   0.8316
Familywise Error Rate (SE)   0.0752 (0.0005)
-------------------------------------------------------------------------
Note: patient accrual stopped at time  6.500
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ----------Stage 1----------
           Overall  Control   Exper.
Arms             3        1        2
Acc. rate      410      137      273
Patients*     1808      603     1205
Events**       143       47       96
           ----------Stage 2----------
           Overall  Control   Exper.
Arms             2        1        1
Acc. rate      410      205      205
Patients*     2665     1031     1634
Events**       487      243      244
```

QUASAR 2_NI Scenario 3

```
nstage, nstage(2) alpha(0.25 0.05) omega(0.9 0.85) hr0(1.3046243 1.3046243)
hr1(1 1) accrue(579 579) arms(3 3) t(3 3) s(0.894 0.894) tstop(6)
```

n-stage trial design version 4.0.1, 2 Nov 2018

Sample size for a 3-arm 2-stage trial with time-to-event outcome based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al. (2019) Clinical Trials 16(2)

Note: I outcome and D outcome are identical
Median survival time: **18.6** time units

Operating characteristics

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.2500	0.902	1.305	1.000	**1.191**	**5.694**	**5.694**
2	0.0500	0.851	1.305	1.000	**1.109**	**2.550**	**8.244**

Pairwise Error Rate	**0.0437**	Pairwise Power	**0.8145**
Familywise Error Rate (SE)	**0.0768** (**0.0005**)		

Note: patient accrual stopped at time 6.000
* All alphas are one-sided
** Length (duration of each stage) is expressed in periods and assumes survival times are exponentially distributed. Time is expressed in cumulative periods.

Sample size and number of events

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	579	193	386
Patients*	**3297**	**1099**	**2198**
Events**	**329**	**109**	**220**
	Stage 2		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	579	193	386
Patients*	**3474**	**1158**	**2316**
Events**	**614**	**204**	**410**

```
nstage, nstage(2) alpha(0.25 0.05) omega(0.9 0.85) hr0(1.3046243 1.3046243)
hr1(1 1) accrue(579 579) arms(3 3) t(3 2) s(0.894 0.894) tstop(6)
```

n-stage trial design version 4.0.1, 2 Nov 2018

Sample size for a 3-arm 2-stage trial with time-to-event outcome based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al. (2019) Clinical Trials 16(2)

Note: I outcome and D outcome are identical
Median survival time: **18.6** time units

Operating characteristics

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.2500	0.902	1.305	1.000	**1.191**	**5.694**	**5.694**
2	0.0500	0.851	1.305	1.000	**1.109**	**2.481**	**8.176**

Pairwise Error Rate **0.0437** Pairwise Power **0.8145**
Familywise Error Rate (SE) **0.0778** (**0.0005**)

Note: patient accrual stopped at time 6.000
* All alphas are one-sided
** Length (duration of each stage) is expressed in periods and assumes survival times are exponentially distributed. Time is expressed in cumulative periods.

Sample size and number of events

	Stage 1		
	Overall	Control	Exper.
Arms	3	1	2
Acc. rate	579	193	386
Patients*	**3297**	**1099**	**2198**
Events**	**329**	**109**	**220**
	Stage 2		
	Overall	Control	Exper.
Arms	2	1	1
Acc. rate	579	290	290
Patients*	**3474**	**1188**	**2286**
Events**	**409**	**204**	**205**

Appendix 8: Annotated Stata code used to implement the simulations

DEVA_NI Scenario 1

```
* What: Generating sample sizes for DEVA NI trial with different alpha and
power values

* Set working directory

cd "C:\Users\pmistry2\OneDrive - JNJ\PhD\3. Post PhD\Thesis\Chapter
6\Analysis\DEVA"

* Generate survival and event data, exponential distribution,
* Control arm = 5 year OS 81.8%, lambda = log(0.818)/-5
* Experimental 1 = 5 year OS 81.8 to test under H0 lambda = log(0.818)/-5 =
0.04017859
* Experimental 2 = 5 year OS 78.8% to test under H1 lambda = log(0.788)/-5
= 0.04765144

**********************************************************************
******
* Scenario 1
* Alpha = 0.5, 0.025
* Power = 0.95, 0.9

nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845)
arms(3 3) t(5 5) s(0.818 0.818) tstop(8)

nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845)
arms(3 2) t(5 5) s(0.818 0.818) tstop(8) x

/*
Alpha = 0.5, 0.025
Power = 0.95, 0.9
Sample size in first stage = 4995
Duration of the first stage including recruitment = 5.912

Sample size in second stage = 1765
Total trial length (no arms dropped) = 13.527
Total trial length 2 (1 arm dropped) = 12.586

Critical HR at stage 1 = 1.188
Critical HR at final stage = 1.070
*/

program drop DEVA_3A2S_NI

prog define DEVA_3A2S_NI, rclass

* Simulate First stage of the trial
clear

local ss_s1 4995 // Set sample size in first stage
local time_s1 5.912 // Set duration of first stage
local lambda_trt1 0.04017859 // Lambda for the control arm
local lambda_trt2 0.04017859 // Lambda for the first exp. arm (Same as
control)
```

```
local lambda_trt3 0.04765144 // Lambda forthe second exp. arm (Alternative
hypothesis)
local trial_length 13.527 // Total trial length if no arms are dropped
local c_hr_s1 1.188 // Critical HR at stage 1

local ss_s2 1765 // Remaining number of patients to be randomised
local trial_length2 12.586 // Total trial length if 1 arm dropped


set obs `ss_s1'

gen trt = runiformint(1,3) // Generate 3 treatment arms
gen rectime = runiform(0, `time_s1') // Generate recruitment (uniform rate)
from 0 to 5.912

* Generate survival time for these patients till end of study, 13.527 years
if arms continue to
very end and if one arm is dropped then trial length will reduce to 12.586

gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt == 3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* Generate NEW survival variable for which events and patients are censored
till time 5.912 - end
of first stage

* Censoring survival time
gen stime_s1 = stime
replace stime_s1 = `time_s1' - rectime if stime_s1+rectime > `time_s1'

* Censoring survival events at stage 1
gen event_s1 = event
replace event_s1 = 0 if rectime+stime > `time_s1'

* Storing number of patients on each arm at stage 1

quietly: tab trt, matcell(trt_s)
quietly: gen trt_s1_1 = trt_s[1,1] in 1
quietly: gen trt_s1_2 = trt_s[2,1] in 1
quietly: gen trt_s1_3 = trt_s[3,1] in 1

* Storing number of events on each arm at stage 1
quietly: tab event_s1 trt if event_s1==1, matcell(event_s1_)
quietly: svmat event_s1_
```

```
* Set survival data to perform analysis between arms 1 and 2
stset stime_s1 if trt==1 | trt==2, failure(event_s1 = 1)

* Perform cox regression (PH Assumption) pairwise comparison and store HR

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s1_12 (exp(_b[2.trt]))
gen hr_s1_12 = `hr_s1_12' in 1 // store the HR

drop _*
stset stime_s1 if trt==1 | trt==3, failure(event_s1 = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s1_13 (exp(_b[3.trt]))
gen hr_s1_13 = `hr_s1_13' in 1 // store the HR

drop _*

* Generating a flag to indicate the outcome of the pairwise comparisons
gen s1_flag = 1 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Both arms continue
replace s1_flag = 2 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Arm 2 continue, Arm 3
stop
replace s1_flag = 3 if hr_s1_12 > `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Arm 2 stop, Arm 3
continue
replace s1_flag = 4 if hr_s1_12 > `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Both arm stop

* Below shows the relevant calculations for each option
if s1_flag == 1 in 1 { // Both arms continue

* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2' // Remaining patients to be randomised
gen trt = runiformint(1,3) // Both arms continue hence equally distributed
between 3 arms
140 gen rectime = runiform(`time_s1', 8) // recruitment time between 6.832
and 8 years

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt ==1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt==3
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
```

```
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* Merge all patients together
tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR

drop _*
stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR

}


if s1_flag == 2 in 1 { // Arm 2 continue, Arm 3 stop

* Censoring survival time & events as arm dropped, trial length reduced and
update to
incorporate recruitment
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'

* Censoring all events on arm 3
replace event = 0 if rectime+stime > `time_s1' & trt==3


* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2'

gen trt = runiformint(1,2)
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'
```

```
* Cenosoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR

gen hr_s2_13 = .

}

if s1_flag == 3 in 1 { // Arm 2 stops, Arm 3 continues

* Censoring survival time as arm dropped and update to incorporate
recruitment &
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'
* Censoring all events on arm 2
replace event = 0 if rectime+stime > `time_s1' & trt==2


tempfile file1
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
replace trt = 3 if trt==2
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt==1
replace lambda = `lambda_trt3' if trt==3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Censoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'
```

```
stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR

gen hr_s2_12 = .

}


if s1_flag == 4 in 1 { // Both arm stops

gen hr_s2_12 = .
gen hr_s2_13 = .

}


* Storing number of patients and events on each arm at stage 1
* Patients
quietly: tab trt, matcell(trt_s2_)
quietly: gen trt_s2_1 = trt_s2_[1,1] in 1
quietly: gen trt_s2_2 = trt_s2_[2,1] in 1
quietly: gen trt_s2_3 = trt_s2_[3,1] in 1

* Events
quietly: tab event trt if event==1, matcell(event_s2_)
quietly: svmat event_s2_

end


* Simulate the data 100 times, show number of patients, treatment, hr and
p-values
set seed 130933
simulate trt_s1_1 = trt_s1_1 trt_s1_2 = trt_s1_2 trt_s1_3 = trt_s1_3 ///
event_s1_1 = event_s1_1 event event_s1_2 = event_s1_2 event_s1_3 =
event_s1_3 ///
hr_s1_12 = hr_s1_12 hr_s1_13 = hr_s1_13 ///
hr_s2_12 = hr_s2_12 hr_s2_13 = hr_s2_13 ///
s1_flag = s1_flag ///
trt_s2_1 = trt_s2_1 trt_s2_2 = trt_s2_2 trt_s2_3 = trt_s2_3 ///
event_s2_1 = event_s2_1 event event_s2_2 = event_s2_2 event_s2_3 =
event_s2_3, reps(10000) :
DEVA_3A2S_NI

save DEVA_NI_Scenario1.dta,replace

* * * * * * *
use DEVA_NI_Scenario1.dta,clear

* Look at the means
ci means trt_s1_1 trt_s1_2 trt_s1_3 event_s1_1 event_s1_2 event_s1_3
event_s2_1 event_s2_2
event_s2_3 hr_s1_12 hr_s1_13
```

```
* Assessing the simulations out of total simulations

* Type 1 error under H0 and Type II error under H1 at stage 1

count if hr_s1_12 <= 1.188

count if hr_s1_13 <= 1.188

* Type 1 error under H0 and Type II error under H1 at final stage

count if hr_s2_12 <= 1.070

count if hr_s2_13 <= 1.070


* Looking at both arms continues (s1_flag =1 )
ci  means  trt_s2_1  trt_s2_2  trt_s2_3  event_s2_1  event_s2_2  event_s2_3
hr_s2_12 hr_s2_13 p_s2_12
p_s2_13 if s1_flag==1

* Looking at only arm 2 continues (s1_flag = 2)
ci  means  trt_s2_1  trt_s2_2  event_s2_1  event_s2_2  hr_s2_12  p_s2_12  if
s1_flag==2

* Looking at only arm 3 continue (s1_flag = 3)
ci  means  trt_s2_1  trt_s2_3  event_s2_1  event_s2_3  hr_s2_13  p_s2_13  if
s1_flag==3
```

DEVA_NI Scenario 2

```
* What: Generating sample sizes for DEVA NI trial with different alpha and
power values

* Set working directory

cd "C:\Users\pmistry2\OneDrive - JNJ\PhD\3. Post PhD\Thesis\Chapter
6\Analysis\DEVA"

* Generate survival and event data, exponential distribution,
* Control arm = 5 year OS 81.8%, lambda = log(0.818)/-5
* Experimental 1 = 5 year OS 81.8 to test under H0 lambda = log(0.818)/-5 =
0.04017859
* Experimental 2 = 5 year OS 78.8% to test under H1 lambda = log(0.788)/-5
= 0.04765144

**********************************************************************
******
* Scenario 2
* Alpha = 0.5, 0.05
* Power = 0.9, 0.9

nstage, nstage(2) alpha(0.5 0.05) omega(0.9 0.9) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 3) t(5 5) s(0.818 0.818) tstop(8)

nstage, nstage(2) alpha(0.5 0.05) omega(0.9 0.9) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 2) t(5 5) s(0.818 0.818) tstop(8)

/*
Alpha = 0.5, 0.05
Power = 0.9, 0.9
Sample size in first stage = 3857
Duration of the first stage including recruitment = 4.564

Sample size in second stage = 2903
Total trial length (no arms dropped) = 11.499
Total trial length 2 (1 arm dropped) = 10.425

Critical HR at stage 1 = 1.188
Critical HR at final stage = 1.078
*/

prog define DEVA_3A2S_NI, rclass

* Simulate First stage of the trial
clear

local ss_s1 3857 // Set sample size in first stage
local time_s1 4.564 // Set duration of first stage
local lambda_trt1 0.04017859 // Lambda for the control arm
local lambda_trt2 0.04017859 // Lambda for the first exp. arm (Same as
control)
local lambda_trt3 0.04765144 // Lambda forthe second exp. arm (Alternative
hypothesis)
local trial_length 11.499 // Total trial length if no arms are dropped
local c_hr_s1 1.188 // Critical HR at stage 1

local ss_s2 2903 // Remaining number of patients to be randomised
```

```
local trial_length2 10.425 // Total trial length if 1 arm dropped


set obs `ss_s1'

gen trt = runiformint(1,3) // Generate 3 treatment arms
gen rectime = runiform(0, `time_s1') // Generate recruitment (uniform rate)
from 0 to 5.912

* Generate survival time for these patients till end of study, 13.527 years
if arms continue to
very end and if one arm is dropped then trial length will reduce to 12.586

gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1

replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt == 3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* Generate NEW survival variable for which events and patients are censored
till time 5.912 - end
of first stage

* Censoring survival time
gen stime_s1 = stime
replace stime_s1 = `time_s1' - rectime if stime_s1+rectime > `time_s1'

* Censoring survival events at stage 1
gen event_s1 = event
replace event_s1 = 0 if rectime+stime > `time_s1'

* Storing number of patients on each arm at stage 1

quietly: tab trt, matcell(trt_s)
quietly: gen trt_s1_1 = trt_s[1,1] in 1
quietly: gen trt_s1_2 = trt_s[2,1] in 1
quietly: gen trt_s1_3 = trt_s[3,1] in 1

* Storing number of events on each arm at stage 1
quietly: tab event_s1 trt if event_s1==1, matcell(event_s1_)
quietly: svmat event_s1_

* Set survival data to perform analysis between arms 1 and 2
stset stime_s1 if trt==1 | trt==2, failure(event_s1 = 1)

* Perform cox regression (PH Assumption) pairwise comparison and store HR,
p-value which obtained
```

```
from LR test however it is equivalent to the log-rank test

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s1_12 (exp(_b[2.trt]))
gen hr_s1_12 = `hr_s1_12' in 1 // store the HR
sts test trt, logrank
* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_12 > 1 in 1
replace p_s1_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime_s1 if trt==1 | trt==3, failure(event_s1 = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s1_13 (exp(_b[3.trt]))
gen hr_s1_13 = `hr_s1_13' in 1 // store the HR
sts test trt, logrank

* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)

gen p_s1_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_13 > 1 in 1
replace p_s1_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_13 < 1 in 1 //
Store p-value according to the value of HR

drop _*

* Generating a flag to indicate the outcome of the pairwise comparisons
gen s1_flag = 1 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Both arms continue
replace s1_flag = 2 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Arm 2 continue, Arm 3
stop
replace s1_flag = 3 if hr_s1_12 > `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Arm 2 stop, Arm 3
continue
replace s1_flag = 4 if hr_s1_12 > `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Both arm stop

* Below shows the relevant calculations for each option

if s1_flag == 1 in 1 { // Both arms continue

* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2' // Remaining patients to be randomised
gen trt = runiformint(1,3) // Both arms continue hence equally distributed
between 3 arms
gen rectime = runiform(`time_s1', 8) // recruitment time between 6.832 and
8 years

* Generate survival data for these patients
gen u = runiform(0,1)
```

```
gen lambda = `lambda_trt1' if trt ==1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt==3
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Censoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* Merge all patients together
tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value
according to the value of HR

drop _*
stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR


}


if s1_flag == 2 in 1 { // Arm 2 continue, Arm 3 stop

* Censoring survival time & events as arm dropped, trial length reduced and
update to
incorporate recruitment
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'

* Censoring all events on arm 3
replace event = 0 if rectime+stime > `time_s1' & trt==3
```

```
* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Censoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value according to the value of HR

gen p_s2_13 = .
gen hr_s2_13 = .

}

if s1_flag == 3 in 1 { // Arm 2 stops, Arm 3 continues

* Censoring survival time as arm dropped and update to incorporate
recruitment &
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'
* Censoring all events on arm 2
replace event = 0 if rectime+stime > `time_s1' & trt==2


tempfile file1
save `file1'
```

```
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
replace trt = 3 if trt==2
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt==1
replace lambda = `lambda_trt3' if trt==3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Censoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'


stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR


gen p_s2_12 = .
gen hr_s2_12 = .

}


if s1_flag == 4 in 1 { // Both arm stops

gen hr_s2_12 = .
gen p_s2_12 = .
gen hr_s2_13 = .
gen p_s2_13 = .

}

* Storing number of patients and events on each arm at stage 1
* Patients
quietly: tab trt, matcell(trt_s2_)
quietly: gen trt_s2_1 = trt_s2_[1,1] in 1
```

```
quietly: gen trt_s2_2 = trt_s2_[2,1] in 1
quietly: gen trt_s2_3 = trt_s2_[3,1] in 1

* Events
quietly: tab event trt if event==1, matcell(event_s2_)
quietly: svmat event_s2_

end


* Simulate the data 100 times, show number of patients, treatment, hr and
p-values
set seed 130933
simulate trt_s1_1 = trt_s1_1 trt_s1_2 = trt_s1_2 trt_s1_3 = trt_s1_3 ///
event_s1_1 = event_s1_1 event event_s1_2 = event_s1_2 event_s1_3 =
event_s1_3 ///
hr_s1_12 = hr_s1_12 hr_s1_13 = hr_s1_13 ///
p_s1_12 = p_s1_12 p_s1_13 = p_s1_13 ///
hr_s2_12 = hr_s2_12 hr_s2_13 = hr_s2_13 ///
p_s2_12 = p_s2_12 p_s2_13 = p_s2_13 ///
s1_flag = s1_flag ///
trt_s2_1 = trt_s2_1 trt_s2_2 = trt_s2_2 trt_s2_3 = trt_s2_3 ///
event_s2_1 = event_s2_1 event event_s2_2 = event_s2_2 event_s2_3 =
event_s2_3, reps(10000) :DEVA_3A2S_NI

save DEVA_NI_Scenario2.dta,replace

use DEVA_NI_Scenario2.dta,clear

* Look at the means for different s1_flags

* First stage
ci means trt_s1_1 trt_s1_2 trt_s1_3 event_s1_1 event_s1_2 event_s1_3
event_s2_1 event_s2_2
event_s2_3 hr_s1_12 hr_s1_13 p_s1_12 p_s1_13

* Looking at both arms continues (s1_flag =1 )
ci means trt_s2_1 trt_s2_2 trt_s2_3 event_s2_1 event_s2_2 event_s2_3
hr_s2_12 hr_s2_13 p_s2_12
p_s2_13 if s1_flag==1

* Looking at only arm 2 continues (s1_flag = 2)
ci means trt_s2_1 trt_s2_2 event_s2_1 event_s2_2 hr_s2_12 p_s2_12 if
s1_flag==2

* Looking at only arm 3 continue (s1_flag = 3)
ci means trt_s2_1 trt_s2_3 event_s2_1 event_s2_3 hr_s2_13 p_s2_13 if
s1_flag==3

* Assessing the simulations out of total simulations
* Type 1 error under H0 and Type II error under H0 at stage 1

count if hr_s1_12 <= 1.188
count if hr_s1_13 <= 1.188
* Type 1 error under H0 and Type II error under H0 at final stage
count if hr_s2_12 <= 1.078
count if hr_s2_13 <= 1.078
```

DEVA NI Scenario 3

```
* What: Generating sample sizes for DEVA NI trial with different alpha and
power values

* Set working directory

cd "C:\Users\pmistry2\OneDrive - JNJ\PhD\3. Post PhD\Thesis\Chapter
6\Analysis\DEVA"

* Generate survival and event data, exponential distribution,
* Control arm = 5 year OS 81.8%, lambda = log(0.818)/-5
* Experimental 1 = 5 year OS 81.8 to test under H0 lambda = log(0.818)/-5
=.04017859
* Experimental 2 = 5 year OS 78.8% to test under H1 lambda = log(0.788)/-5
= 0.04765144

**********************************************************************
******
* Scenario 3
* Alpha = 0.25, 0.05
* Power = 0.9, 0.85

nstage, nstage(2) alpha(0.25 0.05) omega(0.9 0.85) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 3) t(5 5) s(0.818 0.818) tstop(8)

nstage, nstage(2) alpha(0.25 0.05) omega(0.9 0.85) hr0(1.1878 1.1878) hr1(1
1) accrue(845 845) arms(3 2) t(5 5) s(0.818 0.818) tstop(8)

/*
Alpha = 0.25, 0.05
Power = 0.9, 0.85
Sample size in first stage = 5988
Duration of the first stage including recruitment = 7.086

Sample size in second stage = 772
Total trial length (no arms dropped) = 10.154
Total trial length 2 (1 arm dropped) = 9.985

Critical HR at stage 1 = 1.119
Critical HR at final stage = 1.069
*/

program drop DEVA_3A2S_NI

prog define DEVA_3A2S_NI, rclass

* Simulate First stage of the trial
clear

local ss_s1 5988 // Set sample size in first stage
local time_s1 7.086 // Set duration of first stage
local lambda_trt1 0.04017859 // Lambda for the control arm
local lambda_trt2 0.04017859 // Lambda for the first exp. arm (Same as
control)
local lambda_trt3 0.04765144 // Lambda forthe second exp. arm (Alternative
hypothesis)
local trial_length 10.154 // Total trial length if no arms are dropped
local c_hr_s1 1.119 // Critical HR at stage 1
```

```
local ss_s2 772 // Remaining number of patients to be randomised
local trial_length2 9.985 // Total trial length if 1 arm dropped


set obs `ss_s1'

gen trt = runiformint(1,3) // Generate 3 treatment arms
gen rectime = runiform(0, `time_s1') // Generate recruitment (uniform rate)
from 0 to 5.912

* Generate survival time for these patients till end of study, 13.527 years
if arms continue to
very end and if one arm is dropped then trial length will reduce to 12.586

gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt == 3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* Generate NEW survival variable for which events and patients are censored
till time 5.912 - end
of first stage

* Censoring survival time
gen stime_s1 = stime
replace stime_s1 = `time_s1' - rectime if stime_s1+rectime > `time_s1'

* Censoring survival events at stage 1
gen event_s1 = event
replace event_s1 = 0 if rectime+stime > `time_s1'

* Storing number of patients on each arm at stage 1

quietly: tab trt, matcell(trt_s)
quietly: gen trt_s1_1 = trt_s[1,1] in 1
quietly: gen trt_s1_2 = trt_s[2,1] in 1
quietly: gen trt_s1_3 = trt_s[3,1] in 1

* Storing number of events on each arm at stage 1
quietly: tab event_s1 trt if event_s1==1, matcell(event_s1_)
quietly: svmat event_s1_

* Set survival data to perform analysis between arms 1 and 2
stset stime_s1 if trt==1 | trt==2, failure(event_s1 = 1)
```

```
* Perform cox regression (PH Assumption) pairwise comparison and store HR,
p-value which obtained from LR test however it is equivalent to the log-
rank test

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s1_12 (exp(_b[2.trt]))
gen hr_s1_12 = `hr_s1_12' in 1 // store the HR
sts test trt, logrank
* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_12 > 1 in 1
replace p_s1_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime_s1 if trt==1 | trt==3, failure(event_s1 = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s1_13 (exp(_b[3.trt]))
gen hr_s1_13 = `hr_s1_13' in 1 // store the HR
sts test trt, logrank

* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_13 > 1 in 1
replace p_s1_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_13 < 1 in 1 //
Store p-value according to the value of HR

drop _*

* Generating a flag to indicate the outcome of the pairwise comparisons
gen s1_flag = 1 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Both arms continue
replace s1_flag = 2 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Arm 2 continue, Arm 3 stop
replace s1_flag = 3 if hr_s1_12 > `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Arm 2 stop, Arm 3 continue
replace s1_flag = 4 if hr_s1_12 > `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Both arm stop

* Below shows the relevant calculations for each option

if s1_flag == 1 in 1 { // Both arms continue

* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2' // Remaining patients to be randomised
gen trt = runiformint(1,3) // Both arms continue hence equally distributed
between 3 arms
gen rectime = runiform(`time_s1', 8) // recruitment time between 6.832 and
8 years

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt ==1
```

```
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt==3
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* Merge all patients together
tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR
if s1_flag == 2 in 1 { // Arm 2 continue, Arm 3 stop

* Censoring survival time & events as arm dropped, trial length reduced and
update to incorporate recruitment
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'

* Censoring all events on arm 3
replace event = 0 if rectime+stime > `time_s1' & trt==3

* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
gen rectime = runiform(`time_s1', 8)
```

```
* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Cenosoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value according to the value of HR

gen p_s2_13 = .
gen hr_s2_13 = .

}

if s1_flag == 3 in 1 { // Arm 2 stops, Arm 3 continues

* Censoring survival time as arm dropped and update to incorporate
recruitment &
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'
* Censoring all events on arm 2
replace event = 0 if rectime+stime > `time_s1' & trt==2

tempfile file1
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
replace trt = 3 if trt==2
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt==1
replace lambda = `lambda_trt3' if trt==3
```

```
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Censoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'

stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR


gen p_s2_12 = .
gen hr_s2_12 = .

}


if s1_flag == 4 in 1 { // Both arm stops

gen hr_s2_12 = .
gen p_s2_12 = .
gen hr_s2_13 = .
gen p_s2_13 = .

}


* Storing number of patients and events on each arm at stage 1
* Patients
quietly: tab trt, matcell(trt_s2_)
quietly: gen trt_s2_1 = trt_s2_[1,1] in 1
quietly: gen trt_s2_2 = trt_s2_[2,1] in 1
quietly: gen trt_s2_3 = trt_s2_[3,1] in 1

* Events
quietly: tab event trt if event==1, matcell(event_s2_)

quietly: svmat event_s2_

end
```

```
* Simulate the data 100 times, show number of patients, treatment, hr and
p-values
set seed 130933
simulate trt_s1_1 = trt_s1_1 trt_s1_2 = trt_s1_2 trt_s1_3 = trt_s1_3 ///
event_s1_1 = event_s1_1 event event_s1_2 = event_s1_2 event_s1_3 =
event_s1_3 ///
hr_s1_12 = hr_s1_12 hr_s1_13 = hr_s1_13 ///
p_s1_12 = p_s1_12 p_s1_13 = p_s1_13 ///
hr_s2_12 = hr_s2_12 hr_s2_13 = hr_s2_13 ///
p_s2_12 = p_s2_12 p_s2_13 = p_s2_13 ///
s1_flag = s1_flag ///
trt_s2_1 = trt_s2_1 trt_s2_2 = trt_s2_2 trt_s2_3 = trt_s2_3 ///
event_s2_1 = event_s2_1 event event_s2_2 = event_s2_2 event_s2_3 =
event_s2_3, reps(10000) :
DEVA_3A2S_NI

save DEVA_NI_Scenario3.dta,replace

use DEVA_NI_Scenario3.dta,clear


* Look at the means for different s1_flags

* First stage
ci means trt_s1_1 trt_s1_2 trt_s1_3 event_s1_1 event_s1_2 event_s1_3
event_s2_1 event_s2_2
event_s2_3 hr_s1_12 hr_s1_13 p_s1_12 p_s1_13

* Looking at both arms continues (s1_flag =1 )
ci means trt_s2_1 trt_s2_2 trt_s2_3 event_s2_1 event_s2_2 event_s2_3
hr_s2_12 hr_s2_13 p_s2_12
p_s2_13 if s1_flag==1

* Looking at only arm 2 continues (s1_flag = 2)
ci means trt_s2_1 trt_s2_2 event_s2_1 event_s2_2 hr_s2_12 p_s2_12 if
s1_flag==2

* Looking at only arm 3 continue (s1_flag = 3)
ci means trt_s2_1 trt_s2_3 event_s2_1 event_s2_3 hr_s2_13 p_s2_13 if
s1_flag==3


* Assessing the simulations out of total simulations

* Type 1 error under H0 and Type II error under H0 at stage 1

count if hr_s1_12 <= 1.119
count if hr_s1_13 <= 1.119

* Type 1 error under H0 and Type II error under H0 at final stage

count if hr_s2_12 <= 1.069
count if hr_s2_13 <= 1.069
```

QUASAR2_NI Scenario 1

```
* What: Generating sample sizes for QUASAR NI trial with different alpha
and power values and simulating to check Type I & Power
* Set working directory

cd "C:\Users\pmistry2\OneDrive - JNJ\PhD\3. Post PhD\Thesis\Chapter
6\Analysis\QUASAR"

* Generate survival and event data, exponential distribution,
* Control arm = 3 year OS 89.4%, lambda = log(0.894)/-3 = 0.03734983
* Experimental 1 = 3 year OS 89.4 to test under H0 lambda = log(0.894)/-3 =
0.03734983
* Experimental 2 = 3 year OS 86.4% to test under H1 lambda = log(0.864)/-3
= 0.0487275


**********************************************************************
******
* Scenario 1
* Alpha = 0.5, 0.025
* Power = 0.95, 0.9

nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1.3046243 1.3046243)
///
hr1(1 1) accrue(579 579) arms(3 3) t(3 3) s(0.894 0.894) tstop(5.5)

* DROP ONE AT FIRST STAGE
nstage, nstage(2) alpha(0.5 0.025) omega(0.95 0.9) hr0(1.3046243 1.3046243)
///
hr1(1 1) accrue(579 579) arms(3 2) t(3 3) s(0.894 0.894) tstop(5.5)

/*
Alpha = 0.5, 0.025
Power = 0.95, 0.9
Sample size in first stage = 2755
Duration of the first stage including recruitment = 4.759

Sample size in second stage = 430
Total trial length (no arms dropped) = 11.620
Total trial length 2 (1 arm dropped) = 11.122

Critical HR at stage 1 = 1.305
Critical HR at final stage = 1.111
*/

*program drop QUASAR2_3A2S_NI

prog define QUASAR2_3A2S_NI, rclass

* Simulate First stage of the trial
clear

local ss_s1 2755 // Set sample size in first stage
local time_s1 4.759 // Set duration of first stage
local lambda_trt1 0.03734983 // Lambda for the control arm
local lambda_trt2 0.03734983 // Lambda for the first exp. arm (Same as
control)
```

```
local lambda_trt3 0.0487275 // Lambda forthe second exp. arm (Alternative
hypothesis)
local trial_length 11.620 // Total trial length if no arms are dropped
local c_hr_s1 1.305 // Critical HR at stage 1

local ss_s2 430 // Remaining number of patients to be randomised
local trial_length2 11.122 // Total trial length if 1 arm dropped

set obs `ss_s1'

gen trt = runiformint(1,3) // Generate 3 treatment arms
gen rectime = runiform(0, `time_s1') // Generate recruitment (uniform rate)
from 0 to 5.912

* Generate survival time for these patients till end of study, 13.527 years
if arms continue to very end and if one arm is dropped then trial length
will reduce to 12.586

gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt == 3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* Generate NEW survival variable for which events and patients are censored
till time 5.912 – end of first stage

* Censoring survival time
gen stime_s1 = stime
replace stime_s1 = `time_s1' - rectime if stime_s1+rectime > `time_s1'

* Censoring survival events at stage 1
gen event_s1 = event
replace event_s1 = 0 if rectime+stime > `time_s1'

* Storing number of patients on each arm at stage 1

quietly: tab trt, matcell(trt_s)
quietly: gen trt_s1_1 = trt_s[1,1] in 1
quietly: gen trt_s1_2 = trt_s[2,1] in 1
quietly: gen trt_s1_3 = trt_s[3,1] in 1

* Storing number of events on each arm at stage 1
quietly: tab event_s1 trt if event_s1==1, matcell(event_s1_)
quietly: svmat event_s1_

* Set survival data to perform analysis between arms 1 and 2
stset stime_s1 if trt==1 | trt==2, failure(event_s1 = 1)
```

```
* Perform cox regression (PH Assumption) pairwise comparison and store HR,
p-value which obtained from LR test however it is equivalent to the log-
rank test

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s1_12 (exp(_b[2.trt]))
gen hr_s1_12 = `hr_s1_12' in 1 // store the HR
sts test trt, logrank

* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_12 > 1 in 1
replace p_s1_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime_s1 if trt==1 | trt==3, failure(event_s1 = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s1_13 (exp(_b[3.trt]))
gen hr_s1_13 = `hr_s1_13' in 1 // store the HR
sts test trt, logrank
* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_13 > 1 in 1
replace p_s1_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_13 < 1 in 1 //
Store p-value according to the value of HR

drop _*

* Generating a flag to indicate the outcome of the pairwise comparisons
gen s1_flag = 1 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Both arms continue
replace s1_flag = 2 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Arm 2 continue, Arm 3 stop
replace s1_flag = 3 if hr_s1_12 > `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Arm 2 stop, Arm 3 continue
replace s1_flag = 4 if hr_s1_12 > `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Both arm stop

* Below shows the relevant calculations for each option

if s1_flag == 1 in 1 { // Both arms continue

* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2' // Remaining patients to be randomised
gen trt = runiformint(1,3) // Both arms continue hence equally distributed
between 3 arms
gen rectime = runiform(`time_s1', 8) // recruitment time between 6.832 and
8 years

* Generate survival data for these patients
gen u = runiform(0,1)
```

```
gen lambda = `lambda_trt1' if trt ==1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt==3
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* Merge all patients together
tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR

}


if s1_flag == 2 in 1 { // Arm 2 continue, Arm 3 stop

* Censoring survival time & events as arm dropped, trial length reduced and
update to incorporate recruitment
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'

* Censoring all events on arm 3
replace event = 0 if rectime+stime > `time_s1' & trt==3

* Creating a new file for the remaining patients
tempfile file1
```

```
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Cenosoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value according to the value of HR

gen p_s2_13 = .
gen hr_s2_13 = .

}

if s1_flag == 3 in 1 { // Arm 2 stops, Arm 3 continues

*  Censoring  survival  time  as  arm  dropped  and  update  to  incorporate
recruitment &
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'
* Censoring all events on arm 2

replace event = 0 if rectime+stime > `time_s1' & trt==2


tempfile file1
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
```

```
replace trt = 3 if trt==2
gen rectime = runiform(`time_s1', 8)


* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt==1
replace lambda = `lambda_trt3' if trt==3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Censoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'


stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR

gen p_s2_12 = .
gen hr_s2_12 = .

}

if s1_flag == 4 in 1 { // Both arm stops

gen hr_s2_12 = .
gen p_s2_12 = .
gen hr_s2_13 = .
gen p_s2_13 = .

}


* Storing number of patients and events on each arm at stage 1
* Patients
quietly: tab trt, matcell(trt_s2_)
quietly: gen trt_s2_1 = trt_s2_[1,1] in 1
quietly: gen trt_s2_2 = trt_s2_[2,1] in 1
quietly: gen trt_s2_3 = trt_s2_[3,1] in 1
```

```
* Events
quietly: tab event trt if event==1, matcell(event_s2_)
quietly: svmat event_s2_

end


* Simulate the data 100 times, show number of patients, treatment, hr and
p-values
set seed 130933
simulate trt_s1_1 = trt_s1_1 trt_s1_2 = trt_s1_2 trt_s1_3 = trt_s1_3 ///
event_s1_1 = event_s1_1 event event_s1_2 = event_s1_2 event_s1_3 =
event_s1_3 ///
hr_s1_12 = hr_s1_12 hr_s1_13 = hr_s1_13 ///
p_s1_12 = p_s1_12 p_s1_13 = p_s1_13 ///
hr_s2_12 = hr_s2_12 hr_s2_13 = hr_s2_13 ///
p_s2_12 = p_s2_12 p_s2_13 = p_s2_13 ///
s1_flag = s1_flag ///
trt_s2_1 = trt_s2_1 trt_s2_2 = trt_s2_2 trt_s2_3 = trt_s2_3 ///
event_s2_1 = event_s2_1 event event_s2_2 = event_s2_2 event_s2_3 =
event_s2_3, reps(10000):

QUASAR2_3A2S_NI

save QUASAR2_NI_Scenario1.dta,replace

use QUASAR2_NI_Scenario1.dta,clear
* Look at the means for different s1_flags

* First stage
ci means trt_s1_1 trt_s1_2 trt_s1_3 event_s1_1 event_s1_2 event_s1_3
event_s2_1 event_s2_2
event_s2_3 hr_s1_12 hr_s1_13 p_s1_12 p_s1_13
* Looking at both arms continues (s1_flag =1 )
ci means trt_s2_1 trt_s2_2 trt_s2_3 event_s2_1 event_s2_2 event_s2_3
hr_s2_12 hr_s2_13 p_s2_12
p_s2_13 if s1_flag==1

* Looking at only arm 2 continues (s1_flag = 2)
ci means trt_s2_1 trt_s2_2 event_s2_1 event_s2_2 hr_s2_12 p_s2_12 if
s1_flag==2

* Looking at only arm 3 continue (s1_flag = 3)
ci means trt_s2_1 trt_s2_3 event_s2_1 event_s2_3 hr_s2_13 p_s2_13 if
s1_flag==3

* Assessing the simulations out of total simulations

* Type 1 error under H0 and Type II error under H0 at stage 1

count if hr_s1_12 <= 1.305
count if hr_s1_13 <= 1.305

* Type 1 error under H0 and Type II error under H0 at final stage

count if hr_s2_12 <= 1.111
count if hr_s2_13 <= 1.111
```

QUASAR2_NI Scenario 2

```
* What: Generating sample sizes for QUASAR NI trial with different alpha
and power values and simulating to check Type I & Power

* Set working directory

cd "C:\Users\pmistry2\OneDrive - JNJ\PhD\3. Post PhD\Thesis\Chapter
6\Analysis\QUASAR"

* Generate survival and event data, exponential distribution,
* Control arm = 3 year OS 89.4%, lambda = log(0.894)/-3 = 0.03734983
* Experimental 1 = 3 year OS 89.4 to test under H0 lambda = log(0.894)/-3 =
0.03734983
* Experimental 2 = 3 year OS 86.4% to test under H1 lambda = log(0.864)/-3
= 0.0487275

**********************************************************************
******
* Scenario 2
* Alpha = 0.5, 0.05
* Power = 0.9, 0.9

nstage, nstage(2) alpha(0.5 0.05) omega(0.9 0.9) hr0(1.3046243 1.3046243)
///
hr1(1 1) accrue(410 410) arms(3 3) t(3 3) s(0.894 0.894) tstop(6.5)

* DROP ONE AT FIRST STAGE
nstage, nstage(2) alpha(0.5 0.05) omega(0.9 0.9) hr0(1.3046243 1.3046243)
///
hr1(1 1) accrue(410 410) arms(3 2) t(3 3) s(0.894 0.894) tstop(6.5)

/*
Alpha = 0.5, 0.05
Power = 0.9, 0.9
Sample size in first stage = 1808
Duration of the first stage including recruitment = 4.409

Sample size in second stage = 857
Total trial length (no arms dropped) = 11.872
Total trial length 2 (1 arm dropped) = 10.018

Critical HR at stage 1 = 1.305
Critical HR at final stage = 1.124
*/

*program drop QUASAR2_3A2S_NI

prog define QUASAR2_3A2S_NI, rclass

* Simulate First stage of the trial
clear

local ss_s1 1808 // Set sample size in first stage
local time_s1 4.409 // Set duration of first stage
local lambda_trt1 0.03734983 // Lambda for the control arm
local lambda_trt2 0.03734983 // Lambda for the first exp. arm (Same as
control)
```

```
local lambda_trt3 0.0487275 // Lambda forthe second exp. arm (Alternative
hypothesis)
local trial_length 11.872 // Total trial length if no arms are dropped
local c_hr_s1 1.305 // Critical HR at stage 1

local ss_s2 857 // Remaining number of patients to be randomised
local trial_length2 10.018 // Total trial length if 1 arm dropped


set obs `ss_s1'

gen trt = runiformint(1,3) // Generate 3 treatment arms
gen rectime = runiform(0, `time_s1') // Generate recruitment (uniform rate)
from 0 to 5.912

* Generate survival time for these patients till end of study, 13.527 years
if arms continue to very end and if one arm is dropped then trial length
will reduce to 12.586

gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt == 3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* Generate NEW survival variable for which events and patients are censored
till time 5.912 – end of first stage

* Censoring survival time
gen stime_s1 = stime
replace stime_s1 = `time_s1' - rectime if stime_s1+rectime > `time_s1'

* Censoring survival events at stage 1
gen event_s1 = event
replace event_s1 = 0 if rectime+stime > `time_s1'

* Storing number of patients on each arm at stage 1

quietly: tab trt, matcell(trt_s)
quietly: gen trt_s1_1 = trt_s[1,1] in 1
quietly: gen trt_s1_2 = trt_s[2,1] in 1
quietly: gen trt_s1_3 = trt_s[3,1] in 1

* Storing number of events on each arm at stage 1
quietly: tab event_s1 trt if event_s1==1, matcell(event_s1_)
quietly: svmat event_s1_

* Set survival data to perform analysis between arms 1 and 2
```

```
stset stime_s1 if trt==1 | trt==2, failure(event_s1 = 1)

* Perform cox regression (PH Assumption) pairwise comparison and store HR,
p-value which obtained from LR test however it is equivalent to the log-
rank test

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s1_12 (exp(_b[2.trt]))
gen hr_s1_12 = `hr_s1_12' in 1 // store the HR
sts test trt, logrank

* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_12 > 1 in 1
replace p_s1_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime_s1 if trt==1 | trt==3, failure(event_s1 = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s1_13 (exp(_b[3.trt]))
gen hr_s1_13 = `hr_s1_13' in 1 // store the HR
sts test trt, logrank

* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_13 > 1 in 1
replace p_s1_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_13 < 1 in 1 //
Store p-value according to the value of HR

drop _*

* Generating a flag to indicate the outcome of the pairwise comparisons
gen s1_flag = 1 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Both arms continue
replace s1_flag = 2 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Arm 2 continue, Arm 3 stop
replace s1_flag = 3 if hr_s1_12 > `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Arm 2 stop, Arm 3 continue
replace s1_flag = 4 if hr_s1_12 > `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Both arm stop

* Below shows the relevant calculations for each option

if s1_flag == 1 in 1 { // Both arms continue

* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2' // Remaining patients to be randomised
gen trt = runiformint(1,3) // Both arms continue hence equally distributed
between 3 arms
gen rectime = runiform(`time_s1', 8) // recruitment time between 6.832 and
8 years
```

```
* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt ==1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt==3
gen stime = -log(u)/lambda
* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* Merge all patients together
tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR

}


if s1_flag == 2 in 1 { // Arm 2 continue, Arm 3 stop

* Censoring survival time & events as arm dropped, trial length reduced and
update to incorporate recruitment
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'

* Censoring all events on arm 3
replace event = 0 if rectime+stime > `time_s1' & trt==3

* Creating a new file for the remaining patients
```

```
tempfile file1
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
gen rectime = runiform(`time_s1', 8)
* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Cenosoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-valueaccording to the value of HR

gen p_s2_13 = .
gen hr_s2_13 = .

}

if s1_flag == 3 in 1 { // Arm 2 stops, Arm 3 continues

* Censoring survival time as arm dropped and update to incorporate
recruitment &
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'
* Censoring all events on arm 2
replace event = 0 if rectime+stime > `time_s1' & trt==2


tempfile file1
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
replace trt = 3 if trt==2
```

```
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt==1
replace lambda = `lambda_trt3' if trt==3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Censoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'


stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR


gen p_s2_12 = .
gen hr_s2_12 = .

}


if s1_flag == 4 in 1 { // Both arm stops

gen hr_s2_12 = .
gen p_s2_12 = .
gen hr_s2_13 = .
gen p_s2_13 = .

}


* Storing number of patients and events on each arm at stage 1
* Patients
quietly: tab trt, matcell(trt_s2_)
quietly: gen trt_s2_1 = trt_s2_[1,1] in 1
quietly: gen trt_s2_2 = trt_s2_[2,1] in 1
quietly: gen trt_s2_3 = trt_s2_[3,1] in 1
```

```
* Events
quietly: tab event trt if event==1, matcell(event_s2_)
quietly: svmat event_s2_

end


* Simulate the data 100 times, show number of patients, treatment, hr and
p-values
set seed 130933
simulate trt_s1_1 = trt_s1_1 trt_s1_2 = trt_s1_2 trt_s1_3 = trt_s1_3 ///
event_s1_1 = event_s1_1 event event_s1_2 = event_s1_2 event_s1_3 =
event_s1_3 ///
hr_s1_12 = hr_s1_12 hr_s1_13 = hr_s1_13 ///
p_s1_12 = p_s1_12 p_s1_13 = p_s1_13 ///
hr_s2_12 = hr_s2_12 hr_s2_13 = hr_s2_13 ///
p_s2_12 = p_s2_12 p_s2_13 = p_s2_13 ///
s1_flag = s1_flag ///
trt_s2_1 = trt_s2_1 trt_s2_2 = trt_s2_2 trt_s2_3 = trt_s2_3 ///
event_s2_1 = event_s2_1 event event_s2_2 = event_s2_2 event_s2_3 =
event_s2_3, reps(10000):

QUASAR2_3A2S_NI

save QUASAR2_NI_Scenario2.dta,replace

use QUASAR2_NI_Scenario2.dta,clear

* Look at the means for different s1_flags

* First stage
ci means trt_s1_1 trt_s1_2 trt_s1_3 event_s1_1 event_s1_2 event_s1_3
event_s2_1 event_s2_2
event_s2_3 hr_s1_12 hr_s1_13 p_s1_12 p_s1_13
* Looking at both arms continues (s1_flag =1 )
ci means trt_s2_1 trt_s2_2 trt_s2_3 event_s2_1 event_s2_2 event_s2_3
hr_s2_12 hr_s2_13 p_s2_12
p_s2_13 if s1_flag==1

* Looking at only arm 2 continues (s1_flag = 2)
ci means trt_s2_1 trt_s2_2 event_s2_1 event_s2_2 hr_s2_12 p_s2_12 if
s1_flag==2

* Looking at only arm 3 continue (s1_flag = 3)
ci means trt_s2_1 trt_s2_3 event_s2_1 event_s2_3 hr_s2_13 p_s2_13 if
s1_flag==3


* Assessing the simulations out of total simulations

* Type 1 error under H0 and Type II error under H0 at stage 1

count if hr_s1_12 <= 1.305
count if hr_s1_13 <= 1.305

* Type 1 error under H0 and Type II error under H0 at final stage

count if hr_s2_12 <= 1.124
count if hr_s2_13 <= 1.124
```

QUASAR2_NI Scenario 3

```
* What: Generating sample sizes for QUASAR NI trial with different alpha
and power values and simulating to check Type I & Power

* Set working directory

cd "C:\Users\pmistry2\OneDrive - JNJ\PhD\3. Post PhD\Thesis\Chapter
6\Analysis\QUASAR"

* Generate survival and event data, exponential distribution,
* Control arm = 3 year OS 89.4%, lambda = log(0.894)/-3 = 0.03734983
* Experimental 1 = 3 year OS 89.4 to test under H0 lambda = log(0.894)/-3 =
0.03734983
* Experimental 2 = 3 year OS 86.4% to test under H1 lambda = log(0.864)/-3
= 0.0487275

**********************************************************************
******
* Scenario 3
* Alpha = 0.25, 0.05
* Power = 0.9, 0.85

nstage, nstage(2) alpha(0.25 0.05) omega(0.9 0.85) hr0(1.3046243 1.3046243)
///
hr1(1 1) accrue(579 579) arms(3 3) t(3 3) s(0.894 0.894) tstop(6)

* DROP ONE AT FIRST STAGE
nstage, nstage(2) alpha(0.25 0.05) omega(0.9 0.85) hr0(1.3046243 1.3046243)
///
hr1(1 1) accrue(579 579) arms(3 2) t(3 3) s(0.894 0.894) tstop(6)

/*
Alpha = 0.25, 0.05
Power = 0.9, 0.85
Sample size in first stage = 3297
Duration of the first stage including recruitment = 5.694

Sample size in second stage = 177
Total trial length (no arms dropped) = 8.244
Total trial length 2 (1 arm dropped) = 8.176

Critical HR at stage 1 = 1.191
Critical HR at final stage = 1.109
*/

*program drop QUASAR2_3A2S_NI

prog define QUASAR2_3A2S_NI, rclass

* Simulate First stage of the trial
clear

local ss_s1 3297 // Set sample size in first stage
local time_s1 5.694 // Set duration of first stage
local lambda_trt1 0.03734983 // Lambda for the control arm
local lambda_trt2 0.03734983 // Lambda for the first exp. arm (Same as
control)
```

```
local lambda_trt3 0.0487275 // Lambda forthe second exp. arm (Alternative
hypothesis)
local trial_length 8.244 // Total trial length if no arms are dropped
local c_hr_s1 1.191 // Critical HR at stage 1

local ss_s2 177 // Remaining number of patients to be randomised
local trial_length2 8.176 // Total trial length if 1 arm dropped


set obs `ss_s1'

gen trt = runiformint(1,3) // Generate 3 treatment arms
gen rectime = runiform(0, `time_s1') // Generate recruitment (uniform rate)
from 0 to 5.912

* Generate survival time for these patients till end of study, 13.527 years
if arms continue to very end and if one arm is dropped then trial length
will reduce to 12.586

gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt == 3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* Generate NEW survival variable for which events and patients are censored
till time 5.912 – end of first stage

* Censoring survival time
gen stime_s1 = stime
replace stime_s1 = `time_s1' - rectime if stime_s1+rectime > `time_s1'

* Censoring survival events at stage 1
gen event_s1 = event
replace event_s1 = 0 if rectime+stime > `time_s1'

* Storing number of patients on each arm at stage 1

quietly: tab trt, matcell(trt_s)
quietly: gen trt_s1_1 = trt_s[1,1] in 1
quietly: gen trt_s1_2 = trt_s[2,1] in 1
quietly: gen trt_s1_3 = trt_s[3,1] in 1

* Storing number of events on each arm at stage 1
quietly: tab event_s1 trt if event_s1==1, matcell(event_s1_)
quietly: svmat event_s1_

* Set survival data to perform analysis between arms 1 and 2
```

```
stset stime_s1 if trt==1 | trt==2, failure(event_s1 = 1)

* Perform cox regression (PH Assumption) pairwise comparison and store HR,
p-value which obtained from LR test however it is equivalent to the log-
rank test

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s1_12 (exp(_b[2.trt]))
gen hr_s1_12 = `hr_s1_12' in 1 // store the HR
sts test trt, logrank
* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_12 > 1 in 1
replace p_s1_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime_s1 if trt==1 | trt==3, failure(event_s1 = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s1_13 (exp(_b[3.trt]))
gen hr_s1_13 = `hr_s1_13' in 1 // store the HR
sts test trt, logrank
* Storing p-value according to HR - these differ based on hypothesis (sup
or ni)
gen p_s1_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s1_13 > 1 in 1
replace p_s1_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s1_13 < 1 in 1 //
Store p-value according to the value of HR

drop _*

* Generating a flag to indicate the outcome of the pairwise comparisons
gen s1_flag = 1 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Both arms continue
replace s1_flag = 2 if hr_s1_12 <= `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Arm 2 continue, Arm 3 stop
replace s1_flag = 3 if hr_s1_12 > `c_hr_s1' & hr_s1_13 <= `c_hr_s1' in 1 //
Arm 2 stop, Arm 3 continue
replace s1_flag = 4 if hr_s1_12 > `c_hr_s1' & hr_s1_13 > `c_hr_s1' in 1 //
Both arm stop

* Below shows the relevant calculations for each option

if s1_flag == 1 in 1 { // Both arms continue

* Creating a new file for the remaining patients
tempfile file1
save `file1'
clear
set obs `ss_s2' // Remaining patients to be randomised
gen trt = runiformint(1,3) // Both arms continue hence equally distributed
between 3 arms
gen rectime = runiform(`time_s1', 8) // recruitment time between 6.832 and
8 years

* Generate survival data for these patients
gen u = runiform(0,1)
```

```
gen lambda = `lambda_trt1' if trt ==1
replace lambda = `lambda_trt2' if trt == 2
replace lambda = `lambda_trt3' if trt==3
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at max FU time
replace event = 0 if stime + rectime > `trial_length'

* Cenosoring the survival time
replace stime = `trial_length'-rectime if stime+rectime > `trial_length'

* Merge all patients together
tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value according to the value of HR

drop _*
stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR


}

if s1_flag == 2 in 1 { // Arm 2 continue, Arm 3 stop

* Censoring survival time & events as arm dropped, trial length reduced and
update to incorporate recruitment
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'

* Censoring all events on arm 3
replace event = 0 if rectime+stime > `time_s1' & trt==3


* Creating a new file for the remaining patients
```

```
tempfile file1
save `file1'
clear
set obs `ss_s2'
gen trt = runiformint(1,2)
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt == 1
replace lambda = `lambda_trt2' if trt == 2
gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Cenosoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'

* Set survival data to perform analysis
stset stime if trt==1 | trt==2, failure(event = 1)

* BETWEEN ARMS 1 AND ARM 2
stcox i.trt // perform cox regression for arms 1 and 2
local hr_s2_12 (exp(_b[2.trt]))
gen hr_s2_12 = `hr_s2_12' in 1 // store the HR
sts test trt, logrank
gen p_s2_12 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_12 > 1 in 1
replace p_s2_12 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_12 < 1 in 1 //
Store p-value according to the value of HR

gen p_s2_13 = .
gen hr_s2_13 = .

}

if s1_flag == 3 in 1 { // Arm 2 stops, Arm 3 continues

*  Censoring  survival  time  as  arm  dropped  and  update  to  incorporate
recruitment &
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'
replace event = 0 if stime + rectime > `trial_length2'

* Censoring all events on arm 2
replace event = 0 if rectime+stime > `time_s1' & trt==2


tempfile file1
save `file1'
clear
set obs `ss_s2'
```

```
gen trt = runiformint(1,2)
replace trt = 3 if trt==2
gen rectime = runiform(`time_s1', 8)

* Generate survival data for these patients
gen u = runiform(0,1)
gen lambda = `lambda_trt1' if trt==1
replace lambda = `lambda_trt3' if trt==3

gen stime = -log(u)/lambda

* Generate variable for events
gen event = 1

* Censoring number of events at 12.914 years as trial length has reduced
replace event = 0 if stime + rectime > `trial_length2'

* Censoring the survival time
replace stime = `trial_length2'-rectime if stime+rectime > `trial_length2'

tempfile file2
save `file2'
use `file1',clear
append using `file2'


stset stime if trt==1 | trt==3, failure(event = 1)

* BETWEEN ARMS 1 & ARM 3
stcox i.trt // perform cox regression for arms 1 and 3
local hr_s2_13 (exp(_b[3.trt]))
gen hr_s2_13 = `hr_s2_13' in 1 // store the HR
sts test trt, logrank
gen p_s2_13 = (chi2tail(r(df), r(chi2)))/2 if hr_s2_13 > 1 in 1
replace p_s2_13 = 1-((chi2tail(r(df), r(chi2)))/2) if hr_s2_13 < 1 in 1 //
Store p-value according to the value of HR

gen p_s2_12 = .
gen hr_s2_12 = .

}


if s1_flag == 4 in 1 { // Both arm stops

gen hr_s2_12 = .
gen p_s2_12 = .
gen hr_s2_13 = .
gen p_s2_13 = .

}


* Storing number of patients and events on each arm at stage 1
* Patients
quietly: tab trt, matcell(trt_s2_)
quietly: gen trt_s2_1 = trt_s2_[1,1] in 1
quietly: gen trt_s2_2 = trt_s2_[2,1] in 1
quietly: gen trt_s2_3 = trt_s2_[3,1] in 1
```

```
* Events
quietly: tab event trt if event==1, matcell(event_s2_)
quietly: svmat event_s2_

end

* Simulate the data 100 times, show number of patients, treatment, hr and
p-values
set seed 130933
simulate trt_s1_1 = trt_s1_1 trt_s1_2 = trt_s1_2 trt_s1_3 = trt_s1_3 ///
event_s1_1 = event_s1_1 event event_s1_2 = event_s1_2 event_s1_3 =
event_s1_3 ///
hr_s1_12 = hr_s1_12 hr_s1_13 = hr_s1_13 ///
p_s1_12 = p_s1_12 p_s1_13 = p_s1_13 ///
hr_s2_12 = hr_s2_12 hr_s2_13 = hr_s2_13 ///
p_s2_12 = p_s2_12 p_s2_13 = p_s2_13 ///
s1_flag = s1_flag ///
trt_s2_1 = trt_s2_1 trt_s2_2 = trt_s2_2 trt_s2_3 = trt_s2_3 ///
event_s2_1 = event_s2_1 event event_s2_2 = event_s2_2 event_s2_3 =
event_s2_3, reps(10000):
QUASAR2_3A2S_NI

save QUASAR2_NI_Scenario3.dta,replace

use QUASAR2_NI_Scenario3.dta,clear

* Look at the means for different s1_flags

* First stage
ci means trt_s1_1 trt_s1_2 trt_s1_3 event_s1_1 event_s1_2 event_s1_3
event_s2_1 event_s2_2
event_s2_3 hr_s1_12 hr_s1_13 p_s1_12 p_s1_13

* Looking at both arms continues (s1_flag =1 )
ci means trt_s2_1 trt_s2_2 trt_s2_3 event_s2_1 event_s2_2 event_s2_3
hr_s2_12 hr_s2_13 p_s2_12
p_s2_13 if s1_flag==1

* Looking at only arm 2 continues (s1_flag = 2)
ci means trt_s2_1 trt_s2_2 event_s2_1 event_s2_2 hr_s2_12 p_s2_12 if
s1_flag==2

* Looking at only arm 3 continue (s1_flag = 3)
ci means trt_s2_1 trt_s2_3 event_s2_1 event_s2_3 hr_s2_13 p_s2_13 if
s1_flag==3


* Assessing the simulations out of total simulations

* Type 1 error under H0 and Type II error under H0 at stage 1

count if hr_s1_12 <= 1.191
count if hr_s1_13 <= 1.191

* Type 1 error under H0 and Type II error under H0 at final stage

count if hr_s2_12 <= 1.109
count if hr_s2_13 <= 1.109
```

Appendix 9: Simulation results for DEVA_SUP

Trial	Scenario	Stage	Nstage Type I error	Simulated Type I error	Nstage Power	Simulated Power	Nstage Control arm events	No. Control arm events
DEVA_SUP	1	1	0.50	0.498 (0.488 - 0.508)	0.950	96.1% (95.7% - 96.5%)	12	12
		Final*	0.024	0.028 (0.025 - 0.031)	0.880	89.0% (88.4% - 89.6%)	35	34
	2	1	0.05	0.051 (0.047 - 0.055)	0.900	90.6% (90.0% - 91.2%)	7	7
		Final*	0.044	0.050 (0.046 - 0.054)	0.844	84.6% (83.9% - 85.3%)	29	29
	3	1	0.25	0.256 (0.247 - 0.265)	0.900	90.6% (90.0% - 91.2%)	14	14
		Final*	0.044	0.054 (0.050 - 0.058)	0.823	82.3% (81.6% - 83.0%)	24	23

*Showing final pairwise error rate and pairwise power

Appendix 10: Results from the sample size calculation for DEVA_NI and QUASAR 2_NI MAMS trials.

Trial	Results	3A2S		4A3S			5A4S			
		Stage		Stage			Stage			
		1	2	1	2	3	1	2	3	4
DEVA_NI	PWER		0.023			0.021				0.021
	Pairwise Power		0.87			0.86				0.85
	Critical HR	1.188	1.070	1.188	1.13	1.07	1.188	1.13	1.107	1.07
	Time	5.9	7.6	5.6	8.0	12.3	4.6	6.6	8.0	9.5
	Sample Size	4995	6760	7043	10087	10120	10505	15005	18236	18240
	CA Patients	1665	2253	1761	2522	2530	2101	3001	3647	3648
	CA Events	183	710	183	364	710	183	364	528	710
QUASAR 2_NI	PWER		0.023			0.021				0.021
	Pairwise Power		0.87			0.86				0.85
	Critical HR	1.305	1.111	1.305	1.208	1.111	1.305	1.208	1.169	1.111
	Time	4.8	11.6	4.6	6.5	9.9	4.3	6.1	7.3	8.7
	Sample Size	2755	3185	3828	5459	5460	5105	7276	8811	9000
	CA Patients	918	1062	957	1365	1365	1021	1455	1762	1800
	CA Events	77	298	77	153	298	77	153	221	298

Appendix 11: MAMS designs with strong control of the FWER applied with different absolute NI margins

2% absolute NI margin

```
. nstage, nstage(3) alpha(0.5 0.25 0.02) omega(0.95 0.95 0.9) hr0(1.21 1.21) hr1(1 1
> ) accrue(1500 1500 1500) arms(4 4 4) t(3 3) s(0.9 0.9) tstop(7)


n-stage trial design                        version 4.0.1, 2 Nov 2018
─────────────────────────────────────────────────────────────────────────────
Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
─────────────────────────────────────────────────────────────────────────────

Note: I outcome and D outcome are identical
Median survival time: 19.7 time units

Operating characteristics
──────────────────────────────────────────────────────────────────────────────
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
──────────────────────────────────────────────────────────────────────────────
1           0.5000    0.950    1.210    1.000    1.210    4.893    4.893
2           0.2500    0.950    1.210    1.000    1.145    2.098    6.991
3           0.0200    0.900    1.210    1.000    1.076    4.154   11.144
──────────────────────────────────────────────────────────────────────────────
Pairwise Error Rate          0.0171              Pairwise Power   0.8562
Familywise Error Rate (SE)   0.0443 (0.0004)
──────────────────────────────────────────────────────────────────────────────
Note: patient accrual stopped at time  7.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ─────────Stage 1─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate     1500      375     1125
Patients*     7340     1835     5505
Events**       599      149      450
           ─────────Stage 2─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate     1500      375     1125
Patients*    10485     2621     7864
Events**      1188      297      891
           ─────────Stage 3─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate     1500      375     1125
Patients*    10500     2625     7875
Events**      2452      613     1839
           ──────────────────────────
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE

.
end of do-file
```

3% absolute NI margin

```
. nstage, nstage(3) alpha(0.5 0.25 0.02) omega(0.95 0.95 0.9) hr0(1.32 1.32) hr1(1 1
> ) accrue(710 710 710) arms(4 4 4) t(3 3) s(0.9 0.9) tstop(7)

n-stage trial design                        version 4.0.1, 2 Nov 2018
─────────────────────────────────────────────────────────────────────

Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
─────────────────────────────────────────────────────────────────────

Note: I outcome and D outcome are identical
Median survival time: 19.7 time units

Operating characteristics
──────────────────────────────────────────────────────────────────────────────
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
──────────────────────────────────────────────────────────────────────────────
1           0.5000    0.952    1.320    1.000    1.320    4.910    4.910
2           0.2500    0.950    1.320    1.000    1.218    2.066    6.976
3           0.0200    0.900    1.320    1.000    1.113    4.134   11.110
──────────────────────────────────────────────────────────────────────────────
Pairwise Error Rate          0.0171              Pairwise Power   0.8572
Familywise Error Rate (SE)   0.0441 (0.0004)
──────────────────────────────────────────────────────────────────────────────
Note: patient accrual stopped at time  7.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ─────────Stage 1─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      710      178      533
Patients*     3487      872     2615
Events**       287       71      216
           ─────────Stage 2─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      710      178      533
Patients*     4952     1238     3714
Events**       560      140      420
           ─────────Stage 3─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      710      178      533
Patients*     4971     1243     3728
Events**      1156      289      867
           ──────────────────────────
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE

.
end of do-file
```

4% absolute NI margin

```
. nstage, nstage(3) alpha(0.5 0.25 0.02) omega(0.95 0.95 0.9) hr0(1.43 1.43) hr1(1 1
> ) accrue(430 430 430) arms(4 4 4) t(3 3) s(0.9 0.9) tstop(7)


n-stage trial design                        version 4.0.1, 2 Nov 2018
─────────────────────────────────────────────────────────────────────

Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
─────────────────────────────────────────────────────────────────────

Note: I outcome and D outcome are identical
Median survival time: 19.7 time units

Operating characteristics
──────────────────────────────────────────────────────────────────────────────
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
──────────────────────────────────────────────────────────────────────────────
1           0.5000    0.952    1.430    1.000    1.430    4.910    4.910
2           0.2500    0.951    1.430    1.000    1.289    2.075    6.985
3           0.0200    0.900    1.430    1.000    1.147    4.075   11.059
──────────────────────────────────────────────────────────────────────────────
Pairwise Error Rate          0.0172              Pairwise Power   0.8580
Familywise Error Rate (SE)   0.0446 (0.0004)
──────────────────────────────────────────────────────────────────────────────
Note: patient accrual stopped at time  7.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ─────────Stage 1─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      430      108      323
Patients*     2111      528     1583
Events**       175       43      132
           ─────────Stage 2─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      430      108      323
Patients*     3004      751     2253
Events**       340       85      255
           ─────────Stage 3─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      430      108      323
Patients*     3010      752     2258
Events**       696      174      522
           ──────────────────────────
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE

.
end of do-file
```

5% absolute NI margin

```
. nstage, nstage(3) alpha(0.5 0.25 0.02) omega(0.95 0.95 0.9) hr0(1.54 1.54) hr1(1 1
> ) accrue(295 295 295) arms(4 4 4) t(3 3) s(0.9 0.9) tstop(7)


n-stage trial design                        version 4.0.1, 2 Nov 2018
─────────────────────────────────────────────────────────────────────

Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
─────────────────────────────────────────────────────────────────────

Note: I outcome and D outcome are identical
Median survival time: 19.7 time units

Operating characteristics
──────────────────────────────────────────────────────────────────────────────
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
──────────────────────────────────────────────────────────────────────────────
1           0.5000    0.954    1.540    1.000    1.540    4.953    4.953
2           0.2500    0.951    1.540    1.000    1.359    2.012    6.965
3           0.0200    0.902    1.540    1.000    1.181    4.139   11.104
──────────────────────────────────────────────────────────────────────────────
Pairwise Error Rate          0.0172              Pairwise Power   0.8600
Familywise Error Rate (SE)   0.0450 (0.0004)
──────────────────────────────────────────────────────────────────────────────
Note: patient accrual stopped at time  7.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ─────────Stage 1─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      295       74      221
Patients*     1461      365     1096
Events**       123       30       93
           ─────────Stage 2─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      295       74      221
Patients*     2055      514     1541
Events**       232       58      174
           ─────────Stage 3─────────
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      295       74      221
Patients*     2065      516     1549
Events**       480      120      360
           ──────────────────────────
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE

.
end of do-file
```

Appendix 12: MAMS designs with strong control of the PWER applied with different absolute NI margins

2% absolute NI margin

```
. nstage, nstage(3) alpha(0.5 0.25 0.05) omega(0.95 0.95 0.9) hr0(1.21 1.21) hr1
> (1 1) accrue(1500 1500 1500) arms(4 4 4) t(3 3) s(0.9 0.9) tstop(7)


n-stage trial design                        version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 19.7 time units

Operating characteristics

------------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
------------------------------------------------------------------------------
1           0.5000    0.950    1.210    1.000    1.210    4.893    4.893
2           0.2500    0.950    1.210    1.000    1.145    2.098    6.991
3           0.0500    0.900    1.210    1.000    1.087    2.225    9.216
------------------------------------------------------------------------------
Pairwise Error Rate          0.0433              Pairwise Power   0.8628
Familywise Error Rate (SE)   0.1043 (0.0006)
------------------------------------------------------------------------------
Note: patient accrual stopped at time  7.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
                ----------Stage 1----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate     1500      375     1125
Patients*     7340     1835     5505
Events**       599      149      450
                ----------Stage 2----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate     1500      375     1125
Patients*    10485     2621     7864
Events**      1188      297      891
                ----------Stage 3----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate     1500      375     1125
Patients*    10500     2625     7875
Events**      1891      472     1419
                ------------------------------
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE

.
end of do-file
```

3% absolute NI margin

```
. nstage, nstage(3) alpha(0.5 0.25 0.05) omega(0.95 0.95 0.9) hr0(1.32 1.32) hr1
> (1 1) accrue(710 710 710) arms(4 4 4) t(3 3) s(0.9 0.9) tstop(7)

n-stage trial design                        version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 19.7 time units

Operating characteristics
------------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
------------------------------------------------------------------------------
1           0.5000    0.952    1.320    1.000    1.320    4.910    4.910
2           0.2500    0.950    1.320    1.000    1.218    2.066    6.976
3           0.0500    0.901    1.320    1.000    1.130    2.229    9.204
------------------------------------------------------------------------------
Pairwise Error Rate          0.0433              Pairwise Power   0.8640
Familywise Error Rate (SE)   0.1041 (0.0006)
------------------------------------------------------------------------------

Note: patient accrual stopped at time  7.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ----------Stage 1----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      710      178      533
Patients*     3487      872     2615
Events**       287       71      216
           ----------Stage 2----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      710      178      533
Patients*     4952     1238     3714
Events**       560      140      420
           ----------Stage 3----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      710      178      533
Patients*     4971     1243     3728
Events**       892      223      669
           ------------------------------

 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE

.
end of do-file
```

4% absolute NI margin

```
. nstage, nstage(3) alpha(0.5 0.25 0.05) omega(0.95 0.95 0.9) hr0(1.43 1.43) hr1
> (1 1) accrue(430 430 430) arms(4 4 4) t(3 3) s(0.9 0.9) tstop(7)

n-stage trial design                        version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 19.7 time units

Operating characteristics

------------------------------------------------------------------------------
Stage  Alpha(LOB)*    Power    HR|H0    HR|H1  Crit.HR Length**   Time**
------------------------------------------------------------------------------
1           0.5000    0.952    1.430    1.000    1.430    4.910    4.910
2           0.2500    0.951    1.430    1.000    1.289    2.075    6.985
3           0.0500    0.901    1.430    1.000    1.170    2.171    9.155
------------------------------------------------------------------------------
Pairwise Error Rate          0.0435             Pairwise Power   0.8648
Familywise Error Rate (SE)   0.1046 (0.0006)
------------------------------------------------------------------------------
Note: patient accrual stopped at time  7.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
           ----------Stage 1----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      430      108      323
Patients*     2111      528     1583
Events**       175       43      132
           ----------Stage 2----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      430      108      323
Patients*     3004      751     2253
Events**       340       85      255
           ----------Stage 3----------
           Overall  Control   Exper.
Arms             4        1        3
Acc. rate      430      108      323
Patients*     3010      752     2258
Events**       539      134      405
           ---------------------------
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE

.
end of do-file
```

5% absolute NI margin

```
. nstage, nstage(3) alpha(0.5 0.25 0.05) omega(0.95 0.95 0.9) hr0(1.54 1.54) hr1
> (1 1) ///
> accrue(294 294 294) arms(4 4 4) t(3 3) s(0.9 0.9) tstop(7)


n-stage trial design                        version 4.0.1, 2 Nov 2018
-------------------------------------------------------------------------------
Sample size for a 4-arm 3-stage trial with time-to-event outcome
based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al.
(2019) Clinical Trials 16(2)
-------------------------------------------------------------------------------

Note: I outcome and D outcome are identical
Median survival time: 19.7 time units

Operating characteristics
```

Stage	Alpha(LOB)*	Power	HR|H0	HR|H1	Crit.HR	Length**	Time**
1	0.5000	0.954	1.540	1.000	**1.540**	**4.961**	**4.961**
2	0.2500	0.951	1.540	1.000	**1.359**	**2.016**	**6.977**
3	0.0500	0.901	1.540	1.000	**1.208**	**2.204**	**9.181**

Pairwise Error Rate **0.0435** Pairwise Power **0.8661**
Familywise Error Rate (SE) **0.1052 (0.0006)**

```
Note: patient accrual stopped at time  7.000
 *   All alphas are one-sided
 **  Length (duration of each stage) is expressed in  periods and
     assumes survival times are exponentially distributed. Time is
     expressed in cumulative periods.

Sample size and number of events
```

	Overall	Stage 1 Control	Stage 1 Exper.
Arms	4	1	3
Acc. rate	294	74	221
Patients*	**1459**	**365**	**1094**
Events**	**123**	**30**	**93**
	Overall	**Stage 2 Control**	**Stage 2 Exper.**
Arms	4	1	3
Acc. rate	294	74	221
Patients*	**2051**	**513**	**1538**
Events**	**232**	**58**	**174**
	Overall	**Stage 3 Control**	**Stage 3 Exper.**
Arms	4	1	3
Acc. rate	294	74	221
Patients*	**2059**	**515**	**1544**
Events**	**371**	**92**	**279**

```
 *  Patients are cumulative across stages
 ** Events are cumulative across stages, but are only displayed
    for those arms to which patients are still being recruited
 ** Events are for the same outcome at all 3 stages
END OF NSTAGE

.
.
end of do-file
```

10 Bibliography

Abraham, J. (2017). Platinum and PARP inhibitor for Neoadjuvant treatment of Triple NEgative and/or BRCA positive breast cancer (On Behalf of the Partner TMG), The University of Cambridge.

Abraham, J., Vallier, A., Qian, W., Machin, A., Grybowicz, L., Thomas, S., Weiss, M., Harvey, C., McAdam, K. and Hughes-Davies, L. (2019). Abstract OT3-03-03: PARTNER: Randomised, phase II/III trial to evaluate the safety and efficacy of the addition of olaparib to platinum-based neoadjuvant chemotherapy in triple negative and/or germline BRCA mutated breast cancer patients, American Association of Cancer Research.

Ajithkumar, T. and Gilbert, D. (2017). "Modern challenges of cancer clinical trials." Clinical Oncology **29**(12): 767-769.

Akobeng, A. (2005). "Understanding randomised controlled trials." Archives of diseases in childhood **90**(8): 840-844.

Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A. and Esserman, L. J. (2009). "I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy." Clinical Pharmacology & Therapeutics **86**(1): 97-100.

Barthel, F., Parmar, M. K. and Royston, P. (2009). "How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design–a reanalysis of 4 trials." Trials **10**(1): 21.

Barthel, F., Royston, P. and Parmar, M. K. (2009). "A menu-driven facility for sample-size calculation in novel multiarm, multistage randomized controlled trials with a time-to-event outcome." Stata Journal **9**(4): 505.

Bauer, P. (1997). "A note on multiple testing procedures in dose finding." Biometrics: 1125-1128.

Bauer, P., Bretz, F., Dragalin, V., König, F. and Wassmer, G. (2016). "Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls." Statistics in Medicine **35**(3): 325-347.

Bauer, P. and Einfalt, J. (2006). "Application of adaptive designs–a review." Biometrical journal **48**(4): 493-506.

Bauer, P. and Kohne, K. (1994). "Evaluation of experiments with adaptive interim analyses." Biometrics: 1029-1041.

Becker, B. J. (1994). "Combining significance levels." The handbook of research synthesis: 215-230.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F. and Simel, D. (1996). "Improving the quality of reporting of randomized controlled trials: the CONSORT statement." Jama **276**(8): 637-639.

Berchialla, P., Gregori, D. and Baldi, I. (2019). "The Role of Randomization in Bayesian and Frequentist Design of Clinical Trial." Topoi **38**(2): 469-475.

Berry Consultants. (2012). "Fixed and Adaptive Clinical Trial Simulator, FACTS." Retrieved February 18, 2021, from https://www.berryconsultants.com/software/.

Berry, D. A. (2012). "Adaptive clinical trials in oncology." Nature Reviews Clinical Oncology **9**(4): 199-207.

Berry, D. A. (2015). "The Brave New World of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research." Molecular oncology **9**(5): 951-959.

Berry, S. M., Carlin, B. P., Lee, J. J. and Muller, P. (2010). Bayesian adaptive methods for clinical trials, CRC press.

Berry, S. M., Spinelli, W., Littman, G. S., Liang, J. Z., Fardipour, P., Berry, D. A., Lewis, R. J. and Krams, M. (2010). "A Bayesian dose-finding trial with adaptive dose expansion to flexibly assess efficacy and safety of an investigational drug." Clinical Trials **7**(2): 121-135.

Bhatt, D. L. and Mehta, C. (2016). "Adaptive designs for clinical trials." New England Journal of Medicine **375**(1): 65-74.

Birkett, M. A. and Day, S. J. (1994). "Internal pilot studies for estimating sample size." Statistics in medicine **13**(23-24): 2455-2463.

Bland, J. M. and Bland, D. (1994). "Statistics Notes: One and two sided tests of significance." BMJ **309**(6949): 248.

Blenkinsop, A. and Choodari-Oskooei, B. (2019). "Multiarm, multistage randomized controlled trials with stopping boundaries for efficacy and lack of benefit: An update to nstage." The Stata Journal **19**(4): 782-802.

Blenkinsop, A., Parmar, M. K. and Choodari-Oskooei, B. (2019). "Assessing the impact of efficacy stopping rules on the error rates under the multi-arm multi-stage framework." Clinical Trials **16**(2): 132-141.

Bolognese, J. (2017). "Case Study: TRV130 Phase 2 Adaptive Dose-Finding Trial in Bunionectomy (via COMPASS®)." Retrieved 9 September, 2020, from https://www.cytel.com/hubfs/2017_Events/EUGM%2017/EUGM-2017-Adaptive-Design-Monitoring-Bolognese.pdf.

Bond, S., Mander, A., Jodrell, D. and Cheryian, J. (2011). "Model selection and sample size adaptation: HYPAZ trial." Trials **12**.

Booth, C. M., Cescon, D. W., Wang, L., Tannock, I. F. and Krzyzanowska, M. K. (2008). "Evolution of the randomized controlled trial in oncology over three decades." Journal of Clinical Oncology **26**(33): 5458.

Bothwell, L. E., Avorn, J., Khan, N. F. and Kesselheim, A. S. (2018). "Adaptive design clinical trials: a review of the literature and ClinicalTrials. gov." BMJ Open **8**(2): e018320.

Brannath, W., Posch, M. and Bauer, P. (2002). "Recursive combination tests." Journal of the American Statistical Association **97**(457): 236-244.

Bratton, D. J., Choodari-Oskooei, B. and Royston, P. (2015). "A menu-driven facility for sample-size calculation in multiarm, multistage randomized controlled trials with time-to-event outcomes: Update." Stata Journal **15**(2): 350-368.

Bratton, D. J., Parmar, M. K., Phillips, P. P. and Choodari-Oskooei, B. (2016). "Type I error rates of multi-arm multi-stage clinical trials: strong control and impact of intermediate outcomes." Trials **17**(1): 309.

Bratton, D. J., Phillips, P. P. J. and Parmar, M. K. B. (2013). "A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis." Bmc Medical Research Methodology **13**: 14.

Breast Cancer. (2016, July 15, 2016). "Breast Cancer Tests: Screening, Diagnosis, and Monitoring." Retrieved 20 September, 2017, from http://www.breastcancer.org/symptoms/testing/types.

Bresalier, R. S., Sandler, R. S., Quan, H., Bolognese, J. A., Oxenius, B., Horgan, K., Lines, C., Riddell, R., Morton, D. and Lanas, A. (2005). "Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial." New England Journal of Medicine **352**(11): 1092-1102.

Bretz, F. and Koenig, F. (2020). "Commentary on Parker and Weir." Clinical Trials **17**(5): 567-569.

Bretz, F., Koenig, F., Brannath, W., Glimm, E. and Posch, M. (2009). "Adaptive designs for confirmatory clinical trials." Statistics in medicine **28**(8): 1181-1217.

Bruno, A., Durkalski, V. L., Hall, C. E., Juneja, R., Barsan, W. G., Janis, S., Meurer, W. J., Fansler, A. and Johnston, K. C. (2014). "The Stroke Hyperglycemia Insulin Network Effort (SHINE) trial protocol: a randomized, blinded, efficacy trial of standard vs. intensive hyperglycemia management in acute stroke." International Journal of Stroke **9**(2): 246-251.

Brutti, P., De Santis, F. and Gubbiotti, S. (2009). "Mixtures of prior distributions for predictive Bayesian sample size calculations in clinical trials." Statistics in medicine **28**(17): 2185-2201.

Burki, T. K. (2019). "NHS England statistics and long-term plan." The Lancet Oncology **20**(2): 187.

Burton, A., Altman, D. G., Royston, P. and Holder, R. L. (2006). "The design of simulation studies in medical statistics." Statistics in medicine **25**(24): 4279-4292.

Cameron, D., Piccart-Gebhart, M. J., Gelber, R. D., Procter, M., Goldhirsch, A., de Azambuja, E., Castro Jr, G., Untch, M., Smith, I. and Gianni, L. (2017). "11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial." The Lancet **389**(10075): 1195-1205.

Campbell, M. K., Elbourne, D. R. and Altman, D. G. (2004). "CONSORT statement: extension to cluster randomised trials." BMJ **328**(7441): 702-708.

Cancer Research UK. (2018, 23 January 2018). "Cancer incidence for common cancers." Retrieved 4 January, 2019, from http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared#heading-Zero.

Cancer Research UK. (2018). "Find a clinical trial." Retrieved 8 February, 2018, from http://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial.

Cancer Research UK. (2020). "A trial looking at different treatments for people with oropharyngeal cancer (CompARE)." Retrieved 19 October, 2020, from https://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/a-trial-looking-at-different-treatments-for-people-with-oropharyngeal-cancer-compare#undefined.

Carthon, B. C. and Antonarakis, E. S. (2016). "The STAMPEDE trial: paradigm-changing data through innovative trial design." Translational cancer research **5**(3 Suppl): S485.

Casali, P. G., Le Cesne, A., Velasco, A. P., Kotasek, D., Rutkowski, P., Hohenberger, P., Fumagalli, E., Judson, I. R., Italiano, A., Gelderblom, H., Adenis, A., Hartmann, J. T., Duffaud, F., Goldstein, D., Broto, J. M., Gronchi, A., Dei Tos, A. P., Marreaud, S., Van Der Graaf, W. T. A., Zalcberg, J. R., Litiere, S. and Blay, J. Y. (2015). "Time to definitive failure to the first tyrosine kinase inhibitor in localized GI stromal tumors treated with imatinib as an adjuvant: A European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group Intergroup Randomized Trial in collaboration with the Australasian Gastro-Intestinal Trials Group, UNICANCER, French Sarcoma Group, Italian Sarcoma Group, and Spanish Group for Research on Sarcomas." Journal of Clinical Oncology **33**(36): 4276-4283.

Chan, A., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., Hróbjartsson, A., Mann, H., Dickersin, K. and Berlin, J. A. (2013). "SPIRIT 2013 statement: defining standard protocol items for clinical trials." Annals of internal medicine **158**(3): 200-207.

Chang, J. C., Makris, A., Gutierrez, M. C., Hilsenbeck, S. G., Hackett, J. R., Jeong, J., Liu, M., Baker, J., Clark-Langone, K. and Baehner, F. L. (2008). "Gene expression patterns in formalin-fixed, paraffin-embedded core biopsies predict docetaxel chemosensitivity in breast cancer patients." Breast cancer research and treatment **108**(2): 233-240.

Chang, M. (2014). Adaptive design theory and implementation using SAS and R, CRC Press.

Chen, T. T. (1997). "Optimal three-stage designs for phase II cancer clinical trials." Statistics in medicine **16**(23): 2701-2711.

CHMP. (2012). "CHMP reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design." Retrieved 8 March, 2016, from https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf.

Choodari-Oskooei, B., Bratton, D. J., Gannon, M. R., Meade, A. M., Sydes, M. R. and Parmar, M. K. (2020). "Adding new experimental arms to randomised clinical trials: Impact on error rates." Clinical Trials: 1740774520904346.

Chow, S. and Chang, M. (2008). "Adaptive design methods in clinical trials -A Review." Orphanet Journal Rare Disease **3**(11): 169-190.

Chow, S. C. (2014). Adaptive clinical trial design, Annual Reviews Inc.**:** 405-415.

Chow, S. C. and Chang, M. (2011). Adaptive Design Methods in Clinical Trials, Second Edition, CRC Press.

Chuang-Stein, C., Anderson, K., Gallo, P. and Collins, S. (2006). "Sample size reestimation: a review and recommendations." Therapeutic Innovation & Regulatory Science **40**(4): 475.

Clark, T., Berger, U. and Mansmann, U. (2013). "Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees." BMJ **346**.

ClinicalTrials.gov. (2020). "US National Library of Medicine, ClinicalTrials.gov." Retrieved 28 March, 2020, from https://clinicaltrials.gov/ct2/home.

Collett, D. (2015). Modelling survival data in medical research, Chapman and Hall/CRC.

Comprehensive Cancer Centers. (2017). "Why Clinical Research Is Important for the Future of Cancer Treatment." Retrieved 19 April, 2020, from https://cccnevada.com/clinical-research-important-future-cancer-treatment/.

CONSORT. (2010). "Consolidated Standards of Reporting Trials." Retrieved 10 April, 2017, from http://www.consort-statement.org/.

Cook, J. A., Julious, S. A., Sones, W., Hampson, L. V., Hewitt, C., Berlin, J. A., Ashby, D., Emsley, R., Fergusson, D. A. and Walters, S. J. (2018). "DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial." BMJ **363**: k3750.

Coombes, R. C., Bliss, J. M., Espie, M., Erdkamp, F., Wals, J., Tres, A., Marty, M., Coleman, R. E., Tubiana-Mathieu, N. and den Boer, M. O. (2011). "Randomized, phase III trial of sequential epirubicin and docetaxel versus epirubicin alone in postmenopausal patients with node-positive breast cancer." Journal of Clinical Oncology **29**(24): 3247-3254.

Corporation, C. S. (2000). "EaSt 2000: A Software Package for the Design and Interim Monitoring of Group-Sequential Clinical Trials."

Corrie, P. G., Marshall, A., Dunn, J. A., Middleton, M. R., Nathan, P. D., Gore, M., Davidson, N., Nicholson, S., Kelly, C. G. and Marples, M. (2014). "Adjuvant bevacizumab in patients with melanoma at high risk of recurrence (AVAST-M): preplanned interim results from a multicentre, open-label, randomised controlled phase 3 study." The Lancet Oncology **15**(6): 620-630.

Cox, C., Chu, H., Schneider, M. F. and Munoz, A. (2007). "Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution." Statistics in medicine **26**(23): 4352-4374.

Coyle, C., Cafferty, F. H., Rowley, S., MacKenzie, M., Berkman, L., Gupta, S., Pramesh, C., Gilbert, D., Kynaston, H. and Cameron, D. (2016). "ADD-ASPIRIN: A phase III, double-blind, placebo controlled, randomised trial assessing the effects of aspirin on disease recurrence and survival after primary therapy in common non-metastatic solid tumours." Contemporary clinical trials **51**: 56-64.

CPMP (2001). "Points to consider on switching between superiority and non-inferiority." British journal of clinical pharmacology **52**(3): 223-228.

Cui, L., Hung, H. and Wang, S. J. (1999). "Modification of sample size in group sequential clinical trials." Biometrics **55**(3): 853-857.

De Santis, A., Villani, R., Sinisi, M., Stocchetti, N. and Perucca, E. (2002). "Add-on phenytoin fails to prevent early seizures after surgery for supratentorial brain tumors: a randomized controlled study." Epilepsia **43**(2): 175-182.

De Santis, F. (2006). "Sample size determination for robust Bayesian analysis." Journal of the American Statistical Association **101**(473): 278-291.

Deng, H., Du, X., Wang, L. and Chen, M. (2020). "Six Months vs. 12 Months of Adjuvant Trastuzumab Among Women With HER2-Positive Early-Stage Breast Cancer: A Meta-Analysis of Randomized Controlled Trials." Frontiers in oncology **10**: 288.

Department of Health. (2000). "The NHS Cancer Plan: a plan for investment, a plan for reform." Retrieved 21 January, 2019, from https://webarchive.nationalarchives.gov.uk/20120503231511/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4009609?PageOperation=email.

Dimairo, M., Coates, E., Pallmann, P., Todd, S., Julious, S. A., Jaki, T., Wason, J., Mander, A. P., Weir, C. J. and Koenig, F. (2018). "Development process of a consensus-driven CONSORT extension for randomised trials using an adaptive design." BMC Medicine **16**(1): 210.

Dimairo, M., Pallmann, P., Wason, J., Todd, S., Jaki, T., Julious, S. A., Mander, A. P., Weir, C. J., Koenig, F. and Walton, M. K. (2019). Introducing the adaptive designs CONSORT extension (ACE) statement to improve reporting of randomised trials that use an adaptive design. Trials, BioMed Central.

Dimairo, M., Pallmann, P., Wason, J., Todd, S., Jaki, T., Julious, S. A., Mander, A. P., Weir, C. J., Koenig, F. and Walton, M. K. (2020). "The Adaptive designs CONSORT Extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design." BMJ **369**.

Drayson, M. T., Bowcock, S., Planche, T., Iqbal, G., Pratt, G., Yong, K., Wood, J., Raynes, K., Higgins, H. and Dawkins, B. (2019). "Prophylactic levofloxacin to prevent

infections in newly diagnosed symptomatic myeloma: the TEAMM RCT." Open Research Exeter.

Dunn, D. T., Copas, A. J. and Brocklehurst, P. (2018). "Superiority and non-inferiority: two sides of the same coin?" Trials **19**(1): 499.

Dunnett, C. W. (1955). "A multiple comparison procedure for comparing several treatments with a control." Journal of the American Statistical Association **50**(272): 1096-1121.

Earl, H., Hiller, L., Vallier, A.-L., Loi, S., McAdam, K., Hughes-Davies, L., Rea, D., Howe, D., Raynes, K. and Higgins, H. B. (2020). "Six versus 12 months' adjuvant trastuzumab in patients with HER2-positive early breast cancer: the PERSEPHONE non-inferiority RCT." Health Technology Assessment (Winchester, England) **24**(40): 1.

Earl, H. M., Hiller, L., Dunn, J. A., Blenkinsop, C., Grybowicz, L., Vallier, A.-L., Abraham, J., Thomas, J., Provenzano, E. and Hughes-Davies, L. (2015). "Efficacy of neoadjuvant bevacizumab added to docetaxel followed by fluorouracil, epirubicin, and cyclophosphamide, for women with HER2-negative early breast cancer (ARTemis): an open-label, randomised, phase 3 trial." The Lancet Oncology **16**(6): 656-666.

Earl, H. M., Hiller, L., Vallier, A.-L., Loi, S., McAdam, K., Hughes-Davies, L., Harnett, A. N., Ah-See, M.-L., Simcock, R. and Rea, D. (2019). "6 versus 12 months of adjuvant trastuzumab for HER2-positive early breast cancer (PERSEPHONE): 4-year disease-free survival results of a randomised phase 3 non-inferiority trial." The Lancet.

Ensign, L. G., Gehan, E. A., Kamen, D. S. and Thall, P. F. (1994). "An optimal three-stage design for phase II clinical trials." Statistics in medicine **13**(17): 1727-1736.

Esserman, L. J., Berry, D. A., Cheang, M. C., Yau, C., Perou, C. M., Carey, L., DeMichele, A., Gray, J. W., Conway-Dorsey, K. and Lenburg, M. E. (2012). "Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657)." Breast cancer research and treatment **132**(3): 1049-1062.

European Medicines Agency, E. (1998). "ICH Topic E9 Statistical Principles for Clinical Trials." Retrieved 23 November, 2018, from https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf.

European Medicines Agency, E. (2002). "Points to Consider on Multiplicity Issues in Clinical Trials." Retrieved 1 May, 2021, from https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-multiplicity-issues-clinical-trials_en.pdf.

Everitt, B. S. and Palmer, C. (2011). Encyclopaedic companion to medical statistics, John Wiley & Sons.

Ferrarotto, R., Redman, M. W., Gandara, D. R., Herbst, R. S. and Papadimitrakopoulou, V. A. (2015). "Lung-MAP—framework, overview, and design principles." Chinese clinical oncology **4**(3).

Filipits, M., Nielsen, T. O., Rudas, M., Greil, R., Stöger, H., Jakesz, R., Bago-Horvath, Z., Dietze, O., Regitnig, P. and Gruber-Rossipal, C. (2014). "The PAM50 risk-of-recurrence score predicts risk for late distant recurrence after endocrine therapy in postmenopausal women with endocrine-responsive early breast cancer." Clinical Cancer Research **20**(5): 1298-1305.

Fisher, L. D. (1998). "Self-designing clinical trials." Statistics in medicine **17**(14): 1551-1562.

Fisher, R. A. (1992). Statistical methods for research workers. Breakthroughs in statistics, Springer**:** 66-70.

FOCUS4. (2014). "Molecular selection of therapy in colorectal cancer: a molecularly-stratified randomised controlled trial programme." Retrieved 4 April, 2017, from http://www.focus4trial.org/.

FOCUS4. (2019). "FOCUS4 – Molecular selection of therapy in colorectal cancer: a molecularly-stratified randomised controlled trial programme protocol." Retrieved 23 Feburary, 2020, from http://www.focus4trial.org/centres/protocols/.

Ford, H. E., Marshall, A., Bridgewater, J. A., Janowitz, T., Coxon, F. Y., Wadsley, J., Mansoor, W., Fyfe, D., Madhusudan, S. and Middleton, G. W. (2014). "Docetaxel versus active symptom control for refractory oesophagogastric adenocarcinoma (COUGAR-02): an open-label, phase 3 randomised controlled trial." The Lancet Oncology **15**(1): 78-86.

Frei, E., Holland, J. F., Schneiderman, M. A., Pinkel, D., Selkirk, G., Freireich, E. J., Silver, R. T., Gold, G. L. and Regelson, W. (1958). "A comparative study of two regimens of combination chemotherapy in acute leukemia." Blood **13**(12): 1126-1148.

Gajewski, B. J. and Mayo, M. S. (2006). "Bayesian sample size calculations in phase II clinical trials using a mixture of informative priors." Statistics in medicine **25**(15): 2554-2566.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M. and Pinheiro, J. (2006). "Adaptive designs in clinical drug development—an executive summary of the PhRMA working group." Journal of biopharmaceutical statistics **16**(3): 275-283.

Garrett-Mayer, E. (2006). "The continual reassessment method for dose-finding studies: a tutorial." Clinical Trials **3**(1): 57-71.

Gehan, E. A. (1979). "Clinical trials in cancer research." Environmental Health Perspectives **32**: 31-48.

George, S. L., Wang, X. and Pang, H. (2016). Cancer Clinical Trials: Current and Controversial Issues in Design and Analysis, CRC Press.

Gerber, D., Oxnard, G. and Govindan, R. (2015). "ALCHEMIST: Bringing genomic discovery and targeted therapies to early-stage lung cancer." Clinical Pharmacology & Therapeutics **97**(5): 447-450.

Gianni, L., Dafni, U., Gelber, R. D., Azambuja, E., Muehlbauer, S., Goldhirsch, A., Untch, M., Smith, I., Baselga, J. and Jackisch, C. (2011). "Treatment with trastuzumab for 1 year after adjuvant chemotherapy in patients with HER2-positive early breast cancer: a 4-year follow-up of a randomised controlled trial." The Lancet Oncology **12**(3): 236-244.

Gianni, L., Zambetti, M., Clark, K., Baker, J., Cronin, M., Wu, J., Mariani, G., Rodriguez, J., Carcangiu, M. and Watson, D. (2005). "Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer." Journal of clinical oncology **23**(29): 7265-7277.

Gilbert, D. C., Duong, T., Sydes, M., Bara, A., Clarke, N., Abel, P., James, N., Langley, R. and Parmar, M. (2018). "Transdermal oestradiol as a method of androgen suppression for prostate cancer within the STAMPEDE trial platform." BJU International **121**(5): 680-683.

Goldhirsch, A., Gelber, R. D., Piccart-Gebhart, M. J., De Azambuja, E., Procter, M., Suter, T. M., Jackisch, C., Cameron, D., Weber, H. A. and Heinzmann, D. (2013). "2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): an open-label, randomised controlled trial." The Lancet **382**(9897): 1021-1028.

Gordon Lan, K. and DeMets, D. L. (1983). "Discrete sequential boundaries for clinical trials." Biometrika **70**(3): 659-663.

Gould, A. L. (1995). "Planning and revising the sample size for a trial." Statistics in Medicine **14**(9): 1039-1051.

Goutis, C., Casella, G. and Wells, M. T. (1996). "Assessing evidence in multiple hypotheses." Journal of the American Statistical Association **91**(435): 1268-1277.

Grayling, M. J. (2017). "GROUPSEQ: Stata module providing functions to determine group sequential trial designs." Retrieved 10 April, 2019, from https://ideas.repec.org/c/boc/bocode/s458345.html.

Grayling, M. J. (2019). "DESMA: Stata module to design and simulate (adaptive) multi-arm clinical trials." Retrieved 17 May, 2020, from https://econpapers.repec.org/.

Gulia, S., Kannan, S., Badwe, R. and Gupta, S. (2020). "Evaluation of 1-year vs shorter durations of adjuvant trastuzumab among patients with early breast cancer: An individual participant data and trial-level meta-analysis." JAMA network open **3**(8): e2011777-e2011777.

Hague, D., Townsend, S., Masters, L., Rauchenberger, M., Van Looy, N., Diaz-Montana, C., Gannon, M., James, N., Maughan, T. and Parmar, M. K. (2019). "Changing platforms without stopping the train: experiences of data management

and data management systems when adapting platform protocols by adding and closing comparisons." Trials **20**(1): 294.

Hardwick, J. and Stout, Q. F. (1991). "Bandit strategies for ethical sequential allocation." Computing Science and Statistics **23**(421-424): 6.1.

Hatfield, I., Allison, A., Flight, L., Julious, S. A. and Dimairo, M. (2016). "Adaptive designs undertaken in clinical research: a review of registered clinical trials." Trials **17**(1): 1.

Haybittle, J. (1971). "Repeated assessment of results in clinical trials of cancer treatment." The British journal of radiology **44**(526): 793-797.

Hills, R. K. and Burnett, A. K. (2011). "Applicability of a "Pick a Winner" trial design to acute myeloid leukemia." Blood, The Journal of the American Society of Hematology **118**(9): 2389-2394.

Hind, D., Reeves, B. C., Bathers, S., Bray, C., Corkhill, A., Hayward, C., Harper, L., Napp, V., Norrie, J. and Speed, C. (2017). "Comparative costs and activity from a sample of UK clinical trials units." Trials **18**(1): 203.

Hirakawa, A., Asano, J., Sato, H. and Teramukai, S. (2018). "Master protocol trials in oncology: review and new trial designs." Contemporary clinical trials communications **12**: 1-8.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure." Scandinavian journal of statistics: 65-70.

Hornberger, J., Alvarado, M. D., Rebecca, C., Gutierrez, H. R., Yu, T. M. and Gradishar, W. J. (2012). "Clinical validity/utility, change in practice patterns, and economic implications of risk stratifiers to predict outcomes for early-stage breast cancer: a systematic review." Journal of the National Cancer Institute **104**(14): 1068-1079.

Howlader, N., Altekruse, S. F., Li, C. I., Chen, V. W., Clarke, C. A., Ries, L. A. G. and Cronin, K. A. (2014). "US Incidence of Breast Cancer Subtypes Defined by Joint Hormone Receptor and HER2 Status." JNCI: Journal of the National Cancer Institute **106**(5): dju055-dju055.

Hu, F. and Rosenberger, W. F. (2003). "Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons." Journal of the American Statistical Association **98**(463): 671-678.

Hu, F. and Rosenberger, W. F. (2006). The theory of response-adaptive randomization in clinical trials, John Wiley & Sons.

Hudis, C. A., Barlow, W. E., Costantino, J. P., Gray, R. J., Pritchard, K. I., Chapman, J.-A. W., Sparano, J. A., Hunsberger, S., Enos, R. A. and Gelber, R. D. (2007). "Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system." Journal of Clinical Oncology **25**(15): 2127-2132.

Inoue, L. Y., Thall, P. F. and Berry, D. A. (2002). "Seamlessly expanding a randomized phase II trial to phase III." Biometrics **58**(4): 823-831.

Jaki, T. F., Pallmann, P. S. and Magirr, D. (2019). "The R package MAMS for designing multi-arm multi-stage clinical trials." Journal of Statistical Software **88**(4).

James, N. D., de Bono, J. S., Spears, M. R., Clarke, N. W., Mason, M. D., Dearnaley, D. P., Ritchie, A. W., Amos, C. L., Gilson, C. and Jones, R. J. (2017). "Abiraterone for prostate cancer not previously treated with hormone therapy." New England Journal of Medicine **377**(4): 338-351.

James, N. D., Spears, M. R., Clarke, N. W., Dearnaley, D. P., De Bono, J. S., Gale, J., Hetherington, J., Hoskin, P. J., Jones, R. J. and Laing, R. (2015). "Survival with newly diagnosed metastatic prostate cancer in the "docetaxel era": data from 917 patients in the control arm of the STAMPEDE trial (MRC PR08, CRUK/06/019)." European urology **67**(6): 1028-1038.

James, N. D., Sydes, M. R., Clarke, N. W., Mason, M. D., Dearnaley, D. P., Anderson, J., Popert, R. J., Sanders, K., Morgan, R. C., Stansfeld, J., Dwyer, J., Masters, J. and Parmar, M. K. B. (2008). "STAMPEDE: Systemic Therapy for Advancing or Metastatic Prostate Cancer - A Multi-Arm Multi-Stage Randomised Controlled Trial." Clinical Oncology **20**(8): 577-581.

James, N. D., Sydes, M. R., Clarke, N. W., Mason, M. D., Dearnaley, D. P., Spears, M. R., Ritchie, A. W., Parker, C. C., Russell, J. M. and Attard, G. (2016). "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 **387**(10024): 1163-1177.

James, N. D., Sydes, M. R., Mason, M. D., Clarke, N. W., Dearnaley, D. P., Dwyer, J., Jovic, G., Russell, J. M., Thalmann, G. and Parmar, M. K. B. (2011). "Celecoxib plus hormone therapy vs hormone therapy alone for hormone-sensitive prostate cancer: First results from the STAMPEDE randomised controlled trial (MRC PR08)." European Journal of Cancer **47**: 11.

Jennison, C. and Turnbull, B. W. (1999). Group sequential methods with applications to clinical trials, CRC Press.

Jennison, C. and Turnbull, B. W. (2015). "Adaptive sample size modification in clinical trials: Start small then ask for more?" Statistics in Medicine **34**(29): 3793-3810.

Joensuu, H., Bono, P., Kataja, V., Alanko, T., Kokko, R., Asola, R., Utriainen, T., Turpeenniemi-Hujanen, T., Jyrkkiö, S. and Möykkynen, K. (2009). "Fluorouracil, epirubicin, and cyclophosphamide with either docetaxel or vinorelbine, with or without trastuzumab, as adjuvant treatments of breast cancer: final results of the FinHer Trial." Journal of Clinical Oncology **27**(34): 5685-5692.

Joensuu, H., Fraser, J., Wildiers, H., Huovinen, R., Auvinen, P., Utriainen, M., Nyandoto, P., Villman, K. K., Halonen, P. and Granstam-Björneklett, H. (2018). "Effect of Adjuvant Trastuzumab for a Duration of 9 Weeks vs 1 Year With Concomitant Chemotherapy for Early Human Epidermal Growth Factor Receptor 2–Positive Breast Cancer: The SOLD Randomized Clinical Trial." JAMA Oncology.

Johnston, S. R., Kilburn, L. S., Ellis, P., Dodwell, D., Cameron, D., Hayward, L., Im, Y., Braybrooke, J. P., Brunt, A. M. and Cheung, K. (2013). "Fulvestrant plus anastrozole or placebo versus exemestane alone after progression on non-steroidal aromatase inhibitors in postmenopausal patients with hormone-receptor-positive locally advanced or metastatic breast cancer (SoFEA): a composite, multicentre, phase 3 randomised trial." The Lancet Oncology **14**(10): 989-998.

Kahan, B. C. and Morris, T. P. (2012). "Improper analysis of trials randomised using stratified blocks or minimisation." Statistics in medicine **31**(4): 328-340.

Kairalla, J. A., Coffey, C. S., Thomann, M. A. and Muller, K. E. (2012). "Adaptive trial designs: a review of barriers and opportunities." Trials **13**(1): 145.

Kerr, R. S., Love, S., Segelov, E., Johnstone, E., Falcon, B., Hewett, P., Weaver, A., Church, D., Scudder, C. and Pearson, S. (2016). "Adjuvant capecitabine plus bevacizumab versus capecitabine alone in patients with colorectal cancer (QUASAR 2): an open-label, randomised phase 3 trial." The Lancet Oncology **17**(11): 1543-1557.

Kiche, J., Ngesa, O. and Orwa, G. (2019). "On Generalized Gamma Distribution and Its Application to Survival Data." International Journal of Statistics and Probability **8**(5): 85-102.

Kilpatrick, G. and Oldham, P. (1954). "Calcium chloride and adrenaline as bronchial dilators compared by sequential analysis." BMJ **2**(4901): 1388.

Komen, S. G. (2017, 31 October 2017). "Molecular subtypes of breast cancer." Retrieved 09 April 2018, from https://ww5.komen.org/BreastCancer/SubtypesofBreastCancer.html.

Kyle, W. (2021). Introduction to Platform Trials, VP Science Strategy and Innovation**:** 41.

Langley, R. (2020). "Prostate Adenocarcinoma: TransCutaneous Hormones (PATCH09) Protocol." Retrieved 24 November, 2020, from https://www.ctu.mrc.ac.uk/studies/all-studies/p/patch-pr09/.

Lee, J. J., Gu, X. and Liu, S. (2010). "Bayesian adaptive randomization designs for targeted agent development." Clinical Trials **7**(5): 584-596.

Lee, K. H., Kim, J.-Y., Lee, M. H., Han, H. S., Lim, J. H., Park, K. U., Park, I. H., Cho, E. K., Yoon, S. Y. and Kim, J. H. (2016). "A randomized, multicenter, phase II/III study to determine the optimal dose and to evaluate the efficacy and safety of pegteograstim (GCPGC) on chemotherapy-induced neutropenia compared to pegfilgrastim in breast cancer patients: KCSG PC10-09." Supportive Care in Cancer **24**(4): 1709-1717.

Lee, S. M., Khan, I., Upadhyay, S., Lewanski, C., Falk, S., Skailes, G., Marshall, E., Woll, P. J., Hatton, M. and Lal, R. (2012). "First-line erlotinib in patients with advanced non-small-cell lung cancer unsuitable for chemotherapy (TOPICAL): a double-blind, placebo-controlled, phase 3 trial." The Lancet Oncology **13**(11): 1161-1170.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J. and Moher, D. (2009). "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration." Journal of clinical epidemiology **6**(7).

Liu, Q. and Chi, G. Y. (2001). "On sample size and inference for two-stage adaptive designs." Biometrics **57**(1): 172-177.

Liu, S. and Lee, J. J. (2015). "An overview of the design and conduct of the BATTLE trials." Chinese clinical oncology **4**(3): 33.

Long, E. R. and Ferebee, S. H. (1950). "A controlled investigation of streptomycin treatment in pulmonary tuberculosis." Public Health Reports: 1421-1451.

LUNG-MAP. (2017). "A groundbreaking collaborative approach to clinical trials." Retrieved 17th November, 2017, from http://www.lung-map.org.

Luo, M., Parasrampuria, D. A., Usmani, S. Z., Mateos, M., Nahi, H., Heuck, C., Qi, M., Sun, S., Clemens, P. L. and Zhou, H. (2019). Daratumumab (DARA) Subcutaneous (SC) Delivery in Relapsed or Refractory Multiple Myeloma (RRMM): Population Pharmacokinetics (PPK) and Exposure-Response (ER) Analysis, American Society of Hematology Washington, DC.

Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P. and Krams, M. (2006). "Adaptive seamless phase II/III designs-background, operational aspects, and examples." Therapeutic Innovation & Regulatory Science **40**(4): 463.

Macbeth, F., Noble, S., Evans, J., Ahmed, S., Cohen, D., Hood, K., Knoyle, D., Linnane, S., Longo, M. and Moore, B. (2015). "Randomized phase III trial of standard therapy plus low molecular weight heparin in patients with lung cancer: FRAGMATIC trial." Journal of Clinical Oncology **34**(5): 488-494.

Machin, D., Cheung, Y.B., Parmar, M., (2006). Parametric Modelling. Survival Analysis**:** 91-119.

MacPherson, H., Altman, D. G., Hammerschlag, R., Youping, L., Taixiang, W., White, A. and Moher, D. (2010). "Revised standards for reporting interventions in clinical trials of acupuncture (STRICTA): extending the CONSORT statement." Journal of evidence-based medicine **3**(3): 140-155.

Mammo-50. (2016). "Mammographic surveillance in breast cancer patients aged 50 years or older." Retrieved 07 February, 2018, from https://warwick.ac.uk/fac/med/research/ctu/trials/cancer/mammo50/.

Marchionni, L., Wilson, R. F., Marinopoulos, S. S., Wolff, A. C., Parmigiani, G., Bass, E. B. and Goodman, S. N. (2008). "Impact of gene expression profiling tests on breast cancer outcomes." Evidence report/technology assessment

Mason, M. D., Clarke, N. W., James, N. D., Dearnaley, D. P., Spears, M. R., Ritchie, A. W., Attard, G., Cross, W., Jones, R. J. and Parker, C. C. (2017). "Adding celecoxib with or without zoledronic acid for hormone-naïve prostate cancer: long-term

survival results from an adaptive, multiarm, multistage, platform, randomized controlled trial." Journal of clinical oncology **35**(14): 1530.

Maughan, T. S., Adams, R. A., Smith, C. G., Meade, A. M., Seymour, M. T., Wilson, R. H., Idziaszczyk, S., Harris, R., Fisher, D. and Kenny, S. L. (2011). "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**(9783): 2103-2114.

Mavroudis, D., Saloustros, E., Malamos, N., Kakolyris, S., Boukovinas, I., Papakotoulas, P., Kentepozidis, N., Ziras, N. and Georgoulias, V. (2015). "Six versus 12 months of adjuvant trastuzumab in combination with dose-dense chemotherapy for women with HER2-positive breast cancer: a multicenter randomized study by the Hellenic Oncology Research Group (HORG)." Annals of Oncology **26**(7): 1333-1340.

McCabe, L., White, I. R., Chau, N. V. V., Barnes, E., Pett, S. L., Cooke, G. S. and Walker, A. S. (2020). "The design and statistical aspects of VIETNARMS: a strategic post-licensing trial of multiple oral direct-acting antiviral hepatitis C treatment strategies in Vietnam." Trials **21**: 1-12.

McCaul, J. A., Cymerman, J. A., Hislop, S., McConkey, C., McMahon, J., Mehanna, H., Shaw, R., Sutton, D. N. and Dunn, J. (2013). "LIHNCS-Lugol's iodine in head and neck cancer surgery: a multicentre, randomised controlled trial assessing the effectiveness of Lugol's iodine to assist excision of moderate dysplasia, severe dysplasia and carcinoma in situ at mucosal resection margins of oral and oropharyngeal squamous cell carcinoma: study protocol for a randomised controlled trial." Trials **14**(1): 310.

Mehanna, H. (2015). "CompARE trial: Phase III randomised controlled trial comparing alternative regiments for escalating treatment of intermediate and high-risk oropharyngeal cancer." Retrieved April, from https://www.inhanse.org/patient-research/compare/.

Mehanna, H., McConkey, C. C., Rahman, J. K., Wong, W., Smith, A. F., Nutting, C., Hartley, A. G., Hall, P., Hulme, C. and Patel, D. K. (2017). "PET-NECK HTA Report."

Mehanna, H., Wong, W., McConkey, C. C., Rahman, J. K., Robinson, M., Hartley, A. G., Nutting, C., Powell, N., Al-Booz, H. and Robinson, M. (2016). "PET-CT surveillance versus neck dissection in advanced head and neck cancer." New England Journal of Medicine **374**(15): 1444-1454.

Midgley, R. S., McConkey, C. C., Johnstone, E. C., Dunn, J. A., Smith, J. L., Grumett, S. A., Julier, P., Iveson, C., Yanagisawa, Y. and Warren, B. (2010). "Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial." Journal of Clinical Oncology **28**(30): 4575-4580.

Mistry, P., Dunn, J. A. and Marshall, A. (2017a). "A literature review of applied adaptive design methodology within the field of oncology in randomised controlled trials and a proposed extension to the CONSORT guidelines." BMC Medical Research Methodology **17**(1): 108.

Mistry, P., Dunn, J. A. and Marshall, A. (2017b) "A literature review of the use of adaptive design methods in oncology trials." Meeting abstracts from the 4th International Clinical Trials Methodology Conference (ICTMC) and the 38th Annual Meeting of the Society for Clinical Trials, 40.

Moher, D., Schulz, K. F., Altman, D. G. and Group, C. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials, Elsevier.

Moore, K. N. and Mannel, R. S. (2016). "Is the NCI MATCH trial a match for gynecologic oncology?" Gynecologic oncology **140**(1): 161-166.

Morrell, L., Hordern, J., Brown, L., Sydes, M. R., Amos, C. L., Kaplan, R. S., Parmar, M. K. and Maughan, T. S. (2019). "Mind the gap? The platform trial as a working environment." Trials **20**(1): 1-6.

Mosteller, F., Bush, R. R. and Green, B. F. (1954). Selected quantitative techniques, Addison-Wesley.

National Cancer Institute. (2017). "NCI-MATCH Trial (Molecular Analysis for Therapy Choice)." Retrieved 17th November, 2017, from https://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/nci-match.

Noor, N. M., Pett, S. L., Esmail, H., Crook, A. M., Vale, C. L., Sydes, M. R. and Parmar, M. K. (2020). "Adaptive platform trials using multi-arm, multi-stage protocols: getting fast answers in pandemic settings." F1000Research **9**(1109): 1109.

O'Brien, P. C. and Fleming, T. R. (1979). "A multiple testing procedure for clinical trials." Biometrics: 549-556.

O'Quigley, J., Pepe, M. and Fisher, L. (1990). "Continual reassessment method: a practical design for phase 1 clinical trials in cancer." Biometrics: 33-48.

Office for National Statistics. (2019, 12 August 2019). "Cancer survival in England: adult, stage at diagnosis and childhood - patients followed up to 2018." Retrieved 01 April, 2020, from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancersurvivalinengland/stageatdiagnosisandchildhoodpatientsfollowedupto2018.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D. and Park, T. (2004). "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer." New England Journal of Medicine **351**(27): 2817-2826.

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Sydes, M. R. and Villar, S. S. (2018). "Adaptive designs in clinical trials: why use them, and how to run and report them." BMC medicine **16**(1): 1-15.

Park, J. J., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K. and Mills, E. J. (2019). "Systematic review of basket trials, umbrella

trials, and platform trials: a landscape analysis of master protocols." Trials **20**(1): 1-10.

Parker, C. C., James, N. D., Brawley, C. D., Clarke, N. W., Hoyle, A. P., Ali, A., Ritchie, A. W., Attard, G., Chowdhury, S. and Cross, W. (2018). "Radiotherapy to the primary tumour for newly diagnosed, metastatic prostate cancer (STAMPEDE): a randomised controlled phase 3 trial." The Lancet **392**(10162): 2353-2366.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X. and Hu, Z. (2009). "Supervised risk predictor of breast cancer based on intrinsic subtypes." Journal of clinical oncology **27**(8): 1160-1167.

Parker, R. A. and Weir, C. J. (2020). "Non-adjustment for multiple testing in multi-arm trials of distinct treatments: Rationale and justification." Clinical Trials **17**(5): 562-566.

Parmar, M. K., Barthel, F., Sydes, M., Langley, R., Kaplan, R., Eisenhauer, E., Brady, M., James, N., Bookman, M. A. and Swart, A.-M. (2008). "Speeding up the evaluation of new agents in cancer." Journal of the National Cancer Institute **100**(17): 1204-1214.

Parmar, M. K., Carpenter, J. and Sydes, M. R. (2014). "More multiarm randomised trials of superiority are needed." The Lancet **384**(9940): 283.

Parmar, M. K., Sydes, M. R., Cafferty, F. H., Choodari-Oskooei, B., Langley, R. E., Brown, L., Phillips, P. P., Spears, M. R., Rowley, S. and Kaplan, R. (2017). "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 **14**(5): 451-461.

Persephone. (2007). "A phase III, randomised trial comparing six months Herceptin® treatment with twelve months, in patients with early stage breast cancer." Retrieved 7 August, 2018, from https://warwick.ac.uk/fac/med/research/ctu/trials/cancer/persephone.

Peto, R., Pike, M., Armitage, P., Breslow, N. E., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J. and Smith, P. (1977). "Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples." British journal of cancer **35**(1): 1-39.

Piccart-Gebhart, M. J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., Gianni, L., Baselga, J., Bell, R. and Jackisch, C. (2005). "Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer." New England Journal of Medicine **353**(16): 1659-1672.

Pivot, X., Romieu, G., Debled, M., Pierga, J., Kerbrat, P., Bachelot, T., Lortholary, A., Espié, M., Fumoleau, P. and Serin, D. (2013). "6 months versus 12 months of adjuvant trastuzumab for patients with HER2-positive early breast cancer (PHARE): a randomised phase 3 trial." The Lancet Oncology **14**(8): 741-748.

Pivot, X., Romieu, G., Debled, M., Pierga, J., Kerbrat, P., Bachelot, T., Lortholary, A., Espié, M., Fumoleau, P. and Serin, D. (2019). "6 months versus 12 months of

adjuvant trastuzumab in early breast cancer (PHARE): final analysis of a multicentre, open-label, phase 3 randomised trial." The Lancet **393**(10191): 2591-2598.

Pocock, S. J. (1977). "Group sequential methods in the design and analysis of clinical trials." Biometrika: 191-199.

Pocock, S. J. (1992). "When to stop a clinical trial." BMJ **305**(6847): 235.

Pocock, S. J. (2005). "When (not) to stop a clinical trial for benefit." JAMA **294**(17): 2228-2230.

Poole, C. J., Earl, H. M., Hiller, L., Dunn, J. A., Bathers, S., Grieve, R. J., Spooner, D. A., Agrawal, R. K., Fernando, I. N. and Brunt, A. M. (2006). "Epirubicin and cyclophosphamide, methotrexate, and fluorouracil as adjuvant therapy for early breast cancer." New England Journal of Medicine **355**(18): 1851-1862.

Poole, C. J., Marshall, A., Higgins, H. B., Fletcher, J., Williams, S. J., Lo, N., Fernando, I. N., Osborne, R., Crawford, S. M. and Rafii, S. (2012). "Neo-escape: Neoadjuvant extended sequential chemotherapy with adjuvant postoperative treatment for epithelial nonmucinous advanced inoperable peritoneal malignancy." J Clin Oncol.

Posch, M. and Bauer, P. (1999). "Adaptive two stage designs and the conditional error function." Biometrical Journal **41**(6): 689-696.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C. and Bauer, P. (2005). "Testing and estimation in flexible group sequential designs with adaptive treatment selection." Statistics in medicine **24**(24): 3697-3714.

Prat, A., Galván, P., Jimenez, B., Buckingham, W., Jeiranian, H. A., Schaper, C., Vidal, M., Álvarez, M., Díaz, S. and Ellis, C. (2016). "Prediction of response to neoadjuvant chemotherapy using core needle biopsy samples with the prosigna assay." Clinical Cancer Research **22**(3): 560-566.

Proschan, M. A. and Hunsberger, S. A. (1995). "Designed extension of studies based on conditional power." Biometrics: 1315-1324.

Quaresma, M., Coleman, M. P. and Rachet, B. (2015). "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 **385**(9974): 1206-1218.

Quartagno, M., Walker, A. S., Carpenter, J. R., Phillips, P. P. and Parmar, M. K. (2018). "Rethinking non-inferiority: a practical trial design for optimising treatment duration." Clinical Trials **15**(5): 477-488.

Ramalingam, S., Crawford, J., Chang, A., Manegold, C., Perez-Soler, R., Douillard, J. Y., Thatcher, N., Barlesi, F., Owonikoko, T., Wang, Y., Pultar, P. and Zhu, J. (2013). "Talactoferrin alfa versus placebo in patients with refractory advanced non-small-cell lung cancer (FORTIS-M trial)." Annals of Oncology **24**(11): 2875-2880.

Redman, M. W. and Allegra, C. J. (2015). The master protocol concept. Seminars in oncology, Elsevier.

Reitsma, D., Combest, A., rgen Hummel, J. and Simmons, A. (2015). "Improving Oncology Trials Through Adaptive Designs." Applied Clinical Trials **35**(3).

Renfro, L. A. and Mandrekar, S. J. (2017). "Definitions and statistical properties of master protocols for personalized medicine in oncology." Journal of Biopharmaceutical Statistics: 1-12.

Riechelmann, R., Alex, A., Cruz, L., Bariani, G. and Hoff, P. (2013). "Non-inferiority cancer clinical trials: scope and purposes underlying their design." Annals of Oncology **24**(7): 1942-1947.

Rosner, G. L. and Tsiatis, A. A. (1989). "The impact that group sequential tests would have made on ECOG clinical trials." Statistics in Medicine **8**(4): 505-516.

Royston, P. and Barthel, F. (2010). "Projection of power and events in clinical trials with a time-to-event outcome." Stata Journal **10**(3): 386.

Royston, P., Barthel, F. M., Parmar, M. K., Choodari-Oskooei, B. and Isham, V. (2011). "Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit." Trials **12**(1): 81.

Saad, E. D. (2018). Superiority and Non-inferiority Phase III Oncology Trials. Methods and Biostatistics in Oncology, Springer**:** 203-216.

Sacks, H., Chalmers, T. C. and Smith Jr, H. (1982). "Randomized versus historical controls for clinical trials." The American journal of medicine **72**(2): 233-240.

Sargent, D. J. and Goldberg, R. M. (2001). "A flexible design for multiple armed screening trials." Statistics in Medicine **20**(7): 1051-1060.

Saville, B. R. and Berry, S. M. (2016). "Efficiencies of platform clinical trials: a vision of the future." Clinical Trials **13**(3): 358-366.

ScHARR CTRU. (2020). "A Practical Adaptive & Novel Designs and Analysis (PANDA) toolkit." Retrieved 19 November, 2020, from https://www.sheffield.ac.uk/scharr/research/centres/ctru/panda.

Scher, H. I. and Heller, G. (2002). "Picking the winners in a sea of plenty." Clinical cancer research **8**(2): 400-404.

Schiavone, F., Bathia, R., Letchemanan, K., Masters, L., Amos, C., Bara, A., Brown, L., Gilson, C., Pugh, C. and Atako, N. (2019). "This is a platform alteration: a trial management perspective on the operational aspects of adaptive and platform and umbrella protocols." Trials **20**(1): 264.

Schmidt, C. (2007). "Adaptive design may hasten clinical trials." Journal of the National Cancer Institute **99**(2): 108-109.

Schofield, P., Ugalde, A., Carey, M., Mileshkin, L., Duffy, M., Ball, D. and Aranda, S. (2008). "Lung cancer: challenges and solutions for supportive care intervention research." Palliative & supportive care **6**(3): 281-287.

Schulz, K. F. and Grimes, D. A. (2005). "Multiplicity in randomised trials II: subgroup and interim analyses." The Lancet **365**(9471): 1657-1661.

Sestak, I., Dowsett, M., Zabaglo, L., Lopez-Knowles, E., Ferree, S., Cowens, J. W. and Cuzick, J. (2013). "Factors predicting late recurrence for estrogen receptor–positive breast cancer." Journal of the National Cancer Institute **105**(19): 1504-1511.

Seymour, M. T., Brown, S. R., Middleton, G., Maughan, T., Richman, S., Gwyther, S., Lowe, C., Seligmann, J. F., Wadsley, J. and Maisey, N. (2013). "Panitumumab and irinotecan versus irinotecan alone for patients with KRAS wild-type, fluorouracil-resistant advanced colorectal cancer (PICCOLO): a prospectively stratified randomised trial." The Lancet Oncology **14**(8): 749-759.

Seymour, M. T., Maughan, T. S., Ledermann, J. A., Topham, C., James, R., Gwyther, S. J., Smith, D. B., Shepherd, S., Maraveyas, A. and Ferry, D. R. (2007). "Different strategies of sequential and combination chemotherapy for patients with poor prognosis advanced colorectal cancer (MRC FOCUS): a randomised controlled trial." The Lancet **370**(9582): 143-152.

Shapiro, S. A., Smith, C. G., Arthurs, J. R. and Master, Z. (2019). Preparing regenerative therapies for clinical application: proposals for responsible translation, Future Medicine.

Shiu, K. K., Maughan, T., Wilson, R. H., Adams, R. A., Pugh, C., Brown, L., Fisher, D., Wasan, H., Middleton, G. W., Steward, W. P. and Kaplan, R. S. (2013) "FOCUS4: A prospective molecularly stratified, adaptive multicenter program of randomized controlled trials for patients with colorectal cancer (CRC)." Journal of clinical oncology **31**.

Simon, R. (1989). "Optimal two-stage designs for phase II clinical trials." Controlled clinical trials **10**(1): 1-10.

Smartt, P. (2010). A comparison of gene expression profiling tests for breast cancer, HSAC Report.

Smith, I., Procter, M., Gelber, R. D., Guillaume, S., Feyereislova, A., Dowsett, M., Goldhirsch, A., Untch, M., Mariani, G. and Baselga, J. (2007). "2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial." The Lancet **369**(9555): 29-36.

Smith, J. M. (1958). "Hydrocortisone hemisuccinate by inhalation in children with asthma." The Lancet **272**(7059): 1248-1250.

Snell, E. and Armitage, P. (1957). "Clinical Comparison of Diamorphine and Pholcodine as Cough Suprresants: By a New Method of Sequential Analysis." The Lancet **269**(6974): 860-862.

Stallard, N. (2010). "A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information." Statistics in Medicine **29**(9): 959-971.

Stallard, N., Hampson, L., Benda, N., Brannath, W., Burnett, T., Friede, T., Kimani, P. K., Koenig, F., Krisam, J. and Mozgunov, P. (2020). "Efficient adaptive designs for clinical trials of interventions for COVID-19." Statistics in Biopharmaceutical Research **12**(4): 483-497.

STAMPEDE trial. (2005). "Systemic Therapy in Advancing or Metastatic Prostate cancer: Evaluation of Drug Efficacy." Retrieved 23 Feb, 2017, from http://www.stampedetrial.org/.

StataCorp. (2017). Stata Statistical software: Release 15. College Station, TX: Statacorp LLC.

Stein, C. (1945). "A two-sample test for a linear hypothesis whose power is independent of the variance." The Annals of Mathematical Statistics **16**(3): 243-258.

Stein, R. C., Dunn, J. A., Bartlett, J. M., Campbell, A. F., Marshall, A., Hall, P., Rooshenas, L., Morgan, A., Poole, C. and Pinder, S. E. (2016). "OPTIMA prelim: a randomised feasibility study of personalised care in the treatment of women with early breast cancer." Health technology assessment (Winchester, England) **20**(10): 1.

Stevely, A., Dimairo, M., Todd, S., Julious, S. A., Nicholl, J., Hind, D. and Cooper, C. L. (2015). "An Investigation of the Shortcomings of the CONSORT 2010 Statement for the Reporting of Group Sequential Randomised Controlled Trials: A Methodological Systematic Review." Plos One **10**(11).

Sydes, M. R., James, N. D., Mason, M. D., Clarke, N. W., Amos, C., Anderson, J., De Bono, J., Dearnaley, D. P., Dwyer, J., Jovic, G., Ritchie, A., Russell, M., Sanders, K., Thalmann, G. and Parmar, M. K. B. (2011). "Flexible trial design in practice - Dropping and adding arms in STAMPEDE: A multi-arm multi-stage randomised controlled trial." Trials. Conference: Clinical Trials Methodology Conference **12**.

Sydes, M. R., Parmar, M. K. B., James, N. D., Clarke, N. W., Dearnaley, D. P., Mason, M. D., Morgan, R. C., Sanders, K. and Royston, P. (2009). "Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: The MRC STAMPEDE trial." Trials **10**(39).

Sydes, M. R., Parmar, M. K. B., Mason, M. D., Clarke, N. W., Amos, C., Anderson, J., de Bono, J., Dearnaley, D. P., Dwyer, J., Green, C., Jovic, G., Ritchie, A. W. S., Russell, J. M., Sanders, K., Thalmann, G. and James, N. D. (2012). "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 **13**(168).

Thall, P. F. and Wathen, J. K. (2007). "Practical Bayesian adaptive randomisation in clinical trials." European Journal of Cancer **43**(5): 859-866.

Thompson, W. R. (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples." Biometrika **25**(3/4): 285-294.

Thuss-Patience, P. C., Kretzschmar, A., Bichev, D., Deist, T., Hinke, A., Breithaupt, K., Dogan, Y., Gebauer, B., Schumacher, G. and Reichardt, P. (2011). "Survival advantage for irinotecan versus best supportive care as second-line chemotherapy in gastric cancer–a randomised phase III study of the Arbeitsgemeinschaft Internistische Onkologie (AIO)." European journal of cancer **47**(15): 2306-2314.

Todd, S. (2007). "A 25-year review of sequential methodology in clinical studies." Statistics in medicine **26**(2): 237-252.

Tsiatis, A. A. (1981). "The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time." Biometrika **68**(1): 311-315.

Ullibarri, I. L. d. (2013). LANDEMETS: Stata module to compute boundaries for group sequential clinical trials using alpha spending functions. Statistical Software Components, Boston College Department of Economics**:** Landemets implements the method of Lan and DeMets (Biometrika, 1983) to compute the stopping boundaries in a group sequential clinical trial by using alpha spending functions.

US Food and Drug Administration. (2007). "Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics." Retrieved 1 June, 2016, from https://www.fda.gov/downloads/Drugs/%7B%5Cprotect%5Cnewline%7DGuidanceComplianceRegulatoryInformation/Guidances/ucm071590.pdf.

US Food and Drug Administration. (2010). "Adaptive design clinical trials for drugs and biologics." Biotechnol Law Rep Retrieved 16 July, 2016, from https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.

US Food and Drug Administration. (2016). "Non-Inferiority Clinical Trials to Establish Effectiveness." Retrieved 02 January, 2020, from https://www.fda.gov/media/78504/download.

US Food and Drug Administration. (2018). "Adaptive Designs for Clinical Trials of Drugs and Biologics." Retrieved 14 October, 2015, from https://www.fda.gov/media/78495/download.

US Food and Drug Administration. (2018). "Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics. Guidance for Industry." Retrieved December 20, 2019, from https://www.fda.gov/regulatory-information/search-fda-guidance-documents/master-protocols-efficient-clinical-trial-design-strategies-expedite-development-oncology-drugs-and.

US Food and Drug Administration. (2019). "Statement from FDA Commissioner Scott Gottlieb, MD on FDA's comprehensive new policy approach to facilitating the development of innovative regenerative medicine products to improve human health." Retrieved 4 April, 2020, from https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-new-strategies-modernize-clinical-trials-advance.

Wald, A. (1947). Sequential Analysis, Wiley.

Walker, E. and Nowacki, A. S. (2011). "Understanding equivalence and noninferiority testing." Journal of general internal medicine **26**(2): 192-196.

Waller, D., Peake, M. D., Stephens, R. J., Gower, N. H., Milroy, R., Parmar, M. K. B., Rudd, R. M. and Spiro, S. G. (2004). "Chemotherapy for patients with non-small cell lung cancer: the surgical setting of the Big Lung Trial." European Journal of Cardio-Thoracic Surgery **26**(1): 173-182.

Wang, M. D. (2007). "Sample size reestimation by Bayesian prediction." Biometrical journal **49**(3).

Wang, S. K. and Tsiatis, A. A. (1987). "Approximately optimal one-parameter boundaries for group sequential trials." Biometrics: 193-199.

Ward, S., Scope, A., Rafia, R., Pandor, A., Harnan, S., Evans, P. and Wyld, L. (2013). "Gene expression profiling and expanded immunohistochemistry tests to guide the use of adjuvant chemotherapy in breast cancer management: a systematic review and cost-effectiveness analysis." Health technology assessment **17**(44).

Wason, J. (2015). Multi-Arm Multi-Stage Designs for Clinical Trials with Treatment Selection. Modern Adaptive Randomized Clinical Trials: Statistical and Practical Aspects, Chapman and Hall/CRC**:** 389-409.

Wason, J., Magirr, D., Law, M. and Jaki, T. (2016). "Some recommendations for multi-arm multi-stage trials." Statistical methods in medical research **25**(2): 716-727.

Wason, J. M., Stecher, L. and Mander, A. P. (2014). "Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?" Trials **15**(1): 1-7.

Weiskopf, R. B. and Cook, R. J. (2016). "Understanding non-inferiority trials and tests for the non-statistician: when to use them and how to interpret them." Transfusion **57**(2): 240-243.

West, H. J. (2017). "Novel Precision Medicine Trial Designs: Umbrellas and Baskets." JAMA oncology **3**(3): 423-423.

Wheeler, G. M., Mander, A. P., Bedding, A., Brock, K., Cornelius, V., Grieve, A. P., Jaki, T., Love, S. B., Weir, C. J. and Yap, C. (2019). "How to design a dose-finding study using the continual reassessment method." BMC medical research methodology **19**(1): 18.

Whitehead, J. (1997). The design and analysis of sequential clinical trials, John Wiley & Sons.

Whitehead, J. and Stratton, I. (1983). "Group sequential clinical trials with triangular continuation regions." Biometrics: 227-236.

Wilson, M. K., Collyar, D., Chingos, D. T., Friedlander, M., Ho, T. W., Karakasis, K., Kaye, S., Parmar, M. K., Sydes, M. R. and Tannock, I. F. (2015). "Outcomes and endpoints in cancer trials: bridging the divide." The Lancet Oncology **16**(1): e43-e52.

Wittes, J. and Brittain, E. (1990). "The role of internal pilot studies in increasing the efficiency of clinical trials." Statistics in medicine **9**(1-2): 65-72.

World Health Organisation. (2019). "Clinical Trials." Retrieved 02 April 2019, from https://www.who.int/topics/clinical_trials/en/.

Young, A. M., Marshall, A., Thirlwall, J., Chapman, O., Lokare, A., Hill, C., Hale, D., Dunn, J. A., Lyman, G. H. and Hutchinson, C. (2018). "Comparison of an oral factor Xa inhibitor with low molecular weight heparin in patients with cancer with venous

thromboembolism: results of a randomized trial (SELECT-D)." J Clin Oncol(36 (20)): 2017-2032.

Yu, K., Wang, X., Chen, W., Fan, L., Mo, M. and Chen, H. (2020). "Tailored duration of adjuvant trastuzumab for human epidermal growth factor receptor 2-positive breast cancer." NPJ precision oncology **4**(1): 1-7.

Zhang, W., Sargent, D. J. and Mandrekar, S. (2006). "An adaptive dose-finding design incorporating both toxicity and efficacy." Statistics in Medicine **25**(14): 2365-2383.

Zwarenstein, M., Treweek, S., Gagnier, J. J., Altman, D. G., Tunis, S., Haynes, B., Oxman, A. D. and Moher, D. (2008). "Improving the reporting of pragmatic trials: an extension of the CONSORT statement." BMJ **337**: a2390.