The Utility of Adaptive Designs in Publicly Funded Confirmatory Trials

Munyaradzi Dimairo

Thesis Submitted in Fulfillment for the Degree of Doctor of Philosophy

School of Health and Related Research

(ScHARR)

University of Sheffield

The University Of Sheffield.

Acknowledgements

This work would not have been completed without the support of individuals and organisations. First and foremost, I would like to express my gratitude to my supervisors (Prof Steven Julious, Prof Susan Todd, and Prof Jon Nicholl) for their invaluable advice and unwavering support during the course of this research. Their patience has been tested to the limit. I would like to thank the National Institute for Health Research (NIHR) for providing funding support, which afforded me the opportunity to undertake this research as part of a Doctoral Research Fellowship. I shall therefore remain indebted to the research community and my incredible supervisory team.

I acknowledge the contribution of a number of individuals who gave external advice or support that enhanced successful completion of this research. Mr Mike Bradburn gave continuous external advice and support as my personal tutor. Dr Jonathan Boote, Prof Cindy Cooper, Prof Alicia O'Cathain, and Dr Daniel Hind advised and supported this research at various stages as members of the advisory panel. Specifically, Dr Jonathan Boote and Prof Alicia O'Cathain provided qualitative support during the design and analysis of in-depth interviews. Helen Wakefield, Lauren O'Hara, and Kylie Cross provided in-house support with interview transcription prior to analysis. Dr Tracey Young gave advice on Rasch modelling during the analysis of survey response items. Prof Steve Goodacre, Prof Elizabeth Goyder, and Prof Alasdair Gray for consenting to the use of anonymised participant level trial data for retrospective case studies. For the prospective case studies of two grant applications presented, Prof Robert Storey and Mr Sabapathy Balasubramanian gave their consent and contributions as Chief Investigators among other co-applicants, particularly Dr Judith Cohen. I would like to thank the Sheffield Clinical Trials Research Unit (CTRU) mainly for the infrastructure and additional financial support.

My appreciation goes to a number of collaborators for their contribution, which enabled dissemination of research findings through peer reviewed journal papers. In addition to my supervisors, these include Laura Flight, Annabel Allison, Isabella Hatfield, Abigail Stevely, Dr Jonathan Boote, Dr Daniel Hind, and Prof Cindy Cooper. Their support and contribution is acknowledged in relevant sections of this thesis.

Last but not least, my deepest gratitude goes to participating key stakeholders and organisations both in the private and public sector. Their contribution to this research has been immense and its importance cannot be overemphasised as it facilitated better understanding of the research subject. Private sector support particularly that of Prof Frank Bretz deserves mentioning.

Dedication

To my departed, loving, and caring mother. You worked tirelessly for me to be where I am today. I am eternally grateful and indebted to your love you gave us as a family. I miss you every day and may God rest your soul in eternal peace. Tears may dry but memories will never fade. Until we meet again!!

Disclaimer and Author's Declaration

This research has been fully funded by the NIHR as a Doctoral Research Fellowship (Grant Number: NIHR DRF-2012-05-182). However, the views expressed are those of the author and not necessarily those of the National Health Service (NHS), the NIHR, the Department of Health or organisations the author is affiliated to. The author has also received additional funding support from the Sheffield CTRU and has no competing interests to declare.

The author declare that this thesis is his original work and that none of the material contained in this thesis has previously been submitted for a degree to any awarding institution. The work contained in this thesis has been undertaken by myself, with the support from those individuals or collaborators mentioned in my 'Acknowledgements' section. Their contribution is acknowledged at the beginning of each relevant chapter throughout this thesis and in published papers cited in my 'Research Achievements' section.

Research Achievements

Peer Reviewed Publications (Lead and/or Corresponding Author)

Dimairo M *, Julious SA, Todd S, Nicholl JP, Boote J. (2015) Cross-sector surveys assessing perceptions of key stakeholders towards barriers, concerns and facilitators to the appropriate use of adaptive designs in confirmatory trials. Trials 16 (1): 585. doi:10.1186/s13063-015-1119-x.

Dimairo M *, Boote J, Julious SA, Nicholl JP, Todd S. (2015): Missing steps in a staircase: a qualitative study of the perspectives of key stakeholders on the use of adaptive designs in confirmatory trials. Trials 16: 430. doi:10.1186/s13063-015-0958-9.

Stevely A, **Dimairo M** *, Todd S, Julious SA, Nicholl J, et al. (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: e0141104. doi:10.1371/journal.pone.0141104.

Hatfield I, Allison A, Flight L, Julious SA, **Dimairo M *** (2016): Adaptive designs undertaken in clinical research: a review of registered clinical trials. Trials 17 (1): 150. doi:10.1186/s13063-016-1273-9.

Relevant Peer Reviewed Publications (Co-author)

Teare MD, **Dimairo M**, Shephard N, Hayman A, Whitehead A, Walters SJ (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. Trials 15:264. doi: 10.1186/1745-6215-15-264.

Invited Oral Conferences or Meeting Presentations

Dimairo M *, Todd S, Julious S, Nicholl J: Meandering journey towards routine trial adaptation: Survey results on the use of adaptive designs in confirmatory trials. *PSI*. London, 10 - 13 May 2015. http://goo.gl/p2SPW1

Dimairo M *, Todd S, Julious S, Nicholl J: Roadblocks, concerns and potential facilitators to the appropriate use of adaptive designs in confirmatory trials. *UKCRC Registered CTU Network: Bi-annual Statisticians' Operational Group Meeting*. Sheffield, 5 October 2015.

Dimairo M *: Adaptive designs in confirmatory clinical trials: opportunities in investigating complex interventions. *Researching Complex Interventions in Health: The State of the Art*. Exeter, 14 - 15 October 2015. http://goo.gl/6AwF08

Contributed Oral Conference or Meeting Presentations

Dimairo M *, Stevely A, Julious S, Todd S, Cooper C, Hind D, Nicholl J: Differential reporting of group sequential randomised controlled trials: shortcomings of the CONSORT 2010 statement. *36th Annual Meeting of the Society for Clinical Trials*; 2015:91.

Dimairo M *, Julious S, Todd S, Nicholl J: Meandering journey towards routine trial adaptation: survey results on barriers to use of adaptive designs in confirmatory trials. *3rd International Clinical Trials Methodology Conference.* Trials 2015, 16(Suppl 2):O20. doi:10.1186/1745-6215-16-S2-O20.

Dimairo M *, Stevely A, Todd S, Julious S, Nicholl J, Hind D, Cooper C: Investigation of the shortcomings of the CONSORT 2010 statement for the reporting of group sequential randomised controlled trials. *3rd International Clinical Trials Methodology Conference*. Trials 2015, 16(Suppl 2):O53. doi:10.1186/1745-6215-16-S2-O53.

Teare D, **Dimairo M** *, Hayman A., Shephard N, Whitehead A, Walters, S: Sample size requirements for pilot randomised controlled trials with binary outcomes: a simulation study. *2nd International Clinical Trials Methodology Conference.* Trials 2013, 14(Suppl 1), O21. doi: 10.1186/1745-6215-14-S1-O21.

White Rose Repository and Commentary Reports

Dimairo M *, Stevely A, Todd S, Julious S, et al. (2014) Reporting issues in group sequential randomised controlled trials: a systematic review protocol of published journal reports. http://eprints.whiterose.ac.uk/88387/

Julious SA and **Dimairo M** (2013). Discussion on the paper by Hampson, L. V. & Jennison, C. (2013) Group sequential tests for delayed responses (with discussion). Journal of the Royal Statistical Society. Series B: Statistical Methodology75 (1), 3–54. doi:10.1111/j.1467-9868.2012.01030.x.

Cytel. (2016), 'Adaptive Designs in Practice: Interview with NIHR Research Fellow **Munya Dimairo'**, Cytel Blog, available at: http://goo.gl/EEZ6Ys (accessed 20 May 2016).

Other Research Contributions

Some research findings of Chapter 3 informed the application of Laura Flight's successful Doctoral Research Fellowship application (Grant Number: DRF-2015-08-013), which has been funded by NIHR.

Supervised a research attachment student (Abigail Stevely) and supported an NIHR Research Methods intern (Isabella Hatifield) as part of knowledge transfer and capacity building.

Abstract

Introduction: Adaptive designs (ADs) are underused, particularly in publicly funded confirmatory trials, despite their promising benefits and methodological prominence given in the statistical literature.

Research Question: This thesis investigates why ADs are underused in the publicly funded setting, explores facilitators, and proposes recommendations to improve their appropriate use.

Methods: Confirmatory ADs are reviewed from a statistical and practical perspective. Cross-disciplinary key stakeholders are then interviewed to explore roadblocks to the use of ADs. Based on the interview findings, follow-up quantitative surveys are undertaken to explore wider perceptions on barriers, concerns, and facilitators aimed to generalise the findings. The surveys targeted CTUs (Clinical Trials Units), private sector organisations, and Public Funders in the UK. In view of some of the findings, case studies of applied confirmatory ADs are reviewed to highlight their scope and characteristic, and to investigate the state of reporting of the most common AD. The design and implementation of selected ADs is demonstrated using retrospective and prospective planned case studies. Lessons learned are highlighted to enhance the design of future trials of similar characteristics.

Results: The main barriers to the use of ADs include the lack of funding support accessible to UK CTUs to aid their design; limited practical knowledge; preference for traditional mainstream designs; difficulties in marketing ADs to key stakeholders; limited time to support ADs relative to other competing priorities; lack of applied training; and insufficient access to case studies of undertaken ADs, which would facilitate practical learning and successful implementation. Researchers' inadequate description of AD-related aspects (such as rationale, scope, and decision-making criteria to guide the planned AD) in grant proposals was viewed among the major obstacles by Public Funders. Suboptimal reporting of the design and conduct of undertaken ADs appears to influence concerns about their robustness in decision-making and credibility to change practice.

Conclusions: Most obstacles appear connected to a lack of practical implementation knowledge and applied training, and limited access to adequately reported case studies to facilitate practical learning. Assurance of scientific rigour through transparent adequate reporting is paramount to the credibility of findings from adaptive trials. There is a need for a consensus guidance document on ADs and an AD-tailored CONSORT statement to enhance their reporting and conduct. This thesis provides detailed recommendations to improve the appropriate use of ADs and areas for future related research.

Thesis Contents

Chapter 1. Introduction 17

1.1 Background 17
1.2 Brief Historical Perspective on Trial Adaptation 18
1.3 Motivation 19
1.4 The Research Question and Rationale 20
1.5 Overarching Specific Objectives 20
1.6 Scope of the Research 21
1.7 Thesis Roadmap 21
1.8 Summary 22

Chapter 2. Literature Review 23

2.1 Introduction 23
2.2 Aims 23
2.3 Literature Search 23
2.4 Characterisation of Reviewed Confirmatory Adaptive Designs 25
2.5 Design 1: Sample Size Re-estimation 27
2.5.1 Introduction to Sample Size Estimation 27
2.5.2 Addressing Uncertainty Around Nuisance Parameters 28
2.5.3 Motivation for Sample Size Re-estimation 29
2.5.4 Early Research on Sample Size Re-estimation 30
2.5.5 Restricted Internal Pilot Concept 30
2.5.6 Unrestricted Internal Pilot Concept 31
2.5.7 When to Conduct Sample Size Re-estimation 31
2.5.8 The Frequency of Sample Size Re-estimation 34
2.5.9 Methods for Estimating Nuisance Design Parameters 35
2.5.10 Blinded Methods for Continuous Outcomes 35
2.5.11 Unblinded Methods for Continuous Outcomes 38
2.5.12 Blinded Methods for Binary Outcomes 40
2.5.13 Unblinded Methods for Binary Outcomes 41
2.5.14 Reflection on Practical Considerations 42
2.5.15 Reflection on Regulatory Considerations 43
2.5.16 Summary 43
2.6 Design 2: Stochastic Curtailment Futility Analysis 45
2.6.1 Motivation 45
2.6.2 Stochastic Curtailment 46
2.6.3 Conditional Power 46
2.6.4 Computation of Conditional Power 47
2.6.5 Statistical Properties of Conditional Power Futility Analysis 48
2.6.6 Futility Criteria for Decision-Making 48

2.6.7 Assumptions Regarding Future Trend of Results 49
2.6.8 Timing of Futility Analysis 50
2.6.9 Frequency of Futility Analysis and Impact on the Type I and II Errors 51
2.6.10 Statistical Software for Implementation 53
2.6.11 Limitations of Conditional Power Futility Analysis 53
2.6.12 Summary 54
2.7 Design 3: Group Sequential Design 55
2.7.1 Motivation 55
2.7.2 Description of the Methodology 55
2.7.3 Expression of Stopping Boundaries 57
2.7.4 Effect of Interim Analyses on Type I Error and Power 58
2.7.5 Stopping Boundaries 58
2.7.6 The Choice of Stopping Boundaries 63
2.7.7 Impact of Altering the Number and Timing of the Interim Analyses 65
2.7.8 Defining the Interim Information Fraction 65
2.7.9 The Number and Timing of Interim Analyses 66
2.7.10 Sample Size Estimation 66
2.7.11 Impact on Statistical Inference 67
2.7.12 Statistical Software for Implementation 73
2.7.13 Practical, Logistical and Administrative Aspects 74
2.7.14 Interim Decision-Making Challenges 75
2.7.15 Preserving Trial Integrity 76
2.7.16 Reflection 77
2.8 Design 4: Information Based Group Sequential Design 78
2.8.1 Motivation 78
2.8.2 Reflection 80
2.9 Design 5: Seamless Design 81
2.9.1 Motivation 81
2.9.2 Reflection 84
2.10 Design 6: Multi-Arm Multi-Stage Design 85
2.10.1 Motivation 85
2.10.2 Statistical and Practical Considerations 86
2.10.3 Reflection 87
Chapter 3. Interviews Exploring Roadblocks to the Use of Adaptive Designs 88
3.1 Introduction and Rationale 88
3.2 Aims and Objectives 89
3.3 Methods 89
3.3.1 Study Design 89
3.3.2 The Choice of the Sample Size 90

3.3.3 Selection of Interview Participants 90
3.3.4 The Process of Approaching Target Participants 91
3.3.5 Research Ethics and Consenting Respondents 92
3.3.6 The Interview Process 92
3.3.7 Analysis of Interviews and Reporting 93
3.3.8 Quality Control Process 94
3.4 Results of In-depth Qualitative Interviews 95
3.4.1 Demographics and Characteristics of Interviewees 95
3.4.2 Clinical Trials Research and Adaptive Designs Experiences of Interviewees 97
3.4.3 General Perceptions of Adaptive Designs in Confirmatory Trials 98
3.4.4 Perception of Themes on Barriers in Confirmatory Trials 103
3.5 Discussion 113

Chapter 4. Surveys on Perceptions of the Use of Confirmatory Adaptive Designs 116

4.1 Introduction 116
4.2 Aims 116
4.3 Methods 117
4.3.1 Study Design and Sampling Frame 117
4.3.2 The Rationale for Sample Size Approach 119
4.3.3 The Design of Online Survey Instruments 119
4.3.4 Use of Quality Control Measures 120
4.3.5 Approaching Target Participants 120
4.3.6 Research Ethics and Consenting Participants 121
4.3.7 Outline of Statistical Analysis and Reporting 121
4.4 Results 122
4.4.1 Response Rates 122
4.4.2 Demographics, Characteristics and Experiences of Respondents 123
4.4.3 Perceptions on Barriers to the Use of Confirmatory Adaptive Designs 124
4.4.4 Cross-sector Relating to Concerns on the Use of Confirmatory Adaptive Designs 130
4.4.5 Cross-sector Perceptions of Possible Facilitators 130
4.4.6 Organisational Priorities on Adaptive Design-Related Aspects 133
4.4.7 Survey Results on the Application of Confirmatory Adaptive Designs in the UK 133
4.5 Discussion 137
4.5.1 Main Findings and Interpretation 137
4.5.2 Relating Findings to Existing Literature 140
4.5.3 Strengths and Limitations 141
4.5.4 Implications for the Work Described in the Remainder of the Thesis 142

Chapter 5. Review of Case Studies of Confirmatory Adaptive Designs 143

5.1 Introduction 143
5.2 Aims and Objectives 143

5.3 Methods 144
5.3.1 The Rationale for the Literature Search 144
5.3.2 Scoping Exercise to Troubleshoot the Literature Search 144
5.3.3 Data Sources 145
5.3.4 Search Strategy 146
5.3.5 Eligibility Criteria 147
5.3.6 Data Extraction, Main Outcomes, and Statistical Analysis 147
5.3.7 Examination of the Adequacy of ClinicalTrials.gov in Capturing Adaptive Designs 148
5.4 Results 149
5.4.1 Trials Eligibility Screening 149
5.4.2 Characteristics of Identified Confirmatory Adaptive Designs 150
5.4.3 Description of the Reasons for Early Stopping 154
5.4.4 Exemplars and Classification of Identified Confirmatory Adaptive Designs 154
5.4.5 The Publication of Confirmatory Adaptive Designs 158
5.4.6 Geographical Distribution of Identified Confirmatory Adaptive Designs 158
5.4.7 Efficiency of ClinicalTrials.gov in Capturing Registered Adaptive Designs 159
5.5 Discussion 160
5.5.1 Main Findings and Implications 161
5.5.2 Main Strengths, Limitations, and Implications on Interpretation 162
5.5.3 Implications for the Research Described in the Remainder of the Thesis 163

Chapter 6. Transparency and Reporting of Confirmatory Adaptive Designs 164

6.1 Introduction 164
6.2 Aims and Objectives 165
6.3 Methods 166
6.3.1 Trials Eligibility Criteria 166
6.3.2 Searching the Literature and Data Sources 166
6.3.3 Data Extraction and Quality Control 168
6.3.4 Researcher-led Proposed CONSORT Items 168
6.3.5 Outcome Measures, Statistical Analysis and Reporting 169
6.4 Results 170
6.4.1 Eligibility Screening 170
6.4.2 Characteristics of Included Group Sequential Trials 171
6.4.3 Reporting of Universal CONSORT 2010 Checklist Items 174
6.4.4 Reporting of Group Sequential Specific Items and Proposed Modifications 179
6.4.5 Exemplars to Enhance the Reporting of Group Sequential Trials 184
6.5 Discussion 185
6.5.1 Main findings 186
6.5.2 Interpretation of the findings 190
6.5.3 Implications to Practice 190

6.5.4 Strengths and Limitations 192
6.5.5 Summary and Direction of the Remainder of the Thesis 193

Chapter 7. Design and Implementation of Retrospective Case Studies 194

7.1 Introduction 194
7.2 Aims and Objectives 194
7.3 Brief Description of the Retrospective Case Studies 195
7.3.1 RATPAC Trial 195
7.3.2 3CPO Trial 196
7.3.3 3Mg Trial 196
7.3.4 Booster Trial 197
7.4 Methods 198
7.4.1 Sample Size Re-estimation for Binary Outcomes 198
7.4.2 Conditional Power Based Stochastic Curtailment Futility Analysis 199
7.4.3 Group Sequential Design for Binary Outcomes 199
7.5 Results 200
7.5.1 Sample Size Re-estimation for Binary Outcomes 200
7.5.2 Conditional Power Based Stochastic Curtailment Futility Analysis 205
7.5.3 Group Sequential Design 214
7.5.4 Information Based Group Sequential Design for RATPAC trial 225
7.6 Discussion 228
7.6.1 Lessons Learned from Sample Size Re-estimation 228
7.6.2 Lessons Learned from Stochastic Curtailment Futility Analysis 230
7.6.3 Lessons Learned from Group Sequential Design 231
7.6.4 Lessons Learned from an Information Based Group Sequential Design 232
7.6.5 Reflection on Limitations 233
7.6.6 Direction of the Remainder of the Thesis 233

Chapter 8. Design and Planning of Prospective Case Studies 235

8.1 Introduction 235
8.2 Aims and Objectives 235
8.3 PENNYWISE Study 236
8.3.1 Brief Background 236
8.3.2 Study Design and Primary Endpoint 237
8.3.3 Primary Hypothesis and Sample Size Estimates 237
8.3.4 Selection of the Desired Design 241
8.3.5 Sensitivity Analysis of the Statistical Properties of the WT ($\boldsymbol{\partial = 0.25}$) Design 241
8.3.6 Grant Submission Exemplar of the Design and Sample Size Estimates 243
8.3.7 Costing of the Grant Application 244
8.4 NERVE BLOCK Study 244
8.4.1 Brief Background 244

8.4.2 Design Issues and Adaptive Aspects 245
8.4.3 Sample Size Estimates and Planned Analysis 248
8.4.4 Exemplar for Rationale and Costing of the Grant Application 249
8.5 Discussion 249
8.5.1 Reflection on Lessons Learned in the Context of Previous Thesis Findings 250
8.5.2 Reflection on Limitations 251
8.5.3 Direction of the Remainder of the Thesis 251

Chapter 9. Discussion and Recommendations 252

9.1 Introduction 252
9.2 The Main Thesis Findings 253
9.2.1 The Perspective of UK CTUs on Roadblocks 253
9.2.2 Cross-sector Differences in Perceptions on Roadblocks 254
9.2.3 The Perspective of UK Public Funders on Roadblocks 255
9.2.4 Paradigm Shift in Perceptions Towards Adaptive Designs 255
9.3 Recommendations for Best Practice 256
9.3.1 Description of Rationale, Type and Scope of the Proposed Adaptive Design 256
9.3.2 Adaptation by Design, Managed Scope and Design Properties 257
9.3.3 The Choice of Decision-Making Criteria 257
9.3.4 Suitability of the Primary Endpoints and Practical Aspects 258
9.3.5 Consideration for Key Secondary Objectives 258
9.3.6 Data Management and Information Sharing Platform 258
9.3.7 Appropriate Regulatory Engagement 259
9.3.8 Engaging Clinical Investigators 260
9.3.9 Addressing Funding and Support Accessible to UK CTUs 260
9.3.10 Pertaining to Public Funders 260
9.3.11 Bridging the Practical Knowledge Gap 261
9.3.12 Adaptive Trials Monitoring Capacity 263
9.3.13 Transparency and Reporting Framework of Adaptive Trials 263
9.3.14 Addressing Credibility of Findings from Adaptive Trials 264
9.3.15 Promising Adaptive Designs in the Public Sector 265
9.4 Main Thesis Strengths and Dissemination Achievements 266
9.5 Key Limitations and Interpretation of Findings 267
9.6 Areas of Future Related Research Beyond This Thesis 269
9.7 Overall Conclusions 270

Chapter 10. References 272

List of Figures

Figure 2.1. Number of Google Scholar hits. 24
Figure 2.2. Broad classification of confirmatory adaptive designs. 26
Figure 2.3. Generalised two-sided group sequential superiority test with efficacy stopping boundaries. 56
Figure 2.4. Pocock and OBF stopping boundaries for a two-sided group sequential test. 60
Figure 2.5. Flowchart for an information based group sequential design. 80
Figure 2.6. Example of an operationally seamless phase 2/3 design. 83
Figure 2.7. Inferential seamless phase 2/3 design. 84
Figure 2.8. An example of a multi-arm multi-stage design. 85
Figure 3.1. Overlap of roles and responsibilities of 27 interviewees in clinical trials research. 96
Figure 4.1. The distribution of the nature of interventions investigated by UK CTUs respondents. 123
Figure 4.2. Ranked perceptions of UK CTUs respondents on important barriers. 125
Figure 4.3. Ranked perceptions of private sector organisations on important barriers. 127
Figure 4.4. Ranked perceptions of Public Funders respondents on important barriers. 129
Figure 4.5. Cross-sector perceptions on proposed key facilitators. 132
Figure 5.1. A flow diagram of the decision-making process used to determine the final search terms. 146
Figure 5.2. A flow diagram showing the review process including reasons for exclusion of trials. 150
Figure 5.3. Trends in the application of confirmatory adaptive designs. 157
Figure 5.4. The use of confirmatory adaptive designs in the public and private sector. 158
Figure 5.5. Geographical distribution of the application of confirmatory adaptive designs. 159
Figure 5.6. Flow diagram investigating the adequacy of ClinicalTrials.gov in capturing adaptive designs. 160
Figure 6.1. PRISMA flow diagram of the screening process. 171
Figure 6.2. Reporting compliance of universal CONSORT 2010 checklist items. 175
Figure 6.3. Trials meeting 'total' reporting compliance of universal CONSORT checklist items. 176
Figure 6.4. Trials meeting 'at least partial' reporting compliance of universal CONSORT checklist items. 178
Figure 6.5. Reporting compliance of group sequential specific checklist items. 180
Figure 7.1. Uncertainty around assumed successful hospital discharge for RATPAC trial. 201
Figure 7.2. Pattern of the re-estimated total sample size for RATPAC trial. 202
Figure 7.3. Uncertainty around assumed hospital admissions for 3Mg trial. 203
Figure 7.4. Pattern of the re-estimated total sample size for 3Mg trial. 204
Figure 7.5. Trends of conditional power and intervention effect for 3CPO trial. 206
Figure 7.6. Approximate type I error committed for conducting one futility analysis. 209
Figure 7.7. Trends of conditional power and intervention effect for RATPAC trial. 211
Figure 7.8. Trends of conditional power and intervention effect for Booster trial. 213
Figure 7.9. Stopping boundaries for RATPAC two-sided group sequential test. 215
Figure 7.10. Interim monitoring for a retrospective group sequential design for RATPAC trial. 219
Figure 7.11. Interim monitoring for a retrospective group sequential design for 3CPO trial. 223
Figure 8.1. Adaptive hierarchical testing strategy within a group sequential test for the NERVE BLOCK trial. 247

List of Tables

Table 2.1. Decision criteria for a generalised two-sided group sequential superiority test. 57
Table 3.1. Characteristics and demographics of interviewed participants. 97
Table 3.2. Inductive themes perceived to influence conservatism to confirmatory adaptive designs use. 105
Table 4.1. Distribution of the type of confirmatory adaptive designs stratified by sector of application. 135
Table 5.1. Characteristics of identified confirmatory adaptive designs stratified by Funder or Sponsor. 152
Table 5.2. Reasons for early stopping of adaptive designs. 154
Table 5.3. Type of identified adaptive designs stratified by the Funder or Sponsor. 155
Table 5.4. Some exemplars of registered confirmatory adaptive designs 156
Table 6.1. Characteristics of eligible reviewed group sequential randomised trials. 172
Table 6.2. Summary data of reporting compliance of group sequential specific aspects. 182
Table 6.3. Type of stopping boundaries utilised in those with complete information. 184
Table 6.4. Modified CONSORT checklist of information to include when reporting an adaptive randomised trial. 187
Table 7.1. Summary statistics of sample size re-estimation for RATPAC trial. 201
Table 7.2. Participants savings under various scenarios of conditional power futility threshold stopping criteria. 208
Table 7.3. Statistical properties of a retrospective group sequential design for RATPAC trial. 217
Table 7.4. Stopping boundary values for a retrospective group sequential design for RATPAC trial. 218
Table 7.5. Benefits of efficacy early stopping for a retrospective group sequential design for RATPAC trial. 219
Table 7.6. Design properties of a retrospective group sequential design for 3CPO trial. 221
Table 7.7. Stopping boundary values of a retrospective group sequential design for 3CPO trial 222
Table 7.8. Conditional simulation results for 3CPO trial after the 2nd interim analysis. 224
Table 7.9. Benefits of futility early stopping at the 2nd interim analysis for 3CPO trial. 224
Table 7.10. Conversion of design information for RATPAC information based group sequential designs. 226
Table 7.11. Stopping boundaries for RATPAC information based group sequential design. 227
Table 8.1. Statistical properties of six group sequential designs for the PENNYWISE trial at 90% power. 239
Table 8.2. Statistical properties of six group sequential designs for the PENNYWISE trial at 85% power. 240
Table 8.3. Study properties for WT ($\partial = 0.25$) for varying SC event rate scenarios. 242
Table 8.4. Study properties of a WT ($\partial = 0.25$) design for varying effectiveness of Prolonged Enoxaparin. 242

List of Abbreviations

AD	adaptive design
ADWG	Adaptive Design Working Group
ANCOVA	analysis of covariance
BSCPB	bilateral superficial cervical plexus block
CBER	Centre for Biologics Evaluation and Research
CHMP	Committee for Medicinal Products for Human Use
CI	confidence interval
CRC	Clinical Research Collaboration
CRO	Contract Research Organisation
CRUK	Cancer Research United Kingdom
CONSORT	CONsolidated Standards Of Reporting Trials
CP	conditional power
CPAP	continuous positive airway pressure
CTR	Clinical Trials Register
CTU	Clinical Trials Unit
CTRU	Clinical Trials Research Unit
DRF	Doctoral Research Fellowship
EM	expectation-maximization
EMA	European Medicines Agency
EME	Efficacy and Mechanism Evaluation
EQUATOR	Enhancing the QUAlity and Transparency Of health Research
EU	European Union
FDA	Food and Drug Administration
GCP	Good Clinical Practice
GSD	group sequential design
HP	Haybittle-Peto
HTA	Health Technology Assessment
JSM	Joint Statistics Meetings
ICH	International Conference on Harmonisation
ICMJE	International Committee of Medical Journal Editors
ICTMC	International Clinical Trials Methodology Conference
ICTRP	International Clinical Trials Registry Platform
IDMC	Independent Data Monitoring Committee
IQR	interquartile range
IV	intravenous magnesium sulphate
LCL	lower confidence limit
LD	Lan-DeMets
LR	Likelihood ratio
LWI	local wound infiltration
MAMS	multi-arm multi-stage
Max	Maximum
MeSH	Medical Subject Headings
MHRA	Medicines and Healthcare products Regulatory Agency
MLE	maximum likelihood estimate
MI	myocardial infarction
Min	Minimum
MRC	Medical Research Council
MUE	median unbiased estimate
NEB	Nebulised magnesium sulphate
NHS	National Health Service
NHTMR	Network of Hubs for Trials Methodology Research
NB	nerve block
NCI	National Cancer Institute
NHLBT	National Heart, Lung, and Blood Institute
NIH	National Institutes of Health
NIHR	National Institute for Health Research

NIM	non-inferiority margin
NIPPV	non-invasive intermittent positive-pressure ventilation
OBF	O'Brien and Fleming
OR	odds ratio
PE	prolonged enoxaparin
PhRMA	Pharmaceutical Research and Manufacturers of America
PPCI	primary percutaneous coronary intervention
PSI	Statistics for the Pharmaceutical Industry
PoC	Point-of-care
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSI	Statistics in the Pharmaceutical Industry
PT	Pampallona and Tsiatis
REC	Research Ethics Committee
RCI	repeated confidence interval
RCT	randomised controlled trial
R & D	Research and Development
RfPB	Research for Patient Benefit
RR	risk ratio
SAACTD	Scientific Advances in Adaptive Clinical Trials Designs Workshop
SC	Standard Care
ScHARR	School of Health and Related Research
SCT	Society for Clinical Trials
SD	standard deviation
SSR	sample size re-estimation
SOT	standard oxygen therapy
STEMI	ST elevation myocardial infarction
TEE	total energy expenditure
TSC	Trial Steering Committee
UCL	upper confidence limit
UK	United Kingdom
UMVUE	uniformly minimum variance unbiased estimator
USA	United States of America
VAS	visual analogue scale
WHO	World Health Organisation
WT	Wang and Tsiatis

Chapter 1. Introduction

1.1 Background

Carefully planned and conducted confirmatory randomised controlled trials (RCTs) are regarded as the gold standard to provide reliable evidence on the effectiveness of investigative interventions (Akobeng, 2005; Pearce et al., 2015). Hence, RCTs are often the cornerstone of the decision-making process to approve and commission investigative interventions for adoption in clinical practice (Tudur Smith, Williamson, et al., 2014).

Traditionally, 'standard' RCTs are designed with a fixed sample size with the aim to recruit until this sample size target is met (Chen et al., 2012). The statistical analysis is then performed based on the outcome data of all the patients at the end of the trial (Kairalla et al., 2012). This is often referred to as a 'fixed sample size design'. The approach makes the assumption that the statistical and operational aspects made at the design stage are accurate and remain fixed throughout the trial (Law and Wason, 2014; Wang, 2010). However, in practice, obtaining the required information to design a trial is challenging, and design aspects are often estimated with varying degrees of uncertainty – meaning that the design assumptions may be inaccurate (Bauer and Kohne, 1994; Charles et al., 2009; Teare et al., 2014).

A great deal of attention has been paid to an alternative class of RCTs, known as adaptive designs (ADs) (Chow, 2014; Lanini et al., 2015). ADs permit modifications to some aspects of the design based on accruing outcome data from an ongoing trial, while preserving the scientific validity and integrity of the trial (Bretz et al., 2009; Kairalla et al., 2012). Carefully planned and executed ADs, when appropriate, may offer many potential advantages over conventional fixed sample size trials (Chow, 2014). For example, by using accruing outcome data, ADs have the potential to effectively assess promising interventions; thereby saving time, trial participants, and limited resources (Chow and Corey, 2011). In addition, ADs may improve the chance that a trial can efficiently answer the research question(s), by mitigating the risks of making inaccurate design assumptions (Bauer and Kohne, 1994; Chow and Corey, 2011).

As highlighted later in Chapter 3, there is a desire within the research community to explore and make use of efficient trial designs, mainly due to poor treatment 'success' rates, and the need to maximise value for money in research. Despite the potential benefits of ADs, their use in practice is lagging behind the methodological prominence they have in the statistical literature (Bauer and Einfalt, 2006; Morgan et al., 2014; Quinlan et al., 2010). There have been a number of initiatives, predominantly from a pharmaceutical industry perspective, to foster discussions, and address some of the associated challenges to improve the adoption of ADs (Burnham et

al., 2014; Chow and Corey, 2011; Morgan et al., 2014; Quinlan and Krams, 2006; Quinlan et al., 2010). In addition, the European Medicines Agency (EMA) and Food and Drug Administration (FDA) have drafted regulatory guidance documents to enhance the appropriate application of ADs by Clinical Trialists (CHMP, 2007; FDA, 2010, 2013, 2015).

There are suggestions that while the use of ADs in the pharmaceutical sector is gaining momentum, their use in the public sector is trailing behind (Morgan et al., 2014). It has been acknowledged that the public sector has some specific challenges which require addressing to improve the uptake of ADs (Kairalla et al., 2012). Recent research has been undertaken focusing on the use of ADs in early phase trials, specifically in the United Kingdom (UK) public sector (Jaki, 2013). The National Institutes of Health (NIH) in the United Stated of America (USA) also launched a cross-sector initiative, at the commencement of this research, to foster discussions and facilitate ways to address obstacles to the use of ADs in the USA (Coffey et al., 2012; Kairalla et al., 2012).

Despite the promising benefits, ADs are not often applied in publicly funded trials in the UK. This thesis therefore aims to investigate why ADs are underused in UK publicly funded confirmatory RCTs, and to make some practical recommendations for their appropriate use.

1.2 Brief Historical Perspective on Trial Adaptation

" … *there can be no objection to the use of data, however meagre, as a guide to action required before more can be collected* …" (Thompson, 1933)

The concept of using accruing outcome data in the decision-making process of an ongoing clinical trial can be traced back to the early 1930s when a method to modify the randomisation in favour of a promising intervention was proposed (Thompson, 1933). Such an approach was recommended for ethical reasons to minimise the number of participants exposed to supposedly inferior interventions and in cases where recruitment of participants is slow. Stein (1945) proposes a one sample two stage internal pilot procedure aimed to reduce uncertainty in the estimation of the sample standard deviation (SD). Armitage (2014) discusses the evolution of adaptive trials motivated by the desire to stop trials earlier than planned, as soon as there is sufficient evidence to address the research question(s). Early examples of discussed trials can be traced back to the 1950s in the UK (Newton and Tanner, 1956; Robertson and Armitage, 1959; Snell and Armitage, 1957; Watkinson, 1958). Since then, there have been numerous developments on ADs-related statistical methods (Bauer et al., 2015; Todd, 2007).

1.3 Motivation

Of late, the 'success' rate of investigative interventions in confirmatory RCTs has been unsatisfactory and disappointing – both in the public and private sector (Dent and Raftery, 2011; Jaki, 2013; Kola and Landis, 2004). Dent and Raftery (2011) found that, just 24% of pragmatic, superiority comparisons of confirmatory RCTs funded by the UK NIHR Health Technology Assessment (HTA) programme yield clinically important and statistically significant results in either direction – with 19% in favour of the investigative interventions. On comparison, these results were consistent with those from National Cancer Institute (NCI) trials. Furthermore, 54% of intervention comparisons were inconclusive in regarding to answering the research question(s). Reasons such as inaccurate assumptions about variability or control event rate, poor recruitment, and overoptimistic planned effect sizes, among others were cited to explain inconclusive results. It is important to note, as discussed in Section 1.4, that a successful trial is not necessarily a statistically significant trial.

In 2000, the average pharmaceutical 'success' rate of investigative interventions from first-in-human to registration was estimated at around 11%, but it varies considerably across therapeutic areas (Djulbegovic et al., 2013; Kola and Landis, 2004). The average clinical approval success rate of investigative interventions based on the review period 1993 to 2009 was 19% (DiMasi et al., 2010).

Reviews of RCTs have found considerable cross-sector discrepancies between assumptions made about design parameters and observed estimates from outcome data (Charles et al., 2009; Clark et al., 2013; Vickers, 2003). Moreover, the recruitment of trial participants in UK publicly funded trials is not improving, with just over 50% of trials failing to recruit as planned (McDonald et al., 2006; Sully et al., 2013). Whilst in the pharmaceutical industry the costs of conducting trials are escalating at a time when available research resources are shrinking (Collier, 2009).

In summary, the issues highlighted raise fundamental ethical and efficiency questions, provoking a rethink of the way RCTs are designed and conducted. These questions include:

- Are participants unnecessarily exposed to inferior or futile interventions in most trials?
- Are participants, the pool of investigators, and available resources efficiently utilised in clinical trials research?
- How can RCTs be designed to answer research question(s) efficiently?
- Is value for money maximised in clinical trials research in the current era of resource constraints?

1.4 The Research Question and Rationale

As highlighted in Section 1.1, well conducted ADs have the potential to improve efficiency in the design and conduct of RCTs, and alleviate some limitations of conventional fixed sample size designs. However, ADs are not appropriate in all research settings, and there are challenges preventing their application (Chow and Corey, 2011). Building on Section 1.1, there is a need for research specific to the use of ADs in the UK publicly funded confirmatory setting.

There are perceived differences between the public and private settings, which may influence receptiveness towards, and barriers to the use of ADs (Kairalla et al., 2012) . For instance, the nature of study interventions is more diverse in the public sector, as highlighted in Section 4.4.2 of Chapter 4. For example, some publicly funded trials may assess interventions that are licensed and used in clinical practice for conditions where there is no evidence of benefit. This is the case for the 3Mg trial to be described in Section 7.3.3 of Chapter 7. In such cases, researchers may regard a 'negative' trial positively as it would lead to the withdrawal of an ineffective intervention from the care pathway. In contrast, trials in the private sector are more likely to be for unlicensed interventions.

This leads to the research question for this thesis, which is to investigate the lack of routine utilisation of ADs in the UK publicly funded confirmatory setting, to identify barriers to their wide and appropriate use, and to explore facilitators to address some of the perceived barriers.

1.5 Overarching Specific Objectives

The specific thesis objectives to address the research question described in Section 1.4 are to:

1) Review and describe ADs with potential to be applied in the publicly funded confirmatory setting, from a statistical and practical perspective;

2) Investigate barriers and facilitators to the appropriate use of ADs based on key stakeholders' experiences, perceptions, and attitudes, through in-depth interviews and subsequent quantitative surveys;

3) Demonstrate statistical implementation and potential opportunities, and highlight pitfalls during the application of some forms of confirmatory ADs;

4) Learn from prospectively and retrospectively planned case studies of ADs to enhance the future planning of trials with similar characteristics;

5) Draw practical recommendations for best practice tailored for Clinical Trialists and Public Funders, to facilitate the appropriate use of ADs.

1.6 Scope of the Research

The thesis shall focus on ADs for confirmatory trials with emphasis on the UK publicly funded setting. Consideration is given to parallel group superiority RCTs reflecting dominant characteristics of confirmatory trials funded by major UK Public Funders, such as the NIHR HTA (Dent and Raftery, 2011). More so, focus is on prospectively planned ADs due to emerging consensus among Clinical Trialists and policymakers that this is a necessary condition for good practice (FDA, 2015; Kairalla et al., 2012). In addition, only ADs where the nature of the adaptation(s) are solely informed based on accruing primary outcome(s) data from that ongoing trial are considered. Thus, trial adaptations based solely on external information to an ongoing trial and/or operational aspects, such as feasibility criteria are out of scope of this thesis. Finally, ADs based on the frequentist paradigm are considered reflecting the current mainstream approach in the design and conduct of confirmatory RCTs.

1.7 Thesis Roadmap

In Chapter 2, types of confirmatory ADs are reviewed, and described from a statistical and practical perspective, starting from simple to more complex forms. Available statistical software or code for implementation are highlighted. In Chapter 3, methods adopted to explore themes on barriers and potential facilitators to ADs use are described and findings presented. This chapter hinges on the experiences, perceptions, and attitudes of key stakeholders explored through in-depth interviews.

Building on generated findings from Chapter 3, Chapter 4 describes methods employed to further explore barriers to the use of ADs through cross-sector quantitative surveys, with the aim being to generalise the findings, and rank barriers for prioritisation. The findings from quantitative surveys are also presented, including main potential facilitators to mitigate some of the uncovered barriers and concerns. The results from Chapters 3 and 4 lay the foundation for the remainder of this thesis. In Chapter 5, case studies of registered ADs are reviewed and presented aimed at finding solutions to overcome some barriers and concerns raised in Chapters 3 and 4. Chapter 6 investigates the state of reporting of the most commonly used confirmatory AD and draws some recommendations.

Chapter 7 demonstrates the design, and statistical implementation of specific ADs using retrospective case studies. In addition, potential opportunities, pitfalls, and lessons learned from using these ADs are highlighted

to help the future planning of similar adaptive trials. Building on lessons learned from Chapter 7, Chapter 8 illustrates the design and planning of ADs using the two case studies of actual grant applications submitted to Public Funders for consideration. In conclusion, Chapter 9 completes the thesis with a discussion, recommendations for best practice tailored for Clinical Trialists and Public Funders, direction of future work, and overall conclusions.

1.8 Summary

Conventional fixed sample size designs are simple, well accepted, and provide unbiased effects of investigative interventions, when properly conducted. Despite this, they suffer from ethical, efficiency and value for money limitations in trials research, due to inability to use accruing outcome data in the decision-making process. Properly planned and executed ADs have the potential to mitigate some of these limitations. However, despite their promising advantages, ADs are not widely undertaken – especially in the UK publicly funded confirmatory setting. A comprehensive investigation of the obstacles to the application of ADs is required and to highlight when they are appropriate and their limitations, as well as to how they can be implemented statistically. This thesis seeks to address some of these issues and make recommendations for best practice to facilitate appropriate wider use of ADs.

Chapter 2. Literature Review

2.1 Introduction

Chapter 1 introduced the traditional approach underpinning the design and conduct of fixed sample size RCTs and its limitations. The concept of ADs has been highlighted with the aim of mitigating some of the limitations and the thesis aims to investigate why ADs are underutilised in practice. This chapter provides a foundation of the types and scope of ADs in the literature with application potential in confirmatory RCTs. The approach adopted in reviewing the literature shall set the scene. The description of reviewed ADs starts with simple and moves to more complex ADs. The next chapter then provides a backbone by exploring barriers, concerns and potential facilitators to the appropriate use of ADs based on interviews of key stakeholders in clinical trials research.

2.2 Aims

In this chapter, the aim is to review the types and scope of ADs from a statistical and practical perspective, guided by the following key aspects:

1. Motivation behind the AD, and its potential benefits compared to traditional fixed sample size designs;
2. Description of statistical methods behind the AD, impact on type I and II errors, and impact on statistical inference and available methods to obtain unbiased results. ADs with binary or continuous outcomes are considered for simplicity although the principles guiding the application of ADs are outcome independent. A reflection on more complex ADs is provided;
3. Practical considerations during implementation of the AD, with emphasis on the publicly funded setting. Available implementation resources such as statistical software or code are highlighted;
4. Limitations of the AD and pitfalls.

2.3 Literature Search

The search for relevant literature through a systematic and exhaustive approach is often ideally preferred. During a preliminary scoping exercise using just two relevant terms “adaptive design” or “group sequential”, a cumulative total of 20,009 articles were identified between 1980 and 2012. Figure 2.1 shows the corresponding

trends of Google Scholar hits. As evident, the scope of the thesis objectives renders a systematic literature review impractical within the time constraints of the PhD.

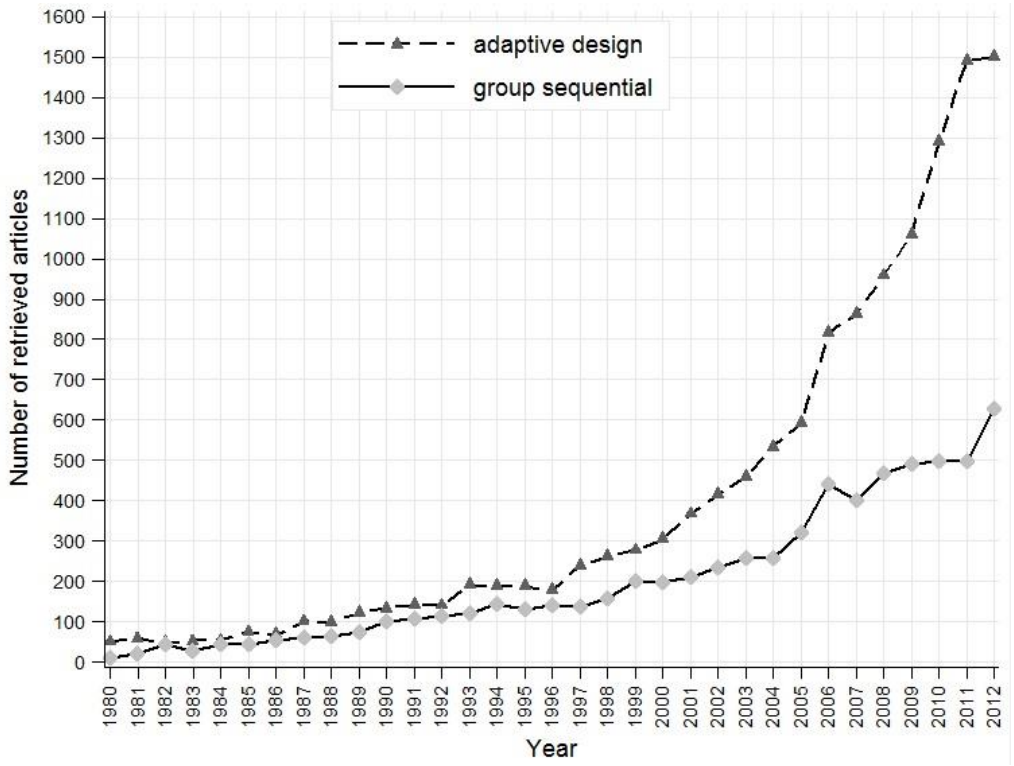

Figure 2.1. Number of Google Scholar hits.

Due to the impracticality of undertaking an exhaustive systematic review, the literature search employed a pearl growing approach (Schlosser et al., 2006) and restricted systematic review. Initial literature search was based on primers of the most relevant peer-reviewed journal publications and textbooks on confirmatory ADs, as recommended by experts on ADs (Bretz et al., 2009; Chow and Chang, 2008, 2011; Jennison and Turnbull, 2000a; Kairalla et al., 2012). Citation tracking of the most relevant publications was also undertaken, guided by key aspects described in Section 2.2. Searching Ovid MEDLINE and PubMed using the following restrictive search algorithm in order to minimise exclusion of key literature complemented the primers for further citation tracking.

"adaptive design" AND ((Clinical Trial[ptyp] OR Clinical Trial, Phase III[ptyp] OR Guideline[ptyp] OR Journal Article[ptyp] OR Review[ptyp] OR systematic[sb] OR Randomized Controlled Trial[ptyp] OR Controlled Clinical Trial[ptyp] OR Pragmatic Clinical Trial[ptyp] OR Practice Guideline[ptyp] OR Scientific Integrity Review[ptyp]) AND (full text[sb] AND hasabstract[text]) AND Humans[Mesh] AND English[lang])

A Cochrane Library search used the term "adaptive design". Relevant publications of interest included: statistical methodology, and related reviews; practical or regulatory documents; scientific integrity reviews; and

commentary or discussions. Of the 484 retrieved records, 151 were relevant publications included for citation tracking. Literature alerts during the course of the thesis were set-up in databases such as PubMed using a restrictive tailored search algorithm. In addition, AD-related "Journal Issues" were tracked in statistical journals such as Journal of Biopharmaceutical Statistics, Pharmaceutical Statistics and Statistics in Medicine, and Sage Journals using the term "adaptive design" in all fields as citation alerts. Literature management was done using Mendeley reference manager.

2.4 Characterisation of Reviewed Confirmatory Adaptive Designs

A number of primer publications review the types of ADs for various trial phases, which are applicable across therapeutic areas as shall be reflected in Section 3.4.3.2 of Chapter 3 (Bauer et al., 2015; Bowalekar, 2011; Bretz et al., 2009; Chow and Chang, 2008; Dragalin, 2006; Jennison and Turnbull, 2000a; Kairalla et al., 2012; Koch, 2006; Lai et al., 2012, 2015; Law and Wason, 2014; Maca et al., 2014; Todd, 2007; Zang and Lee, 2014). Figure 2.2 displays the broad classification of the confirmatory types of ADs in some order of complexity.

There is wide scope of the types of ADs described in the literature. As supported by the results of Chapters 3 to 6, focus in the description of the types of ADs is paid to those that are most commonly used, easier to implement, and attract more attention and receptiveness by decision-makers, researchers, and Public Funders. These include blinded sample size re-estimation (SSR), one stochastic curtailment futility analysis, group sequential design (GSD), and operational seamless ADs. The concept behind the multi-arm multi-stage (MAMS) and inferential seamless designs is highlighted together with the underlying statistical principles guiding the more complex ADs.

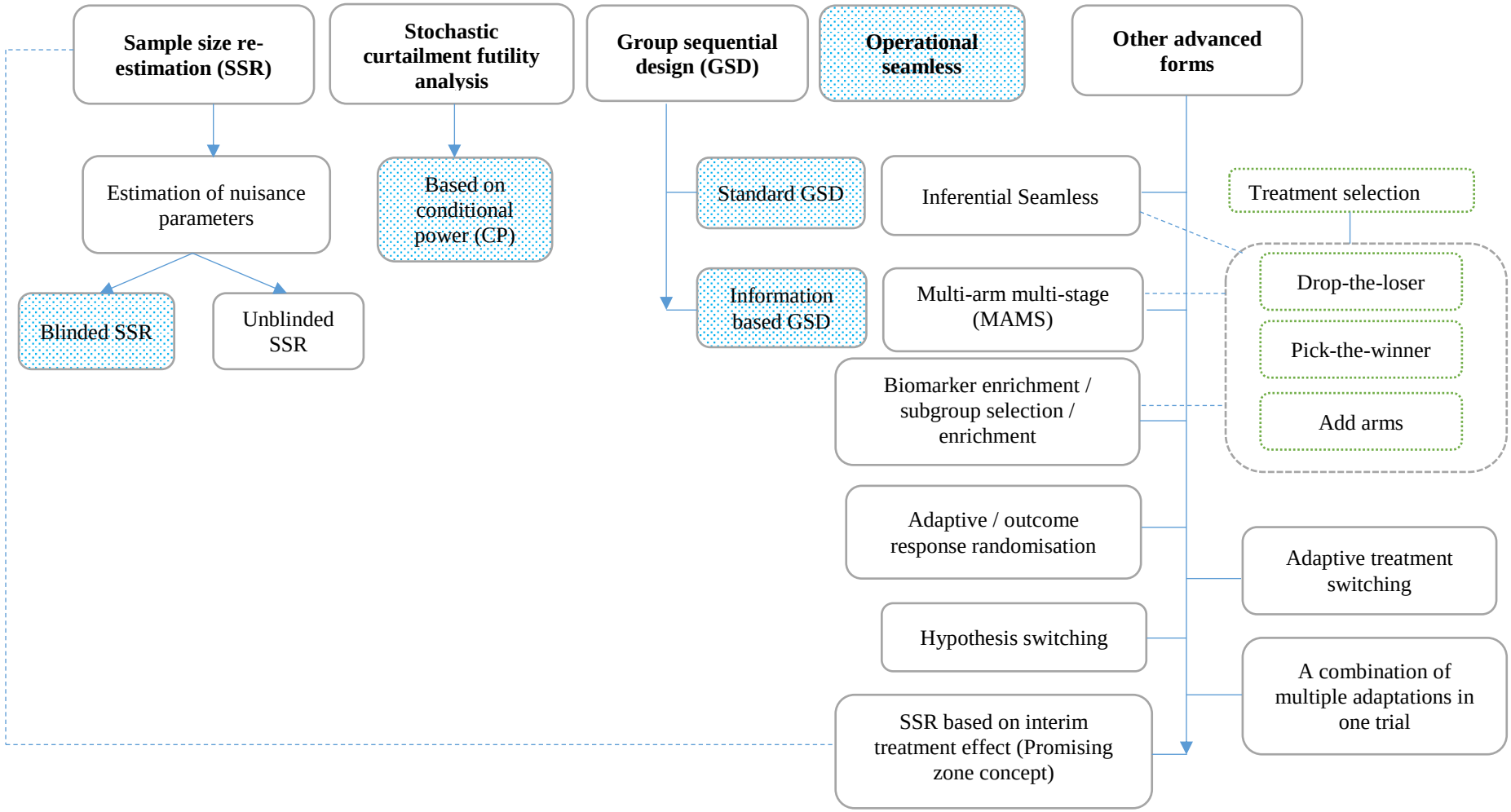

Figure 2.2. Broad classification of confirmatory adaptive designs.

2.5 Design 1: Sample Size Re-estimation

This section deals with SSR methods informed solely by the estimates of nuisance parameters calculated in a blinded or unblinded manner.

2.5.1 Introduction to Sample Size Estimation

Estimating the sample size needed is essential during the planning of a trial to ensure that there are enough participants to address the research question(s) with desired statistical properties (ICH, 1998). This is achieved based on specific primary objective(s) and outcome(s), hypothesis test(s) of interest, trial design, planned statistical analysis, and statistical characteristics of the design (type I and II errors) (Chow et al., 2003; Donner, 1984; Eng, 2003; Julious, 2010). For a balanced, parallel group RCT, the sample size (n_0 per group) required to test a two-sided superiority hypothesis test of difference between two independent means for a continuous Normally distributed outcome can be estimated from

$$n_0 = \frac{2\sigma^2\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2}{{\theta_\delta}^2}, \qquad 2:1$$

where σ^2 is the pooled or within-group variance; $Z_{\frac{\alpha}{2}}$ and Z_β are the Normal distribution percentiles corresponding to type I (α) and II (β) errors, respectively; and θ_δ is the clinically relevant effect size. Similarly, the sample size (per group) for a binary outcome for testing the difference between two proportions based on a Chi-Square test is computed using one of equations (2:2), (2:3) or (2:4) (Chow et al., 2003; Fleiss et al., 2003; Lachin, 1981), given as

$$n_0 = \frac{\left(Z_{\frac{\alpha}{2}}\sqrt{2\bar{p}(1-\bar{p})} + Z_\beta\sqrt{p_C(1-p_C)+p_T(1-p_T)}\right)^2}{{\theta_\delta}^2}, \qquad 2:2$$

$$n_0 = \frac{\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2\left(p_C(1-p_C)+p_T(1-p_T)\right)}{{\theta_\delta}^2}, \qquad 2:3$$

$$n_0 = 2\bar{p}(1-\bar{p})\left(\frac{Z_{\frac{\alpha}{2}} + Z_\beta}{\theta_\delta}\right)^2, \qquad 2:4$$

where p_C and p_T are the control and intervention event rates, respectively; $\bar{p} = \frac{p_C+p_T}{2}$ is the overall event rate; and $\theta_\delta = p_T - p_C$. Equation (2:4) approximates (2:3) and is used for blinded SSR for binary outcomes described in Section 2.5.12.

To calculate the required sample size, essential nuisance parameters (σ^2, p_C or $\bar{p}$) are often estimated based on previous data of similar patient population such as from cohort studies, RCTs or pilot trials (Altman, 1991; Arnold et al., 2009; Bauer and Kohne, 1994; Lancaster et al., 2004). Dantzig (1940) proves the non-existence of statistically useful regions for a single sample t-test such that its power function is independent of σ. Stein (1945) generalises these findings to a two sample t-test and linear hypotheses. The estimation of these nuisance parameters is often challenging, associated with varying degrees of uncertainty, and thus influences the power of a test (Chiang-Stein et al., 2006; Dantzig, 1940; Pritchett et al., 2015; Stein, 1945; Teare et al., 2014). The cited authors underline that the reasons may be due to, but not limited to:

- Estimates based on relatively small studies or pilot trials;
- The shift in the distribution of health outcomes because of healthcare improvements over time;
- Differences in patient populations influenced by aspects such as inclusion and exclusion criteria, diagnosis criteria, and monitoring procedures;
- Estimates based on studies with questionable generalisability, such as from single centre studies when the main trial is intended as a multicentre RCT.

The aim of SSR is therefore, to use interim primary outcome data from an ongoing trial to validate assumptions on nuisance design parameters and take action to revise the sample size of that trial, and not to conduct interim hypothesis testing (Herson and Wittes, 1993).

2.5.2 Addressing Uncertainty Around Nuisance Parameters

Major UK funding bodies such as the NIHR and Medical Research Council (MRC) have been advocating for the conduct of pilot and feasibility studies to inform robust design of confirmatory RCTs, to enhance their quality and success rates (Craig et al., 2013; NIHR, n.d.). The definitions, objectives, and design aspects of pilot and feasibility studies have been well discussed (Arnold et al., 2009; Charlesworth et al., 2013; Lancaster, 2015; Lancaster et al., 2004; Shanyinde et al., 2011; Thabane et al., 2010; Whitehead et al., 2014).

The need to estimate design parameters for sample size estimation is often stated as a justification for the conduct of a pilot trial (Cocks and Torgerson, 2013). An external pilot trial is an approach employed to help improve the accuracy of parameter estimates assumed for sample size estimation (Lancaster et al., 2004). However, as Bauer and Kohne (1994) highlight, the use of external pilots to estimate sample size parameters alone wastes resources, time and trial participants, since data generated cannot be used in the final confirmatory analysis.

In addition, external pilot trials based on relatively small sample sizes are inefficient in reducing uncertainty around design parameters (Sim and Lewis, 2012; Teare et al., 2014). Inflation methods have been proposed to improve efficiency of small external pilot trials (Browne, 1995; Julious and Owen, 2006; Kieser and Wassmer, 1996). However, these methods can yield excessively larger confirmatory RCTs than necessary (Teare et al., 2014).

An internal pilot trial has been proposed as an alternative to an external pilot trial, where the data generated are used in the final confirmatory analysis (Bauer and Kohne, 1994; Wittes and Brittain, 1990). SSR uses the internal pilot concept to mitigate the risk of making inaccurate estimates regarding the actual size of the trial required to address research question(s) (Bauer and Kohne, 1994).

2.5.3 Motivation for Sample Size Re-estimation

Reviews of confirmatory RCTs found marked discrepancies between nuisance parameters used at the design stage, and those observed after trial completion (Charles et al., 2009; Clark et al., 2013; Vickers, 2003). In a survey of 215 published journal reports, just 73(34%) RCTs reported all information required to estimate the sample size, had accurate calculations, and used accurate assumptions about the control event rate (differed by less than 30% of observed data) (Charles et al., 2009). In another review of 28 RCTs, the observed variability was greater than that assumed in 24(80%) of the endpoints published in perceived reputable journals (Vickers, 2003). In contrast, trials tended to be overpowered rather the underpowered based on another review of protocols submitted to the UK research ethics committees (Clark et al., 2013). All these studies highlight that sample size estimates are often based on inaccurate assumptions, regardless of the direction of the bias, as illustrated in Section 7.5.1 of Chapter 7.

The challenges Trialists face to obtain accurate estimates of nuisance parameters to guide sample size estimation are well acknowledged (Bauer and Kohne, 1994; Schulz and Grimes, 2005; Teare et al., 2014; Wittes and Brittain, 1990). Inaccurate design information can have detrimental consequences on the statistical power, as well as trial resources, duration, feasibility, and ethics (Friede and Miller, 2012). For example, underpowered trials may yield inconclusive results unable to answer the primary research objective(s) (Sim and Lewis, 2012) – although results can be used for evidence synthesis (Chalmers et al., 1987). In contrast, although 'overpowered' trials are desired from statistical perspectives, they raise ethical, economic, and feasibility questions (Donner and Makuch, 1985; Sim and Lewis, 2012). It is therefore important to recruit enough participants to address the

research question(s), while minimising unnecessary over recruitment. This is paramount given that, over 50% of publicly funded confirmatory trials struggle to meet target recruitment in time (Farrell et al., 2010; McDonald et al., 2006; Sully et al., 2013).

2.5.4 Early Research on Sample Size Re-estimation

To alleviate uncertainty around the estimate of the population variance, Stein (1945) proposes a two stage design to estimate a population mean within a specified CI (Confidence Interval) limit and significance level for a single sample. The decision on the need for further recruitment is based on the variance estimate at some chosen interim point, referred to as 'stage 1'. Considerable attention has been given to this approach, its variants and extensions studied, and applied in various settings (Anscombe, 1953; Cox, 1952; Hall, 1981; Lohr, 1990; Seelbinder, 1953).

Although Stein's design marked a milestone, it had limited application and is wasteful because it only uses stage 1 data in estimating the variance of the test statistic even if further stage 2 recruitment is required (Denne and Jennison, 1999; Proschan and Wittes, 2000; Wittes and Brittain, 1990; Zucker et al., 1999). In addition, Proschan et al (2006) highlight that, the variance estimate may be unappealing and inefficient, when based on a small 'stage 1' sample. Hence, modified versions have been proposed to improve efficiency, and applicability of Stein's design in clinical trials (Denne and Jennison, 1999; Proschan and Wittes, 2000; Wittes and Brittain, 1990; Zucker et al., 1999). These designs are briefly described in Sections 2.5.5 and 2.5.6.

2.5.5 Restricted Internal Pilot Concept

Wittes and Brittain (1990) describe the concept of internal pilots in clinical trials, which is a modified version of Stein's two stage design. The authors argue in favour of using internal pilots to estimate parameters related to trial administration and process of the disease in order to inform robust re-design of confirmatory RCTs. The implementation of the design is summarised as follows:

1) Estimate the sample size (n_0 per group) required using an appropriate formula such as that described in Section 2.5.1;
2) Choose a proportion π such that the internal pilot sample size (per group) is given by $n_1 = \pi n_0$;
3) Recruit (n_1 per group) participants in the internal pilot and estimate nuisance parameters of interest;

4) Re-estimate the sample size (n^* per group) using an appropriate formula such as that described in Section 2.5.1, but based on re-estimated nuisance parameters;
5) Make an appropriate decision for further recruitment, such that the final sample size per group is given by $maximum(n_0, n^*)$. Thus, additional recruitment (per group) is based on $n_2 = maximum(n_0, n^*) - n_0$ participants.

This design operationally works against revising the sample size downwards, hence, it is referred to as a 'restricted' SSR design (Zucker et al., 1999). As acknowledged by other authors, there are situations when the revised sample size turns out to be much smaller than planned, suggesting the trial could be stopped earlier than planned (Gould and Shih, 1992; Gould, 1992). In such cases, failure to reduce the sample size could result in larger trials than necessary to address the research question(s) (Birkett and Day, 1994). However, Gould and Shih (1992) highlight that careful consideration should be given to the impact of downward revision of the sample size on other important secondary trial objectives. Section 7.5.1 of Chapter 7 illustrates this issue.

2.5.6 Unrestricted Internal Pilot Concept

Citing the limitation of a 'restricted' SSR design, Birkett and Day (1994) propose an approach referred to as an 'unrestricted' design to allow for downward sample size revision when necessary. Building on Section 2.5.5, item 5, the revised sample size bounded by a lower limit n_1 is such that $n_2 = maximum(n_1, n^*) - n_1$. The 'restricted' and 'unrestricted' SSR designs assume a single analysis after complete recruitment of all participants (Day, 2000; Zucker et al., 1999). Moreover, the estimation of the intervention effect is based on pooled data from the two stages, without any statistical correction to the test statistic or significance level, often referred to as a *naïve* approach (Day, 2000; Jennison and Turnbull, 2000b). The statistical properties of the 'restricted' and 'unrestricted' SSR designs are described in subsequent sections. Appendix 2.1 describes other extensions to Stein's two stage internal pilot concept.

2.5.7 When to Conduct Sample Size Re-estimation

The selection of the internal pilot sample size is an important consideration at the design stage of RCTs. If the SSR is performed too early based on relatively few participants, it may produce inaccurate estimates undermining the efficiency of the procedure (FDA, 2015).

2.5.7.1 Continuous Outcomes

Another limitation of Stein's two stage design is that there is no rule guiding the choice of the internal pilot sample size, so its selection is left to the discretion of the Trialist (Moshman, 1958). Bechhofer et al (1954) observed an inverse relationship between the magnitude of the sample variance and internal pilot sample size. Seelbinder (1953) describes a minimax rule such that the internal pilot sample size of Stein's two stage design is chosen to minimise the maximum expected excess in sample size, over a range of the SD and maximum tolerated discrepancy given as a ratio. Moshman (1958) modifies this idea to minimise the total sample size in conjunction with the use of an upper CI limit (such as 95%) of the total sample size distribution, for a given ratio of the critical value to the SD. However, the idea is applicable in settings where trial sample sizes are very constrained – hence, it may be of limited practical application in most confirmatory trials.

Like Stein (1945), Wittes and Brittain (1990) fail to provide an 'optimum' rule guiding the choice of the internal pilot sample size depending on fraction π to estimate nuisance parameters with a reasonable degree of precision. Their choice of 50% of the initial total sample size was arbitrary which may appear suboptimal by ignoring the precision of estimates of nuisance parameters (Sandvik et al., 1996). In contrast, Birkett and Day (1994) found that the absolute sample size of the internal pilot was more important than the choice of π to provide accurate estimates of the variance. For example, they argued that for a small planned sample size, the corresponding internal pilot sample size would be relatively small regardless of how large π is. However, planned sample sizes are relatively large in confirmatory trials – hence this argument may not suffice. A number of authors argue that, in addition to the precise estimation of parameters, the choice of π is also important in planning practicalities (Gould, 1992, 1995; Wittes et al., 1999).

Birkett and Day (1994) suggest a minimum of 20 to 40 degrees of freedom for an internal pilot with a continuous outcome based on minimum change in expected sample size. However, the imprecision of the variance estimate could still be large due to the skewed nature its distribution. For example, with 20 degrees of freedom, the internal pilot variance estimate has a 37% chance of falling outside the range 0.75 to 1.33 of the true variance (Denne and Jennison, 1999; Singer, 1999). Even based on an internal pilot total sample size of 50, the chance of the variance estimate differing from the true parameter by at least 20% is as large as 32% (Sandvik et al., 1996).

Sandvik et al (1996) describe a method based on calculating a constant of proportionality derived by choosing the probability of recommending unnecessary additional participants. Although the idea seems prudent, the constant of proportionality depends on the sample size used in the estimation of initial assumed variance, often

obtained from many studies. In addition, its choice is unclear, although authors recommend avoiding values close to 100%. Singer (1999) describes a modified method to account for recruitment and follow-up assuming a constant recruitment rate. This assumption is limiting since a deterministic constant recruitment rate is rarely tenable in practice.

Zucker et al (1999) found that the type I error inflation is negligible with a minimum internal pilot sample size of 40 (per group), and converges to the desired nominal level as the sample size increases. Denne and Jennison (1999) describe a rule of choosing the internal pilot sample size to minimise the ratio of the expected (based on observed variance) to the pre-planned sample size. However, the method has limitations because the true variance is unknown at the design stage.

Cocks and Torgerson (2013) suggest the use of internal or external pilots yielding a one-sided 80% CI that excludes the effect size for the main trial. They recommend a pilot size of at least 9% of the main trial for both binary and continuous outcomes. In addition, they also argue in favour of a minimum of 20 (per group) based on informal review of previous recommendations for continuous outcomes. Their "9% rule of thumb" appears inefficient, particularly for moderate sized RCTs (such as 164 to 200), hence contradicting other recommendations. Friede and Miller (2012) suggest that the bias of the test size is less of a concern when the internal pilot sample size is at least 50 per group for continuous outcomes.

In the context of external pilots, Browne (1995) studies the chances of the observed power exceeding the desired power and suggested a minimum of 30 participants for relatively large effects. Kieser and Wassmer (1996) argue for a rule based on minimising the size of both the external pilot and main trials. They recommend internal pilot sample sizes of 20 to 40 (per group) for main trials of sizes of 80 to 250 say, when an 80% one-sided upper CI limit is applied to the interim variance estimate. Whitehead et al (2015) apply this minimisation approach and provide recommendations on external pilot sample sizes depending on the effect size sought. For example, they suggest 75 per group for external pilots when a 10% standardised effect size is sought for the main trial. However, although minimising the main trial is appealing for SSR, minimising the size of the internal pilot seems illogical. Hertzog (2008) suggests a pilot size of between 10 and 40 per group as being sufficient to estimate nuisance parameters with a relatively 'high' degree of precision. Sim and Lewis (2012) recommend a pilot trial sample size of at least 55 participants in total (~20 to 30 per group) arguing that, in addition to gains in precision, it minimises the size of the pilot trial and the main trial. Teare et al (2014) suggest a minimum of 35 (per group) to yield less than 10% relative gain in the precision of the variance estimate.

2.5.7.2 Binary Outcomes

Herson and Wittes (1993) investigate the performance of the internal pilot SSR method through simulations and found that the choice of π (for the internal pilot size) between 25% and 75% yields desired statistical power. They state that a 50% threshold seems to be a reasonable choice. In the context of external pilot trials, Teare et al (2014) recommend a minimum of 60 to 100 (per group) when they investigated the percentage gain in precision around the control event rate between 10% and 50% through simulation. The authors argue that the marginal gain in precision beyond this is 'negligible'. Importantly, the type I error inflation was found to be more comparable to the fixed sample size design under a number of scenarios when the internal pilot sample size is above 100 in total (Friede and Kieser, 2004, 2006).

In summary, reviewed literature suggests that, it is advisable to conduct the SSR with as many participants as possible above the minimum internal pilot sample size thresholds proposed. Furthermore, logistical and administrative practicalities for smooth trial conduct should guide the upper bounds of the internal pilot sample sizes. Section 7.5.1 of Chapter 7 demonstrates the accuracy of internal pilot estimates with increases in sample size using case studies with binary outcomes.

2.5.8 The Frequency of Sample Size Re-estimation

The understanding of how frequent to undertake SSR is important at the design stage. Theoretical extensions to multiple and continuous monitoring SSR designs for continuous outcomes have been studied (Betensky and Tierney, 1997; Friede and Miller, 2012; Hall, 1981). For continuous monitoring SSR, nuisance parameters are estimated and sample size re-estimated after every recruited participant. Statistical properties of Stein's 'restricted' two stage (1945), three stage, and continuous monitoring SSR designs were studied with respect to expected estimated sample size, its mean square error and 95% CI (Betensky and Tierney, 1997). The continuous monitoring design was found to be more efficient compared to other designs. Similarly, Friede and Miller (2012) found superiority of the continuous monitoring SSR design in reducing the variability of the estimated sample sizes when compared to the repeated and two stage designs. However, although in theory the continuous monitoring SSR design is the most statistically efficient, it is impractical in most clinical trial settings due to logistical and operational challenges in planning recruitment activities (Betensky and Tierney, 1997; Friede and Miller, 2012). Hence, these authors generally conclude that a two stage design with a single interim is the

most pragmatic in confirmatory clinical trials, provided the sample size of the internal pilot is large enough to give reliable estimates.

2.5.9 Methods for Estimating Nuisance Design Parameters

The need to minimise operational bias in trial conduct is an important consideration to maintain credibility of findings to change clinical practice (FDA, 2010, 2013, 2015; ICH, 1998). Blinding is intended to minimise conscious and unconscious bias during trial conduct arising from the knowledge of which interventions trial participants are allocated to (Day and Altman, 2000). Such knowledge may influence indirect inference of the intervention effect and subsequently impact on aspects such as future recruitment of participants, management of trial participants, and assessment of outcomes. The relevance and degree of blinding varies from trial-to-trial according to design aspects and circumstances (Day, 2000; Schulz and Grimes, 2002). The levels of blinding, its implications on inferential bias and trial credibility have been well discussed (Day and Altman, 2000; FDA, 2010, 2015; ICH, 1998; Schulz and Grimes, 2002).

Some authors broadly classify SSR methods into two categories (Proschan, 2009; Proschan et al., 2006a). Those that are based on:

a) Estimation of nuisance design parameters either in a blinded or unblinded manner, such as pooled or within-group variance and control arm or overall event rate. These methods are described in Sections 2.5.10 and 2.5.11;
b) Estimate of the interim intervention effect, which is out of scope of this thesis.

2.5.10 Blinded Methods for Continuous Outcomes

Authors have reviewed methods for SSR using internal pilots for binary and continuous outcomes (Chiang-Stein et al., 2006; Friede and Kieser, 2006; Pritchett et al., 2015; Proschan, 2005). This section summarises these methods.

2.5.10.1 Estimation of Nuisance Parameters

Gould and Shih (1992) present a one sample variance (s_{L1}^2) formula to estimate the pooled variance computed by ignoring the fact that trial participants belong to different interventions. That is, s_{L1}^2 is computed without the knowledge or use of intervention allocation. This is defined as

$$s_{L1}^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{2}\sum_{j=1}^{n_{1i}}(Y_{ij} - \bar{Y}_1)^2\,, \qquad 2:5$$

where Y_{ij} is the interim outcome of the j^{th} participant in intervention i and $\bar{Y}_1$ is the overall interim mean using aggregated outcome data. Although s_{L1}^2 is simple to estimate and implemented by in-house personnel, it overestimates the true pooled variance (Proschan et al., 2006a). In addition, some authors illustrated that, blinded investigators can mathematically deduce the interim intervention effect from the knowledge of both the one sample and the pooled variances if they have access to individual participants outcome data (Julious, 2010; Proschan et al., 2006a; Proschan, 2005).

To obviate this problem, an improved estimate (s_{1GS}^2) that uses s_{L1}^2 and hypothesised intervention effect θ_δ under H_1 assuming a relatively large sample was proposed (Gould and Shih, 1992). The estimate is computed using

$$s_{1GS}^2 = \frac{N_1 - 1}{N_1 - 2}\left(s_{L1}^2 - \frac{\theta_\delta^2}{4}\right). \qquad 2:6$$

One limitation of s_{1GS}^2 is that it uses θ_δ, which is often specified with a degree of uncertainty in some settings (Chow and Chang, 2012; Proschan et al., 2006a). As a result, if θ_δ is overoptimistic, then s_{1GS}^2 will be too small compared to the true pooled variance. Hence, authors argue in favour of using s_{L1}^2 citing that overestimation of the pooled variance may provide safeguards in cases of small internal pilots. Some authors show that, the ratio of s_{L1}^2 to the true pooled variance is approximated by $(1 + \frac{\hat{e}_1^2}{4})$; where $\hat{e}_1^2$ is the estimated interim standardised effect size (Friede and Kieser, 2006; Proschan et al., 2006a). Authors point out that, the overestimation by using the pooled variance is normally too small with effect sizes of 10% to 50%, often observed in practice.

Citing the aforementioned limitation of s_{1GS}^2, Zucker (1999) presents an alternative modified unbiased estimate calculated under H_1.

$$s_{1MZ}^2 = s_{L1}^2 - \frac{N_1}{2(N_1 - 1)}{\theta_\delta}^2. \qquad 2:7$$

Jensen and Keiser (2010) propose a modified one-sample variance in the presence of centre effects in multicentre trials estimated by

$$s_{c1}^2 = \frac{1}{N_1 - c}\sum_{i=1}^{2}\sum_{k=1}^{c}\sum_{j=1}^{n_{1k}}\left(Y_{ijk} - \bar{Y}_{.k.}\right)^2, \qquad 2:8$$

where $n_1 = \sum_{k=1}^{c} n_{1k}$ and c is the total number of centres.

Authors have proposed methods using the EM (expectation-maximization) algorithm to compute the maximum likelihood estimate (MLE) of the pooled variance in a blinded manner, without providing reliable interim intervention effect (Gould and Shih, 1992; Gould, 1992). A SAS EM algorithm code is available (Zellner et al., 2001). Friede and Keiser (2002) investigate the performance of this EM algorithm method through extensive simulations and concluded its inappropriateness as a blinded method to estimate the pooled variance. Gould and Shih (2005) refuted these conclusions, conducted further simulation work and claimed their EM algorithm method was reliable. However, further research found the method to yield negatively biased and imprecise estimates (Waksman, 2007). In summary, the EM algorithm method has generated controversies and its reliable application is still questionable (Friede and Kieser, 2002; Gould and Shih, 2005; Waksman, 2007).

2.5.10.2 Influence of Blinded Estimation on Type I Error

For continuous outcomes using a t-test, Kieser and Friede (2003) examine the performance of strictly blinded methods in controlling type I error for the 'restricted' and 'unrestricted' SSR designs. They used numerical integration to study s_{L1}^2 (2:5) and s_{1MZ}^2 (2:7) blinded estimators over a range of internal pilot sample sizes (10 to 50), main trial sample sizes (10 to 100) , statistical power (80% and 90%), and fixed effect size. For all the scenarios considered, they found no inflation to type I error for a 'restricted' design. In addition, a maximum inflation of 0.0001 above nominal was observed for an 'unrestricted' design. In conclusion, they stated that no additional measures are required to adjust significance level at the final analysis, when blinded methods are used. They suggest use of a permutation test when there is desire for strict control of the 0.0001% inflation in the case of an 'unrestricted' design. One limitation of their conclusions is that, they failed to investigate the influence of the magnitude of the effect size and variance misspecification ratio.

In the context of multicentre RCTs, Jensen and Kieser (2010) investigate through simulation, the performance of variance estimators s_{L1}^2 (2:5) and s_{c1}^2 (2:8) for the 'restricted' and 'unrestricted' SSR designs, assuming a weighted and unweighted analysis. Under many scenarios considered, no marked difference in the control of type I error between the two variance estimators and designs was found. The maximum type I error inflation was just 0.0004 under the two SSR designs. The performances of the two SSR designs were found to be similar when the total internal pilot sample size was at least 200.

2.5.11 Unblinded Methods for Continuous Outcomes

2.5.11.1 Estimation of Nuisance parameters

The pooled variance can be computed directly using the knowledge of intervention allocation at the interim (Friede and Kieser, 2006). The methods described here require the knowledge of the intervention allocation of trial participants, which may potentially introduce operational bias in trial conduct.

2.5.11.2 Influence of Unblinded Estimation on Type I Error

Many authors highlight that the revision of the sample size based on unblinded interim estimates of nuisance parameters can influence the type I error (Friede and Kieser, 2006; Herson and Wittes, 1993; Wittes and Brittain, 1990; Wittes et al., 1999). The inflation to the type I error is because after SSR, all the data are used to compute the final test statistic assuming independence between 2 stages (internal pilot and after) when in fact this is not the case (Wittes et al., 1999; Zucker et al., 1999). Research has been undertaken to understand the extent of this problem.

Wittes and Brittain (1990) investigate through simulation, the impact of using the pooled variance (estimated using unblinded data) in the 'restricted' internal pilot, over a range of misspecifications of the ratio of the variance for a fixed internal pilot sample size, and single effect size, aided by an example. A small inflation to the type I error was observed, when the pooled variance is largely underestimated. One limitation is that, they did not investigate the influence of the internal pilot sample size and magnitude of the effect size on the type I error, and variance misspecification ratio yielding the maximum type I error. Birkett and Day (1994) also found similar inflation to the type I error for small internal pilot sample size, but decreased asymptotically to desired nominal levels. The authors reach similar conclusions and argue that the type I error inflation is negligible, and can be traded in favour of increased statistical power (Birkett and Day, 1994; Wittes and Brittain, 1990).

Wittes et al (1999) further study the impact of the 'restricted' and 'unrestricted' internal pilot SSR methods on type I error and statistical power. They conducted simulations using the pooled variance estimated using unblinded data and analysis based on a naïve t-test – without any statistical adjustment. The simulations covered a wide range of variance misspecification ratios, effect sizes, and internal pilot sample sizes. They found that, type I error inflation depended on all these factors. In conclusion, they stated that, a 'restricted' design outperformed an 'unrestricted' design by achieving the desired power and satisfactory type I error control, over a broad range of scenarios. When unblinded methods are used, they advise simulation work at the design stage, over

the range of internal pilot sizes, effect sizes and misspecification ratios to select a design with the smallest type I error inflation. Alternatively, to attain the exact desired type I error, they suggested statistical correction or adjustment to the significance level at the end of the trial, determined through simulation at the design stage. Methods introduced in Appendix 2.5 can be used to control the desired statistical properties.

In theory, Denne and Jennison (1999) suggest a modified independent t-test based on Stein's two stage design by calculating the test statistic with inflated degrees of freedom under the t-distribution. They evaluated its performance under a wide range of scenarios and compared it to an 'unrestricted' internal pilot. They conclude that their method is better than the 'unrestricted' design, and type I and II errors are close to the desired nominal level. However, the inflation problem was not eradicated, although it converges to the desired nominal level with increasing size of the internal pilot. In addition, as highlighted in Section 2.5.4, Stein's two stage design is inefficient and rarely applied in clinical trials.

Kieser and Friede (2000) consider circumstances when strict type I error control is of paramount importance using an unblinded SSR. They numerically computed the actual maximum type I error depending on internal pilot sample size. They also noted decreasing type I error inflation with increasing internal pilot sample size. For example, for the 'unrestricted' designs, with a total internal sample size of 150 and 200 participants, they found a maximum type I error inflation of 1.8% (0.0509) and 1.4% (0.0507), respectively. They suggest partitioning the pooled variance of the final test to reflect participants recruited in the internal pilot (n_1) and after (n_2). Under H_0, the modified t-test statistic t^* follows a t-distribution with $n_1 + n_2 - 4$ degrees of freedom. For an 'unrestricted' design, they suggested adjusting significance level or using t^* for a given size of the internal pilot. They also propose using a one-sided 80% of the interim variance estimate for SSR in order to increase the probability of achieving the desired statistical power.

In the context of a 'restricted' design, Friede and Miller (2012) study the performance of the two stage, repeated and continuous monitoring SSR methods applied in a blinded and an unblinded manner, through extensive simulations. Small type I error inflation was observed for small internal pilot sample size (<50 in total) when unblinded SSR methods were used. In conclusion, they claim that, blinded SSR methods do not suffer the same biases and almost no inflation to the type I error was observed.

2.5.12 Blinded Methods for Binary Outcomes

2.5.12.1 Estimation of Nuisance parameters

The overall event rate ignoring intervention allocation can be used together with the design effect size sought (θ_δ) using an appropriate formula such as equation (2:4) (Gould, 1995). Gould also presents a modified method based on the interim overall event rate and its prior information in the Bayesian context.

The complications during SSR for binary outcomes have been highlighted because the (Herson and Wittes, 1993; Proschan et al., 2006a):

a) Effect size depends on the control event rate and as a result, the re-estimated sample size is quite sensitive to the control event rate;
b) Mean and variance are indistinct parameters and hence, the variance increases with increase in the overall event rate – attaining its maximum when $\overline{p} = 0.5$;
c) Inaccurate assumptions may render the assumed design effect size implausible depending on the parameterised measure of effect size.

Wittes (2002) points out that, investigators are able to deduce the interim intervention effect using the overall event rate and the re-estimated sample size. Wittes suggests a partial solution to obviate this problem by not disclosing the revised sample size, but only communicating the decision to stop the trial when recruitment is reached. Implementation of this in practice may be questionable, given that other practical decisions required will depend on the knowledge of the revised sample size.

Shih and Zhao (1997) devise a randomisation method with 'pseudo' stratification to estimate interim event rates in two groups, also applicable in multicentre RCTs. The authors advise inflating the variance estimate by a factor, arguing that it prohibits Trialists from performing hypothesis testing. This method depends on the choice of the random allocation parameter, for which the authors advise Trialists to select values near the middle of the range 0 to 0.5. Whilst the method seems to provide unbiased estimates of the intervention group event rates without breaking the allocation code, it reveals the intervention effect. In addition, they suggested the SSR depends on the interim observed intervention effect. Hence, Trialists can make indirect inference about the intervention effect based on decisions made after SSR.

2.5.12.2 Influence of Blinded Estimation on Type I Error

For a SSR based on overall event rate, Gould (1992) studies the impact of a 'restricted' design on type I error, and expected sample size and power, through simulation. Consideration was given to tests based on risk difference, risk and odds ratios. The expected theoretical and observed cumulative distributions of rejecting H_0 were similar for tests based on ratio effects, but not on risk differences. Under simulation scenarios considered, the maximum type I error observed was 0.054.

For an 'unrestricted' design based on overall event rate, Friede and Kieser (2004) compute the exact type I error for a test comparing two independent proportions using a Chi-Square test and compared with a fixed design. A wide range of scenarios were considered: 5% to 50% event rate, internal pilot sample sizes (20 to 200), and equal and unequal intervention allocation. They concluded similarity in type I error inflation was observed between the fixed and SSR designs, with a maximum inflation of 0.01 above nominal for internal pilot trials of above 60 per group.

In conclusion, for relatively large internal pilots, Friede and Kieser argue that there is no need to be concerned about the actual type I error inflation if one is willing to accept the anti-conservatism of the Chi-Square test for a fixed sample size design. In situations requiring strict type I error control, they proposed statistical adjustment to ensure the actual error does not exceed the desired error, and illustrated this method using a case study. A SAS code for implementation is available on request from the authors. This approach has been described elsewhere (Kieser and Friede, 2000).

2.5.13 Unblinded Methods for Binary Outcomes

2.5.13.1 Estimation of Nuisance Parameters

The estimate of the control event rate uses interim data of participants allocated to the control group only (Wittes and Brittain, 1990). Herson and Wittes (1993) presented this method for SSR together with the reparameterised H_1, which is a function of the interim observed control event rate. They highlighted that the interim event rate may indicate implausibility of the hypothesised effect size θ_δ under H_1. They argue that using the reparameterised H_1 is a reasonable approach to guide investigators about an appropriate effect size. Day (2000) articulates the difficulties in parameterisation of the intervention effect for binary outcomes when the control event rate is inaccurate.

2.5.13.2 Influence of Unblinded Estimation on Type I Error

Herson and Wittes (1993) conclude through simulation that the use of the control event rate has negligible impact on inflation of type I error. Using a simulation approach, Shih and Zhao (1997) found an inflation ranging from 0.005 to 0.015 above nominal level based on a blinded SSR method using a 'pseudo' stratification scheme, compared to 0.014 to 0.022 for an unblinded method. However, they found that both methods increase or reduce the required sample size appropriately to maintain the desired statistical power.

In summary, there appears to be limited literature exploring the impact of unblinded SSR for binary outcomes on the type I error.

2.5.14 Reflection on Practical Considerations

The applicability of SSR is trial depended (Chiang-Stein et al., 2006). Aspects such as the duration of the endpoint relative to recruitment pace, accrual pace of primary outcome data, and recruitment pace influence the applicability of SSR methods. For example, long-term endpoints or poor recruitment may render the implementation of SSR impractical. In circumstances when SSR is administratively or practically challenging to implement, some may argue in favour of designing of a trial with more power than the minimum desired. However, this approach is problematic because it is likely to lead to unnecessary overpowered and costly RCTs.

A decision to increase the sample size necessitates additional resources. There are logistical and administrative costs due to aspects such as addition of recruiting centres and investigators. Some authors suggest setting up maximum thresholds for sample size increase and minimum sample size to ensure accurate contingency planning and costing at the design stage (Gould, 1992, 1995; Shih and Zhao, 1997). Day (2000) highlights that the decision to stop further recruitment at the internal pilot stage is complex and there are numerous possible decision criteria for consideration.

Besides statistical implications, the choice of the SSR method has implications on operational aspects to maintain confidentiality, credibility, and integrity for acceptability of results (Chiang-Stein et al., 2006; FDA, 2015).The need to maintain blinding wherever possible has been well articulated (Day and Altman, 2000; Schulz and Grimes, 2002) and reflected in regulatory documents (FDA, 2010, 2013, 2015; ICH, 1998). To achieve this, some of the following considerations have been highlighted as important:

a) How the SSR will be implemented (blinded or unblinded)?

b) When appropriate, what are the perceptions of Regulators regarding the proposed SSR method?

c) Who are the key stakeholders who need to know the exact SSR implementation and decision rule?

d) Who will conduct the SSR?

e) Which stakeholders should the SSR results be shared with?

f) Who will recommend the SSR decision rule?

g) How will the details of the SSR method be documented?

2.5.15 Reflection on Regulatory Considerations

In general, it appears that, appropriate use of methods for SSR to maintain the desired statistical power to answer research questions is well recommended from a regulatory perspective (FDA, 2010, 2013, 2015; ICH, 1998). Regulators highlight that planning at the design stage is necessary. The FDA (2010) states that blinded examination of interim data does not introduce statistical bias, and no statistical adjustments are required. More so, the FDA states that blinded SSR methods to maintain desired statistical power should generally be considered for most trials. Blinded SSR methods appear to be preferred by Regulators. The FDA also cautions against downward revision of the sample size, arguing that such a decision may be subsequently regrettable when estimates turns out to be inaccurate, when often based on small internal pilot trials.

In summary, regulatory perspective appears to suggests that, regardless of the SSR method proposed, the rationale and measures to minimise operational bias and consequences (if any), control of type I error, and to obtain unbiased trial results should be explained and supported with evidence and documentation.

2.5.16 Summary

Inaccurate assumptions about nuisance design parameters is a prevalent problem in the design of confirmatory trials, as shall be illustrated in Section 7.5.1 of Chapter 7. Properly implemented SSR methods can help provide accurate sample size estimates to achieve the desired statistical power. However, SSR methods are not excuses for inadequate planning of confirmatory trials. There is still the need for the use of available evidence based information, such as from reviews to inform initial sample size estimation. SSR is then used as a tool to validate such assumptions. Importantly, prospectively planning the proposed SSR method is essential.

The SSR methods considered in this section have some limitations. It is assumed that, the effect size sought is fixed and precisely known in advance. However, for some health conditions, the effect size sought is subject to some degree of uncertainty. In addition, a decision might be taken to increase the sample size whilst the

interim effect size may be too small to warrant the need for further resources. In these situations, it appears logical to consider other approaches as discussed in Section 7.6.1 of Chapter 7.

Blinded SSR methods are simple to implement in-house, well accepted, and have negligible effect on type I error inflation to be concerned about, when based on relatively large internal pilots. The methods can be implemented using a 'restricted' or 'unrestricted' design. However, whenever the 'unrestricted' design is contemplated, carefully consideration should be given to the impact of sample size reduction on other important secondary trial objectives.

One of the blinded variance estimators can be used for SSR based on primary outcomes: one sample variance s_{L1}^2 or modified versions (s_{1MZ}^2 or s_{1GS}^2). The situation is more complicated for binary outcomes, because the event rate, intervention effect, and its variance are connected. Hence, even the use of the overall event rate may reveal some information about the intervention effect. More so, the overall or control event rates may indicate the implausibility of the assumed effect size, often problematic when based on absolute relative risk scales. Herson and Wittes (1993) argue in favour of using the reparameterised H_1 as a function of the control event rate. Day (2000) advises definition of the effect size on a constant odds ratio scale invariant from the overall or control event rate. SSR will then be calculated based on a fixed odds ratio as a function of the considered event rate. However, this is more likely to result in marked sample size increase. Careful consideration by clinical investigators in consultation with the Trial Statistician is important at the design stage. Furthermore, it is vital to pre-specify this approach at the design stage with documentation for audit trails.

Regardless of the SSR methods proposed, it is important to consider measures put in place to minimise operational bias, control type I error, and to obtain unbiased results where necessary and possible. In this section, SSR methods based on estimates of nuisance parameters have been described. In addition, the impact of the methods on type I error and inference, practical considerations and regulatory perspective have been discussed to help Trialists choose appropriate methods. The application of SSR methods for binary outcomes based on overall event rate is illustrated in Chapter 7 and benefits and lessons learned highlighted.

2.6 Design 2: Stochastic Curtailment Futility Analysis

2.6.1 Motivation

Well-conducted and generalisable confirmatory RCTs are costly, time consuming, and require involvement of a huge number of participants across multiple centres. Chapter 1 highlighted that the costs of running RCTs are escalating across sectors while recruitment of participants is becoming harder, with significant numbers of RCTs failing to meet recruitment targets on time. This raises the importance of efficient design and conduct of individual RCTs, within the constraints of available resources.

Chapter 1 highlighted that the proportion of confirmatory RCTs yielding statistically and clinically important results in favour of investigative interventions is low across the public sector. Kaplan and Irvin's (2015) review of large trials funded by the National Heart, Lung, and Blood Institute (NHLBI) in the USA (1970 to 2012) showed a huge increase in the primary results failing to show clinically and statistically significant effects. Despite the underlying cause of the trends towards H_0, the findings highlight that most investigative interventions do not translate into or need to be withdrawn from routine medical practice. As reflected in Section 1.3, although obtaining significant results favouring the investigative intervention does not solely define 'success' in all trials, trends towards H_0 raise fundamental ethical, practical, scientific, and economic questions in trials research. The idea of using accumulating outcome data from an ongoing trial to make decisions about early stopping due to poor or disappointing efficacy results, referred to as futility analysis, is motivated to address ethical and economic issues cited, although reasons vary from trial-to-trial (Gallo et al., 2014).

Over a 30-year period, authors have been highlighting the need to discontinue futile trials. Lan *et al* (1982) highlight that one posed question during the monitoring of accruing outcome data could be whether the current observed results are sufficient to answer the research question. Fleming and DeMets (1993) state that if the investigative intervention effect is known with some degree of certainty, then the trial should not be continued longer than necessary to reach intended objectives. Lachin (2009) highlights the possibility to stop a trial earlier than planned if the emerging effectiveness results suggest that the investigative intervention would not produce beneficial effects.

Statistical methods have been developed to assess futility under different circumstances (Gallo et al., 2014). In this section, focus is on futility analysis based on stochastic curtailment methods. Practical implementation of this approach is demonstrated in Section 7.5.2 of Chapter 7.

2.6.2 Stochastic Curtailment

A curtailment method is described as asking at some point in a fixed sample size trial designed without multiple looks, whether the outcome of a hypothesis test at the end of the trial is already determined (Lan et al., 1982). A deterministic curtailment method, where a decision is made on whether to reject H_0 if the interim test statistic falls within the rejection region of the test or 'accept' H_0 if otherwise has been studied in the context of survival outcomes (Alling, 1963; Halperin and Ware, 1974). For this method, an interim decision is made with certainty. However, it is generally felt that evidence observed at an interim look is not that extreme as to be deterministic in decision-making (Davis and Hardy, 1994; Lan et al., 1982).

As an alternative, Lan et al (1982) describe stochastic curtailment where a decision to stop a trial early is considered when a particular decision is highly likely given the observed interim results. The method aims to address a question on whether it is worth continuing a trial – given the observed interim results and projected future trend (Lan and Wittes, 1988). The method provides statistical rationale for early stopping for a trial, even for trials designed as fixed sample size RCTs (Lachin, 2005, 2009; Lan and Wittes, 1988; Lan et al., 1984).

Although in theory stochastic curtailment can be used to stop a trial early to reject H_0 or H_1 (Lan et al., 1982), group sequential methods are recommended when stopping for efficacy (rejecting H_0) is considered (Ellenberg et al., 2003; Lachin, 2005; Whitehead and Matsushita, 2003). Group sequential methods with options for early stopping either futility and/or efficacy are described in Section 2.7. Therefore, this section focuses only on early stopping for futility using Frequentist methods for stochastic curtailment based on conditional power. A brief reflection of the Bayesian equivalent method - predictive power is given for completeness.

2.6.3 Conditional Power

Conditional power (CP) is the probability of rejecting H_0 (finding clinically important and significant results) at the end of the trial – given the interim observed results, under specific assumptions about the pattern of the future unobserved results for the remainder of the trial (Betensky, 2000; Davis and Hardy, 1990, 1994; Lachin, 2005). That is, a very low CP informs Clinical Trialists that the trial is probably futile, and unlikely to yield clinically relevant and statistically significant results (Proschan et al., 2006c).

Lan and Wittes (1988) describe the concept of B-values, that arise from observed trends of effectiveness results. These B-values are assumed to follow a Brownian motion process and are a direct transformation of the interim test statistic $Z(t)$ and how far the trial has progressed ($t \in [0,1]$), as expressed by

$$B(t) = Z(t)\sqrt{t}\,. \qquad 2:9$$

The authors demonstrate that the B-value at the end of the trial $[B(1)]$ can be partitioned to reflect independent increments of the non-random interim observed results $B(t)$ and the random unobserved future trend of results $[B(1) - B(t)]$. The properties of the Normal and joint distributions of the independent increments have been extensively studied (Lachin, 2005; Lai et al., 2000; Lan and Wittes, 1988; Proschan et al., 2006c; Zhang et al., 2015). The use of B-values and their known distributions is central to the calculation of CP (Section 2.6.4) as the functions lead to a linear trend of results with time, which is natural to visualise, and easy to interpret (Lan and Wittes, 1988; Proschan et al., 2006c).

2.6.4 Computation of Conditional Power

Lachin (2005) gives a comprehensive review of futility analysis methods based on CP, and investigates its statistical properties through numerical integration, aided by examples. The methods reviewed are applicable for all outcomes, although this thesis deals with binary and continuous outcomes. The computation of CP is summarised as follows:

a) Calculate the information fraction ($t \in [0,1]$) observed at the interim. Lan and Zucker (1993) highlight that an appropriate statistical measure of how the trial has progressed is the amount of statistical information accumulated, which may be reflected by the sample size in certain circumstances. For comparing two groups of unequal sizes, Lan et al (2005) describe the interim information fraction as

$$t = \frac{\left(\frac{1}{n_C} + \frac{1}{n_T}\right)^{-1}}{\left(\frac{1}{N_C} + \frac{1}{N_T}\right)^{-1}}, \qquad 2:10$$

where (n_C, n_T) and (N_C, N_T) are the interim and planned sample sizes in the control and intervention arms, respectively; $t = 0$ and $t = 1$ corresponds to the trial start and planned end, respectively;

b) Calculate the standardised test statistic $Z(t)$ corresponding to the interim information fraction $t \in (0,1]$. Under $H_0: \theta = 0$, Lan and Zucker (1993) define it as:

$$Z(t) = \frac{\hat{\theta}_t}{se(\hat{\theta}_t)}, \qquad 2:11$$

where $\hat{\theta}_t$ and $se(\hat{\theta}_t)$ are the interim intervention effect and its standard error, respectively;

c) For a two-sided test, Lan and Wittes (1988) describe the formula to calculate CP as

$$CP_\theta(t) = 1 - \Phi\left(\frac{Z_{\frac{\alpha}{2}} - B(t) - \theta_f(1-t)}{\sqrt{1-t}}\right), \qquad 2:12$$

where θ_f is the drift parameter representing the assumed projection of the random future trend of results and $B(t)$ is expressed by equation (2:9).

2.6.5 Statistical Properties of Conditional Power Futility Analysis

The use of CP for futility analysis raises a number of questions pertinent to Clinical Trialists during the design and conduct of trials (Gallo, 2015; Gallo et al., 2014; Herson et al., 2012; Lachin, 2005; Leung et al., 2003).

1) What futility criteria to use for decision-making?
2) What assumptions to make about the future trajectory of results after the interim to calculate CP?
3) When is it appropriate to perform futility analysis based on CP and how many times?
4) How are the design statistical properties (type I and II errors) affected by undertaking futility analysis based on CP?

The next sections review literature to help Clinical Trialists to address some of the raised questions.

2.6.6 Futility Criteria for Decision-Making

It has been suggested that Trialists might stop a trial early for futility when the CP under H_1 given the interim observed results, of accepting H_0 is greater than some arbitrary value $\omega \in (0,1]$ (Halperin et al., 1982; Lan et al., 1982). For this approach, high ω values indicate that it is more likely to 'accept' H_0, even if H_1 is true. That is, low values of $(1-\omega)$ are consistent with H_0, undesirable, and offer a provision for futility early stopping.

A decision-making criterion recommending early stopping for futility when $CP_\theta(t) \leq 1-\omega$ has been proposed (Betensky, 2000; Halperin et al., 1982; Lachin, 2005; Lan et al., 1982). Pre-defined high ω values in the range 50% to 100% are proposed (Lan et al., 1982), although values above 80% or 90% are recommended, citing 'negligible' impact on statistical properties presented in Section 2.6.9 (Davis and Hardy, 1990). The complimentary probability $(1-\omega)$ translates to a low probability of rejecting H_0, even when H_1 is true.

Ware *et al* (1985) suggest that a trial could be stopped for futility if the CP for demonstrating efficacy at the planned end under H_1 falls below an arbitrary threshold of 33%. The authors advised weighing early stopping benefits against the loss of statistical power when choosing the futility threshold. This was illustrated with an example based on limited combinations of power and futility thresholds as functions of the sample size to create futility stopping boundaries. Friedman et al (2010) suggest, aided with an example, that early futility stopping could be considered if the CP falls below a threshold of 10% to 30%. Herson et al (2012) strongly argue against the use of large CP futility thresholds within the range of 40% to 50%, and recommend threshold values between 15% and 25% as the basis for futility decision-making.

In summary, the literature suggests that there is no single clinical threshold applicable for every trial situation. Although futility thresholds ranging from 10% to 30% are suggested in practice, the choice of the threshold to use is trial dependent guided by other factors such as the amount of evidence already in practice, and impact of generated results.

2.6.7 Assumptions Regarding Future Trend of Results

It has been widely highlighted that there is no unique way to calculate the CP for an ongoing trial since it depends on assumptions made regarding the unobserved future results (Betensky, 2000; Lachin, 2005; Proschan et al., 2006c). That is, the CP is calculated under a number of assumptions about the random unobserved future trend of results (θ_f on equation (2:12)) to aid decision-making (Ellenberg et al., 2003; Friedman et al., 2010). For example, Lan and Wittes (1988) show that when the CP is calculated under H_0 ($\theta_f = 0$), interim results ($\theta_f = \hat{\theta}_t$), and H_1 ($\theta_f = \theta_\delta$), equation (2:12) simplifies to

$$CP_\theta(t) = \begin{cases} 1 - \Phi\left(\dfrac{Z_{\alpha/2} - B(t)}{\sqrt{1-t}}\right) \\ 1 - \Phi\left(\dfrac{Z_{\alpha/2} - B(t) - \hat{\theta}_t(1-t)}{\sqrt{1-t}}\right) \\ 1 - \Phi\left(\dfrac{Z_{\alpha/2} - B(t) - \theta_\delta(1-t)}{\sqrt{1-t}}\right) \end{cases}. \qquad 2:13$$

It has been argued that the calculation of the CP under the planned H_1 ($\theta_f = \theta_\delta$) be considered when stopping for futility (Halperin et al., 1982; Lan et al., 1982). However, Jennison and Turnbull (2000) highlight that this approach is problematic when the interim trend results are inconsistent with the planned and assumed H_1 effect. The problem is more profound in trials designed with overoptimistic and unrealistic effect sizes sought under H_1 (Betensky, 2000; Pepe and Anderson, 1992). These authors argue that in such cases, the CP assuming

H_1 ($\theta_f = \theta_\delta$) may not provide a reasonable stopping criteria for decision-making, and make it difficult to stop for futility even if the interim intervention effect is too small.

The use of the upper CI limit of the interim intervention effect to calculate the CP, rather than assuming H_1, has been proposed (Betensky, 2000; Herson et al., 2012; Lachin, 2009; Lan and Wittes, 1988; Pepe and Anderson, 1992). Betensky (2000) argues that this approach is more conservative and yields lower stopping boundaries. Herson et al (2012) suggest the use of the optimistic upper 80% CI point of the interim intervention effect. Although the CP is calculated for a specific alternative hypothesis, wide ranges of scenarios are often considered in routine practice for decision-making by the IDMC (Independent Data Monitoring Committee) say (Ellenberg et al., 2003; Herson et al., 2012; Proschan et al., 2006c).

In conclusion, available literature seems to suggest that whenever stopping early for futility is considered based on CP, it is advisable to calculate the CP under a number of reasonable agreed assumptions consistent with the observed data and planned H_1 to aid robust decision-making.

2.6.8 Timing of Futility Analysis

The influence of the size of the interim information fraction on the statistical properties of CP futility analysis for decision-making has been widely highlighted (Halperin et al., 1982; Lan et al., 1982, 1984; Proschan et al., 2006c). Gallo et al (2014) underline the conflicting interests in stopping the trial very early in order to maximise potential benefits and the availability of 'adequate' information for robust futility decision-making. Ellenberg et al (2003) state that although performing futility analysis very early on produces greatest potential benefits, there is often huge variability reflecting marked uncertainty about the true intervention effect. The authors however, did not suggest optimal criteria for the timing of futility analysis. Gallo et al (2014) demonstrate with an example halfway through the trial and suggest that the timing of futility analysis can be chosen based on the trade-off between potential benefits and expected loss in statistical power. Sully et al (2014) used simulations to predict the proportion of MRC and NIHR HTA funded trials from 2002 to 2008, which could have been stopped early based on futility analysis. The authors assumed futility was undertaken after 50% to 90% of planned recruitment based on futility stopping thresholds of 20%, 30% and 40%. The authors argue in favour of analysis at 75% of target recruitment based on maximum benefits gained.

In summary, although there does not appear to be a consensus on exact timing, reviewed literature advises against conducting a futility analysis with a relatively small interim information fraction or too late in the course

of the trial. The timing of futility analysis could be informed by weighing the chances of futility early stopping and benefits of doing so against the magnitude of loss in power. Futility analysis planned between 50% and 75% of the target recruitment appears to be reasonable, as shall be highlighted using case studies in Section 7.5.2 of Chapter 7.

2.6.9 Frequency of Futility Analysis and Impact on the Type I and II Errors

2.6.9.1 Type I Error

Lan et al (1982) show through numerical integration that the overall maximum type I error committed for performing an infinite number of interim futility analyses based on CP is

$$\alpha = \frac{\alpha_D}{\omega'}, \qquad 2:14$$

where α_D and ω' are the planned desired type I error and the probability of rejecting H_0 at the end of the trial when H_1 is true. For example, for a trial designed with a two-sided 5% type I error, a CP futility threshold of 20%, and power 90% (ω'), the maximum type I error at the end following infinite futility analyses is $\frac{0.05}{0.9} = 0.056$. Using an example, the authors also illustrate the exact type I error for a finite number of equally spaced interim looks. Davis and Hardy (1990) show the exact type I errors using case studies under a range of scenarios; numbers of equally spaced interims (2, 5 and 10) and futility thresholds (50%, 20% and 10%). The authors show that the exact type I errors are much lower than the maximum bounds for low futility thresholds (for example, 10% and 20%). In a later manuscript, Davis and Hardy (1994) further state that the exact overall type I error is 5.1% assuming 5 equally spaced interims for a trial designed with a 5% type I error and futility threshold of 20%.

Chang and Chuang-Stein (2004) and Lachin (2005) numerically studied the influence of one interim futility analysis on the type I error and give numerical formulae to calculate the exact type I error committed. Lachin (2005) states that if the decision is made to stop a trial based on lower CP than the planned futility threshold (CP_L), then there is no type I error committed. However, when a decision is made to continue a trial because the CP is larger than the planned futility threshold (CP_L), then the exact overall type I error for a two-sided test is evaluated by

$$\alpha = \int_{b=B_L}^{\infty} f_0(b) \int_{u=Z_{1-\alpha_D}-b}^{\infty} g_0(u)\, du\, db\,, \qquad 2:15$$

where the distribution functions of $f_0(b)$ and $g_0(u)$ are

$$B(t) = f_0(b) \sim N(t\theta, t)\,, \tag{2:16}$$

$$B(1) - B(t) = g_0(u) \sim N[\theta(1-t), 1-t]\,, \tag{2:17}$$

$$B_L = \phi^{-1}(CP_L)\,. \tag{2:18}$$

All the cited authors emphasise that when a single futility analysis is undertaken, there is no inflation to the overall type I error. In fact, the exact type I error will be lower than the desired nominal level with its size positively correlated with the magnitude of the interim intervention effect or negatively correlated with the probability of stopping for futility. When a trial is continued because the CP is above the threshold set and there is a desire to preserve the type I error, Chang and Chuang-Stein (2004) suggest raising the significance threshold for the final test and provide a numerical formula for its calculation.

2.6.9.2 Type II Error

Lan et al (1982) prove that for an infinite number of looks, the maximum type II error committed is

$$\beta = \frac{\beta_D}{1-\omega}\,, \tag{2:19}$$

where the desired type II error is β_D under the design and $\omega > 1 - \omega'$ is a necessary condition. For example, for a trial planned with 80% power ($\beta_D = 0.2$) and a CP futility threshold of 20%, the maximum type I error when infinite futility analyses are performed is $\frac{0.2}{0.8} = 0.25$. Chang and Chuang-Stein (2004) numerically study the influence of one interim futility analysis on the type II error and give numerical formulae for its calculation. The authors conclude that the study power is relatively comparable to the planned for a futility threshold of ≤40% and after 50% of target recruitment. Lachin (2005) partitions the overall type II error as the sum of the probability for stopping for futility at the interim (and committing a type II error), and the probability of continuation to the planned end and finding statistically non-significant results. Lachin further gives a formula to compute the exact type II error for one futility analysis through numerical integration. Ellenberg et al (2003) state that if a futility CP threshold of 20% is used for a trial designed with 85% to 90% power, the increase in type II error is negligible to be concerned about.

Lachin (2009) argues that an alternative approach to reduce the type II error while providing a high probability of stopping under H_0 is to undertake futility analysis later in the trial. In cases where strict control of statistical power is a necessity, some authors suggest joint numerical integration to evaluate the exact type I and II errors for a given futility stopping boundary, interim information fraction, and nominal critical value to adjust

the critical value and preserve desired errors (Jennison and Turnbull, 2000c; Lachin, 2005). That is, a trial can be *a priori*-designed to yield exact desired statistical properties.

In conclusion, the frequency and timing of the futility analyses and the choice of the CP futility threshold influence the type I and II errors. Reviewed literature suggests that conducting futility analysis only once based on CP reduces the type I error and has a 'negligible' impact on the type II error. This holds for a low CP futility threshold (for example, 10% to 25%), suitably large planned statistical power (for example, above 80%), and relatively large interim information fraction for robust estimation of nuisance parameters and intervention effect (for example, 50% to 75% of target recruitment).

2.6.10 Statistical Software for Implementation

The computation of CP for futility analysis can be easily implemented in a number of statistical software such as Stata, R and SAS by invoking the cited equation and procedure in Section 2.6.4. Similarly, the type I and II error bounds are calculated using equations (2:14) and (2:17). There are open access SAS programs (BSC, n.d.) available to implement CP futility analysis and evaluate its statistical properties described by Lachin (2005, 2009). This thesis will implement the approach described in Section 7.5.2 of Chapter 7 in Stata 14.1. Stata implementation code will be made publicly accessible as a package.

2.6.11 Limitations of Conditional Power Futility Analysis

As highlighted in Section 2.6.7, there is no unique 'optimum' approach to perform CP since different assumptions at the interim are made about the future unobserved results. Thus, for a fixed futility threshold, the CP calculated under different scenarios may produce inconsistent decisions (Gallo et al., 2014). Spiegelhalter et al (1986) propose an alternative Bayesian method referred to as predictive power (PP), which is a weighted function of the CP. Instead of using a single point estimate of the hypothesised intervention effect, a series of hypothesised intervention effects are represented by its prior distribution (Demets, 2006). The prior distribution of θ_δ is then updated using interim outcome data to give a posterior distribution. The PP is then computed as a weighted function of the interim data and the prior distribution of θ_δ averaged over the posterior distribution. Jennison and Turnbull (2000) state that the calculated PP can be used to stop a trial for futility based on stopping decision criteria similar to CP as highlighted in Section 2.6.6. One limitation of the PP is the need to specify the

prior beliefs about θ_δ through a prior distribution, which is not straightforward, and may be controversial in the confirmatory setting, particularly when there is little previous evidence (Fisher, 1996; Friedman et al., 2010).

The CP approach described here assumes that the drift parameter estimate $B(t)$ is a linear function of increasing information fraction. This assumption is untenable in the presence of a population drift, where participants recruited early on differ from later participants during the course of the trial (Gallo et al., 2014; Proschan et al., 2006c; Zhang et al., 2015). In such cases, the method may be unreliable because the intervention effect estimated early on may turn out to be inconsistent later in the trial limiting its application in some settings.

2.6.12 Summary

As a single futility analysis, the CP is a simple method, easy to understand and implement with negligible impact on the statistical properties of the design when performed late on in the trial (such as at 50% to 75% of target recruitment) and with reasonably low futility thresholds (such as 10% to 30%). The influence of timing of futility analysis and the choice of futility threshold on decision-making has been discussed. However, the method has some highlighted limitations. The need for and application of CP futility analysis is trial dependent on aspects such as:

- The research question, the amount of evidence already available in practice; and potential impact of the generated evidence;
- The health condition under investigation and ethical considerations;
- Potential benefits such as savings in terms of financial cost, patients and trial duration;
- Accrual of the primary outcome data relative to the expected recruitment.

Although there are different ways to calculate the CP, values assuming the interim results or upper CL point or H_1 are suitable candidates. The choice of the CP futility threshold (10% to 30%) is trial dependent and should be agreed upon by the IDMC and clinical investigators. For example, a very low futility threshold may be selected for a trial investigating a condition where there is very little or no evidence and the impact of the generated evidence is very important. In addition, it is advisable to avoid conducting futility analysis very early on in the trial when there is huge uncertainty due to aspects, such as learning effects when the research team is grasping trial procedures and processes.

With a sizable proportion of trials seeking funding extensions in the public sector mainly because of poor recruitment, CP based futility can be a useful tool for Public Funders to assess value for money prior to granting

additional funding conditional on prior results observed. Public funders could make this approach mandatory as part of their grant funding process.

2.7 Design 3: Group Sequential Design

Section 2.6 presented stochastic curtailment futility analysis, which is also applicable for trials designed without formal interim analyses. However, it does not preserve the desired statistical properties of the design although the impact is 'negligible' when conducted once with 'adequate' information. This section focuses on methods prospectively planned with interim analyses to accurately preserve the design statistical properties.

In a GSD, hypothesis testing is performed multiple times using primary outcome data from groups of accruing participants during an ongoing clinical trial (Wason, 2015). Here, the focus is on standard two-arm RCTs with options for early stopping for futility and/or efficacy at interim analyses. The practical application of group sequential trials and their state of reporting is reviewed in Chapter 6 building on results from Chapters 3 and 4. Some of the results of this section guide the planning and implementation of the methods using case studies in Chapters 7 and 8. Additional material without direct application in the remainder of this thesis is provided as appendices.

2.7.1 Motivation

The development and testing of investigative interventions is time consuming and expensive (DiMasi et al., 2003). Importantly, the proportion of confirmatory trials yielding clinically relevant and statistically significant results with the potential to translate into practice is very low (Dent and Raftery, 2011; Kaplan and Irvin, 2015). A GSD is primarily motivated by the desire to stop trials early for ethical, trial efficiency, and economic reasons (Jennison and Turnbull, 2000a; Kittelson and Emerson, 1999). In Chapter 3 under Section 3.4.3, perceived potential opportunities resulting from early stopping of ongoing trials as soon as there is sufficient evidence to address the research questions are presented. These are from in-depth interviews of key stakeholders in clinical trials research based on their experiences and perceptions.

2.7.2 Description of the Methodology

General methodology underpinning GSDs has been well articulated in the literature (Emerson and Fleming, 1989; Jennison and Turnbull, 2000a; Pocock, 1977; Todd, 2007; Whitehead, 1997). Let $\{l_1, l_2, \ldots, l_k\}$

and $\{u_1, u_2, \dots, u_k\}$ be the lower and upper stopping boundary values corresponding to interim information fraction $\{t_1, t_2, \dots, t_k\}$ respectively. The area between the sets $\{(l_1, u_1), (l_2, u_2), \dots, (l_k, u_k)\}$ is the trial continuation region. These boundaries are used to base decisions to stop a trial for efficacy in favour of the investigative intervention or comparator, respectively. Here, a positive intervention effect is assumed beneficial. The $\{Z(t_1), Z(t_2), \dots, Z(t_k)\}$ are the corresponding interim standardised Z test statistics such that

$$Z(t_j) = \frac{\hat{\theta}(t_j)}{se(\hat{\theta}(t_j))} = \hat{\theta}(t_j)\sqrt{I(t_k)}\,. \qquad 2:20$$

In addition, the $\{Z(t_1), Z(t_2), \dots, Z(t_k)\}$ follows a joint multivariate normal distribution such that

$$Z(t_k) \sim N\left(\theta\sqrt{I(t_k)}, 1\right), \qquad 2:21$$

$$Cov\left(Z(t_{k-1}), Z(t_k)\right) = \sqrt{\frac{I(t_{k-1})}{I(t_k)}}\,. \qquad 2:22$$

Figure 2.3 is a schematic diagram for a generalised GSD with an option for early stopping for efficacy.

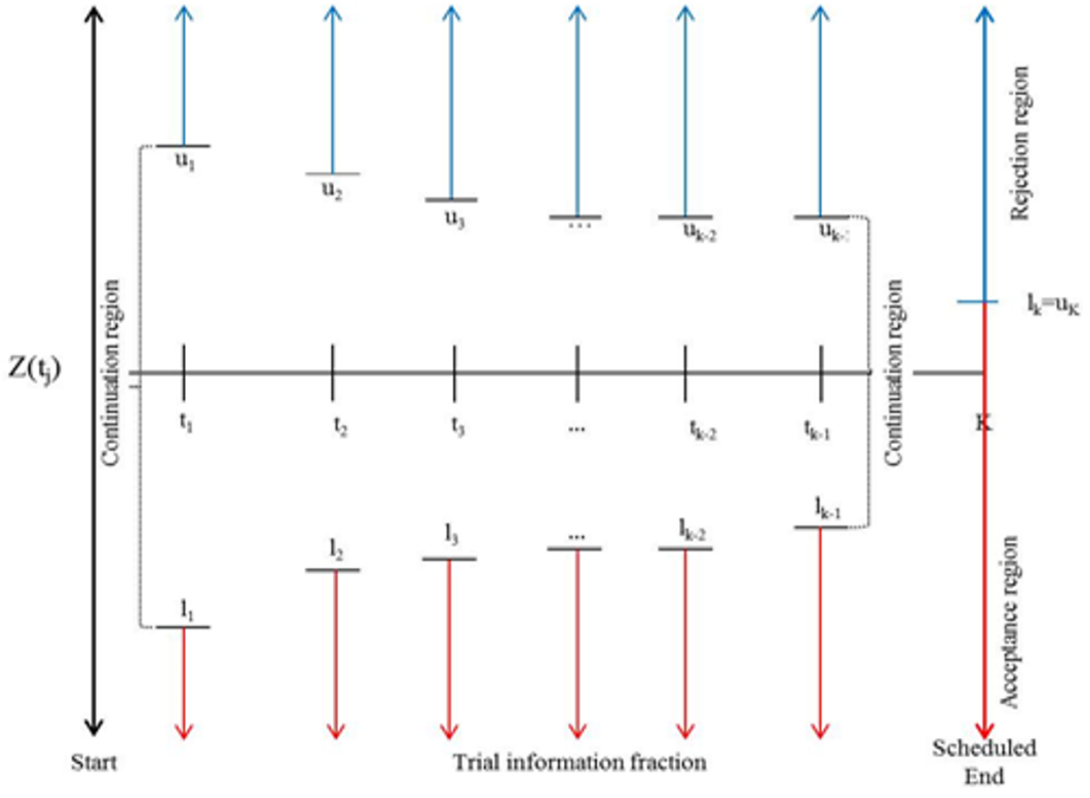

Figure 2.3. Generalised two-sided group sequential superiority test with efficacy stopping boundaries.

At the interim analysis corresponding to information fraction t_j, $Z(t_j)$ value is calculated and compared to the corresponding stopping boundary values $\{l_j, u_j\}$. A decision is then made on whether to stop a trial early for efficacy depending on the research question under consideration. Table 2.1 summarises the decision-making criteria for a schematic GSD illustrated in Figure 2.3.

Table 2.1. Decision criteria for a generalised two-sided group sequential superiority test.

Information fraction	Criteria	Decision rule
At any interim analysis $(t_1, t_2, \dots, t_{k-1})$	If $Z(t_j) \geq u_j$ or $Z(t_j) \leq l_j$	Stop the trial to reject H_0 for efficacy favouring investigative intervention or comparator, respectively
	(Symmetric case)	
	If $Z(t_j) \geq c_j$ or $Z(t_j) \leq -c_j$	Stop the trial to reject H_0 for efficacy favouring investigative intervention or comparator, respectively
	Otherwise	Continue recruitment to interim k
At the scheduled end (t_k)	If $Z(t_k) \geq u_k$ or $Z(t_k) \leq l_k$	Reject or 'accept' H_0, respectively
	(Symmetric case)	
	If $Z(t_k) \geq c_k$ or $Z(t_k) < c_k$	Reject or 'accept' H_0 respectively

$u_j = l_j = \pm c_j$ for a two-sided symmetric test; Note: positive intervention effect is assumed beneficial.

2.7.3 Expression of Stopping Boundaries

The stopping boundaries $\{l_j, u_j\}$ can be expressed in various scales, such as crude estimate of intervention effect, standardised Z test statistic, and p-value (Emerson and Fleming, 1989; Emerson et al., 2007). Whitehead (1997) expressed the test statistic in terms of the Fisher's information. Todd et al (2001) supports this approach by stating that the test statistic measuring the intervention effect may increase or decrease between interim analyses, while the test statistic measuring information will always increase during the course of the trial. Emerson et al (2007) provide mathematical expressions of various stopping boundary scales. Importantly, the authors highlight that stopping boundaries defined on one scale induce the stopping boundaries for all the other scales used to specify decision-making rules.

In summary, the way the stopping boundary scale is defined is immaterial as long as the statistical characteristics of the design are well evaluated. In addition, the stopping boundary scales can be chosen based on easy interpretability by wider members of the research team with diverse research backgrounds (Rudser and Emerson, 2008). Case studies presented in Chapters 7 and 8 use three boundary scales.

2.7.4 Effect of Interim Analyses on Type I Error and Power

Armitage et al (1969) highlight through simulation and numerical integration that multiple hypothesis testing of accumulating outcome data at the same pre-planned level increases the probability of finding a false significant result at some interim analyses. The authors demonstrate that the type I error inflation increases with an increase in the number of interim analyses. The same phenomenon is highlighted by a number of authors (Haybittle, 1971; O'Brien and Fleming, 1979a; Pocock, 1983). Armitage et al (1969) suggest performing interim analyses using adjusted interim type I errors in order to preserve the overall pre-planned type I error. That is, the critical values used for interim analyses are larger than for a corresponding fixed sample size design. Slud and Wei (1982) illustrate this concept of multiple interim hypothesis testing in the context of survival outcomes using a modified Wilcoxon test.

Jennison and Turnbull (2000c) highlight that early stopping because of interim analyses reduces statistical power. Proschan et al (2006) further illustrate the inflation to the fixed sample size design (without interim analysis) to compensate for power loss for a specified number of interims, planned power, and Pocock and O'Brien and Fleming (OBF) stopping boundaries described in Section 2.7.5. In summary, interim analyses increase the type I error and reduce statistical power, unless a trial is designed properly to preserve these statistical properties.

2.7.5 Stopping Boundaries

Section 2.7.2 introduced the concept of stopping boundaries $\{l_j, u_j\}$ without details on how these values are constructed. Section 2.7.4 discussed the impact of interim hypothesis testing on type I error and power. A number of procedures have been proposed to 'raise the bar' of evidence required to reject H_0 based on $\{l_j, u_j\}$ values to preserve these statistical properties. This section describes some of the proposed stopping boundaries for symmetric two-sided tests, unless stated otherwise. In this case, a decision to stop a trial early to reject H_0 is made if $|Z(t_j)| \geq c_j$. In general, the type I error is preserved by solving equation (2:23) though numerical integration (Vandemeulebroecke, 2008).

$$P_0\left(\bigcap_{j=1}^{k}|Z(t_j)| < c_j\right) = 1 - \alpha\,. \qquad 2{:}23$$

The notation $C_{j_{HP}}(k,\alpha)$, $C_{j_P}(k,\alpha)$, $C_{j_{OBF}}(k,\alpha)$, $C_{j_{WT}}(k,\alpha,\Delta)$ and $C_{*_{PT}}(k,\alpha,\beta,\Delta_*)$ is adopted to represent the Haybittle-Peto (HP), Pocock, OBF, Wang and Tsiatis (WT), and Pampallona and Tsiatis (PT) stopping boundaries, respectively. Stopping boundaries which are not frequently used such as the Whitehead Triangular test (Whitehead and Stratton, 1983; Whitehead, 2000), WT (1987) and CP are presented in Appendix 2.2 for completeness.

2.7.5.1 Haybittle-Peto

Haybittle (1971) and Peto *et al* (1976, 1977) suggest an approach that prohibits premature early stopping with insufficient evidence to reject H_0 whenever $|Z(t_j)| \geq 3; \forall\, j \leq k-1$ and $|Z(t_k)| \geq C_{k_{HP}}(k,\alpha)$. However, the constant $C_{k_{HP}}(k,\alpha)$ cannot be chosen to preserve the exact overall type I error for large values of k. Jennison and Turnbull (2000) demonstrate that for $k \geq 7$ and $\alpha = 5\%$, there is no $C_{k_{HP}}(k,\alpha)$ value to preserve an overall α. Nonetheless, the authors acknowledge that the number of interims is often small in practice. The authors state that boundaries do not possess properties that yield the maximum reduction in the expected sample size compared to other boundaries. Fleming *et al* (1984) extend HP ideas and suggest a modification by imposing a constant restriction on the level of significance testing up to interim t_{k-1} obtained through recursive numerical integration, such that the overall significance level is exactly preserved at the pre-specified level regardless of the magnitude of k.

2.7.5.2 Pocock

Pocock (1977) suggests choosing the values $c_j; \forall j$ to be constant ($C_{k_p}(k,\alpha)$), depending only on the total number of planned interim analyses k and α. Thus, at each interim analysis, H_0 is rejected if $|Z(t_j)| \geq C_{k_P}(k,\alpha)$. The original procedure is limiting as it requires equally spaced interims and k to be specified in advance. DeMets and Ware (1980) extend this approach for one-sided and asymmetric tests. The main shortcoming of the Pocock approach is that it requires the same level of evidence at all interims to reject H_0, as illustrated on Figure 2.4. In practice, the research community and decision-makers often require overwhelming evidence at the earlier interim analyses (Ellenberg et al., 2003). Proschan et al (2006b) argue in favour of prohibiting stopping very early during a trial where there is both high statistical and non-statistical variability. The authors state that there are early learning effects during the trial conduct and it takes some time to reach consistency in the delivery and adherence to the trial protocol. This could influence inconsistency in decision-making.

2.7.5.3 O'Brien and Fleming

OBF (1979) propose an approach where the interim nominal significance increases as a function of the information fraction. Thus, prohibiting early stopping at the earliest interim analyses where there is huge uncertainty with respect to clinical effectiveness and safety, and the data are often inadequate to address key model assumptions (Ellenberg et al., 2003; Proschan et al., 2006b). As the information fraction increases the restrictions on type I error spending are relaxed. This fits in with the wishes of Clinical Trialists who do not wish to stop a trial prematurely with insufficient evidence based on less reliable and unrepresentative data (Ellenberg et al., 2003; Mazumdar and Bang, 2011).

The values of c_j are chosen such that at each interim analysis H_0 is rejected if $|Z(t_j)| \geq c_j$ such that

$$c_j = \frac{C_{j_{OBF}}(k,\alpha)}{\sqrt{t_j}} \; ; \forall j. \qquad 2:24$$

Equation (2:24) is substituted in (2:23) to compute the constant $C_{j_{OBF}}(k,\alpha)$. Figure 2.4 illustrates an example of Pocock and OBF stopping boundaries with four equally spaced interim analyses and 5% overall type I error.

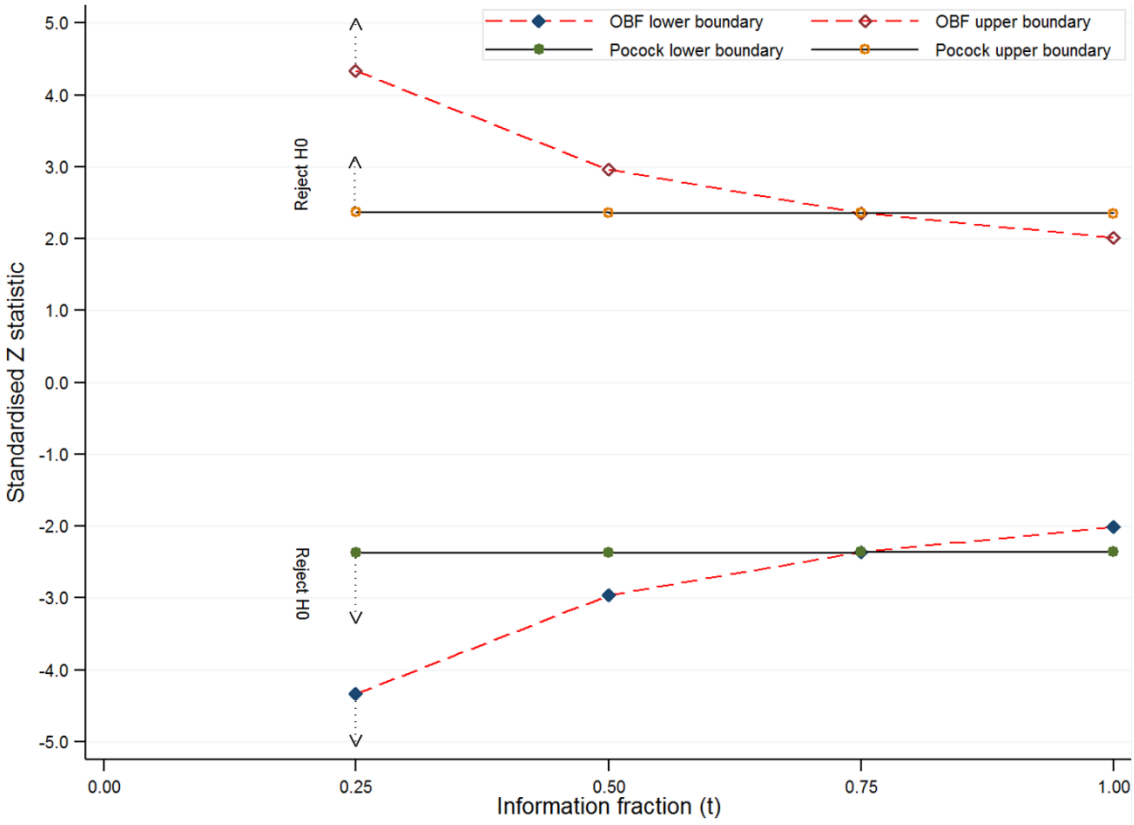

Figure 2.4. Pocock and OBF stopping boundaries for a two-sided group sequential test.

Like the Pocock, the original OBF approach requires k to be fixed and known in advance. However, this assumption can be relaxed for greater flexibility as described in Section 2.7.5.5. Moreover, original OBF stopping rules do not allow for futility early stopping under H_0 (Sebille and Bellissant, 2003).

2.7.5.4 Pampallona and Tsiatis

Pampallona and Tsiatis (PT) (1994) extend the work of Wang and Tsiatis (1987) (described in Appendix 2.2) and propose stopping boundaries for efficacy and/or futility early stopping for one or two-sided tests. For a two-sided test, the PT approach is characterised by positive constants $C_{1_{PT}}(k, \alpha, \beta, \Delta_1)$ and $C_{2_{PT}}(k, \alpha, \beta, \Delta_2)$ which are used to compute the boundaries for early stopping for efficacy (c_j^1) and futility (c_j^0) respectively, such that

$$c_j^1 = C_{1_{PT}}(k, \alpha, \beta, \Delta_1) t_j^{\Delta_1 - \frac{1}{2}}, \qquad 2:25$$
$$c_j^0 = t_j \theta_1 - C_{2_{PT}}(k, \alpha, \beta, \Delta_2) t_j^{\Delta_2 - \frac{1}{2}},$$

where $c_j^0 \leq c_j^1; \forall j$. The constraint $c_k^0 = c_k^1$ is imposed at the scheduled end. A decision is made to stop for efficacy if $\left|Z(t_j)\right| \geq c_j^1$ or for futility if $|Z(t_j)| \leq c_j^0$. The authors highlight that although they assumed that $\Delta_1 = \Delta_2$, the constants Δ_1 and Δ_2 can be chosen differently depending on the desired boundary configurations under H_0 and H_1. For a one-sided test, a decision to stop for efficacy or futility is made if $Z(t_j) \geq c_j^1$ or $Z(t_j) \leq c_j^0$, respectively.

2.7.5.5 Lan and DeMets Spending Functions

The original HP, Pocock, OBF, WT and PT boundaries described so far require pre-specification of the number of equally spaced interim analyses, which is often limiting in practice. For example, the IDMC tasked to make decisions recommending early stopping may decide to change the frequency and timing of interim analyses for some reason (Demets and Ware, 1980; Lan and DeMets, 1983, 1989; Proschan et al., 2006b). These authors cited reasons such as logistics, slower recruitment than anticipated, worrying safety signals, and other emerging external relevant information. Hence, there is a need to allow for some flexibility in the timing and frequency of interim analyses. As an answer to this challenge, Lan and DeMets (LD) (1983) devise stopping boundaries defined through an α spending function.

The LD function describes how the overall type I error is spend during the course of the trial using a standard Brownian motion process. The authors underline that the α spending function is only influenced by the past and current decision times, and not by the unobserved future. The c_j's using α spending functions are

computed using recursive numerical integration under H_0 at each interim t_j to satisfy (Pampallona et al., 2001; Vandemeulebroecke, 2008)

$$P_0(|Z(t_1)| \geq c_1) = \alpha(t_1)\,, \tag{2:26}$$

$$P_0\left(\bigcap_{j=1}^{k-1}\{|Z(t_j) < c_j|\}\bigcap \{|Z(t_j)| \geq c_k\}\right) = \alpha(t_j) - \alpha(t_{j-1}); \forall j \geq 2\,, \tag{2:27}$$

where $\{\alpha(t_j) - \alpha(t_{j-1})\}$ and $\alpha(t_j)$ are the probabilities of making a type I error at and up to interim analysis t_j, respectively. Thus, the overall type I error is preserved such that

$$\alpha(t_1) + \sum_{j=2}^{k}\left(\alpha(t_j) - \alpha(t_{j-1})\right) = \alpha(t_k) = \alpha\,. \tag{2:28}$$

There are several proposed ways to specify the function $\alpha(t_j)$. Some of these functions are expressed in Appendix 2.3. For example, some authors describe Rho (ρ) family (Jennison and Turnbull, 1990, 1989, 2000a; Kim and Demets, 1987). In addition, Hwang et al (1990) describe gamma family. Importantly, the shape parameters of these functions can be selected to resemble the Pocock or OBF type of boundaries described in Sections 2.7.5.2 and 2.7.5.3. A review presented in Chapter 6 (Section 6.4.4.3) shows that the LD spending functions equivalent to OBF type boundaries are the most commonly used in practice.

2.7.5.6 Beta Spending Functions

Pampalloma et al (2001) highlight that most of the stopping boundaries described so far, including α spending functions focus on controlling α, with little attention on β. The authors present an approach to control both α and β by fitting continuous functions or curves to the cumulative errors spent under the type I or II error between two successive equally spaced interim analyses. The extrapolated curves guarantee the overall errors at the scheduled end are as pre-planned and are applicable for any arbitrary number of interim analyses. However, the authors acknowledge that this could result in overpowered or underpowered trials for fixed equally spaced interim analyses. The authors suggest an adjustment to the maximum information to mitigate this problem.

The authors extended the work by Lan and DeMets (1983) on α spending functions such that the efficacy and futility boundaries are computed to guarantee α and β respectively. The generalised procedure with efficacy and futility boundaries is summarised to satisfy the following:

At the first interim analysis ($j = 1$);

$$\begin{aligned}\beta(t_1) &= P_1|(Z(t_1)| \leq l_1),\\ \alpha(t_1) &= P_0|(Z(t_1)| \geq u_1),\end{aligned} \tag{2:29}$$

At any subsequent interim analysis ($\forall j \geq 2$);

$$P_0\left(l_1 < |Z(t_1)| < u_1, l_{j-1} < \left|Z\left(t_{j-1}\right)\right| < u_{j-1}, \ldots, \left|Z\left(t_j\right)\right| \geq u_j\right) = \alpha\left(t_j\right) - \alpha(t_{j-1}),$$
$$P_1\left(l_1 < |Z(t_1)| < u_1, l_{j-1} < \left|Z\left(t_{j-1}\right)\right| < u_{j-1}, \ldots, \left|Z\left(t_j\right)\right| \leq l_j\right) = \beta\left(t_j\right) - \beta(t_{j-1}). \quad 2:30$$

A constraint $l_k = u_k$ is put at the scheduled end. The functions $\alpha(t)$ and $\beta(t)$ are chosen at the design stage to suit some desired stopping criteria under H_0 and H_1. The overall type I and II errors are preserved using equations (2:28) and (2:31), respectively.

$$\beta(t_1) + \sum_{j=2}^{k} \left(\beta\left(t_j\right) - \beta\left(t_{j-1}\right)\right) = \beta(t_k) = \beta\,. \quad 2:31$$

In cases of any changes to the timing and frequency of planned interim monitoring, the future boundaries are recalculated after every interim analysis.

2.7.6 The Choice of Stopping Boundaries

When designing a group sequential trial, which stopping rules to use is one of the most pertinent questions faced by Clinical Trialists. Sebille and Bellissant (2000) study the statistical properties of the triangular test and PT approximations to Pocock and OBF boundaries through simulations. The authors looked at type I and II error control, expected sample size, and 90th percentile of the sample size required to reach a conclusion. Jennison and Turnbull (2000b) describe the contrasting statistical properties of various stopping boundaries through numerical integration, focusing on expected sample size under H_0, H_1 and $H_{0.5}$ (half-way between H_0 and H_1) for one and two-sided tests. Kim and DeMets (1987) investigate the behaviour of α spending function approximations to Pocock and OBF, and a subset of ρ family power functions through numerical integration, with respect to their expected sample size, expected stopping times, and conservatism. All these authors and Emerson *et al* (2007) highlight that the stopping boundaries differ mainly with respect to:

a) Flexibility to alter the number and timing of the interim analyses,

b) Conservatism or liberalism for early stopping decision-making,

c) Expected maximum sample size assuming the trial proceeds to the planned end,

d) Expected sample size under H_0, H_1 and $H_{0.5}$,

e) Expected stopping probabilities at different interim analyses,

f) Appropriateness of the boundary for given study objectives such as efficacy and/or futility.

As highlighted in Section 2.7.5.2, conservative stopping boundaries earlier in the trial are recommended to mitigate the risk of premature early stopping due to the high degree of uncertainty in the effectiveness of the intervention under investigation. Pocock (2006) states that any stopping boundaries for benefit need to be sufficiently tough to relate well to the public health implications of the decision to stop the trial early. Literature including Pocock strongly advised against the use of Pocock boundary to guide early stopping decisions for efficacy due to its less conservative nature (Kim and Demets, 1987; Pocock, 2006; Proschan et al., 2006b).

In practice, several authors recommend the use of α spending functions to allow for some flexibility in the timing and frequency of interim analyses due to unforeseeable circumstances, although the expected number is specified in advance to enhance planning (Ellenberg et al., 2003; Kim and Demets, 1987; Lan and DeMets, 2009; Proschan et al., 2006b). Furthermore, Kim and DeMets (1987) recommend boundaries where the amount of α being spent between interims is steadily increasing with the information fraction such that the resultant boundaries on the Z score scale are steadily decreasing. The authors underscore that the Pocock boundary and its variants do not satisfy these properties. Lan and DeMets (1983) argue that spending functions which approximate the OBF may be a suitable choice when long term intervention effect is the trial objective. DeMets and Lan (1994) emphasise that it is not permissible to change the spending function during the course of the trial. The authors argue that doing so means that there is no longer control over the type I error – hence jeopardising trial credibility.

Lan and DeMets (2009) advise that in addition to other aspects, the research question, its objectives, and the motivation behind the use of a GSD should guide in the choice between symmetric and asymmetric stopping boundaries. For instance, DeMets and Ware (1980) describe a one-sided test for non-inferiority. In such a case, Clinical Trialists may be willing to claim inferiority ('accepting' H_0) with less evidence and otherwise to claim non-inferiority under H_1. Thus, an asymmetric stopping boundary with less and more conservative futility and non-inferiority spending functions respectively, may be desirable.

Several authors highlight the difficulties faced in defining the 'optimality' of a stopping boundary, however, optimality dimensions such as those based on expected sample size and expected sample size saving under H_0 and H_1 assuming the trial is stopped early may be desirable (Hwang et al., 1990; Jennison and Turnbull, 2000a; Kim and Demets, 1987; Lan and DeMets, 1983; Wason, 2015). Hwang et al (1990) highlight that minimisation of the expected stopping times may contradict the fact that Clinical Trialists often avoid premature early stopping, especially when long term efficacy is required.

2.7.7 Impact of Altering the Number and Timing of the Interim Analyses

Whitehead (2000) articulates the fact that even though it is good practice to pre-plan the scheduling of interim analyses, the actual interim analyses may not coincide with those planned due to the practical realities faced in conducting trials, such as uncertainty around expected recruitment. More so, the IDMC may advise alteration of the frequency due to emerging information thereby overruling the initial planned number and/or frequency of interim analyses (Lan and DeMets, 2009). The spending functions approach alleviates this problem. However, DeMets and Lan (1994) state that the main concern is whether spending functions could be abused if the changes are in response to emerging trends. Some authors investigated this through simulation over a range of intervention effects and type I errors using the OBF and Pocock α spending boundaries (DeMets and Lan, 1994; Lan and DeMets, 1989). The authors employ a rule in which the frequency of the interim analyses would be doubled if the emerging trends are within 80% of the critical value (data driven decision-making) and compared the expected and observed type I error. They conclude that the type I error inflation is negligible, although there are noticeable discrepancies in the average stopping times.

2.7.8 Defining the Interim Information Fraction

Todd et al (2001) state that the timing of interim analysis can be directly measured in terms of the number of participants or the information. DeMets and Lan (1994) highlight that the exact amount of information contributed by each participant depends on the nature of the primary endpoint. For continuous and binary endpoints, a number of authors define the information fraction used in the computation of the stopping boundaries c_j's as $t_j = \frac{j}{k}$ (Pocock 1977; Jennison *et al*. 2000b; Proschan *et al*. 2006b). However, an accurate generalised expression of the information fraction depending on interim data relative to the pre-planned sample size per group is given by equation (2:10) (Lachin, 2005; Lan and Zucker, 1993b). Mehta and Tsiatis (2001) also defined the information fraction by

$$t_j = \frac{I(t_j)}{I_{max}}, \qquad 2:32$$

where $I(t_j)$ and I_{max} represent the interim and maximum Fisher's information at the planned end, which is defined for any outcome.

In summary, the expression of the interim information fraction based on Fishers information $I(t_j)$ is the most desirable and applicable regardless of the outcome, particularly when there is little information to inform the design. The approach enables correction of stopping boundaries when the expected interim information fraction does not match the planned information fraction. For example, this happens when assumptions on the expected nuisance parameters such as pooled variance are inaccurate. This is described in Section 2.8 and its application is illustrated in Section 7.5.4 of Chapter 7.

2.7.9 The Number and Timing of Interim Analyses

In choosing the number of interim analyses, Jennison and Turnbull (1989) recommend balancing the benefits of more frequent data examination against the effort required to perform each additional analysis. Pocock (1983) indicates little statistical gain for having too many interim analyses in terms of the expected sample size under H_1 for small to moderate effect sizes often observed in practice. Pocock argues for a sensible general rule to plan a maximum of five interim analyses. Similarly, Pampallona et al (2001) states that clinical trials are rarely designed with more than five interim analyses. A review presented in Chapter 6 Section 6.4.4.4 summarises the number of planned interim analyses observed in practice.

Regarding the timing, Kim and DeMets (1987) demonstrate that the time patterns influence the stopping boundaries. Pinheiro and DeMets (1997) demonstrate through simulation over a wide range of effect sizes that delaying the 1st interim analysis offers protection against bias in the intervention effect estimate. This supports the argument against conducting interim analyses very early on during the trial with a very small amount of information, as this is associated with huge uncertainty (Ellenberg et al., 2003).

In summary, although reviewed literature recommends the delaying of the 1st interim analysis, there does not appear to be a general 'rule of thumb' in terms of the information fraction to guide appropriate timing. A review presented in Section 6.4.4.1 of Chapter 6 shows summary statistics of the interim information fraction where trials are most likely to be stopped early.

2.7.10 Sample Size Estimation

The sample size is estimated to satisfy α and β given by equations (2:23) and (2:33) for a given effect size sought under H_1, other outcome dependent nuisance design parameters, and stopping boundaries c_j's.

$$P_{\theta_\delta}\left(\bigcap_{j=1}^{k}|Z(t_j)| < c_j\right) = \beta\,. \qquad 2:33$$

For a design with k interim analyses, there are $3k-1$ unknown parameters $\{n_j, l_j, u_j\}$ with a constraint $(l_k = u_k)$; where n_j is the group sample size at interim t_j and sets $\{l_j, u_j\}$ are defined by the stopping rules chosen. Jennison and Turnbull (2000b) point out that in practice, the information fraction at each interim analysis does not need to be equal and it is difficult to guarantee this condition. In such situations, additional constraints are imposed such that $n_j = r_j n_1$ for $j \geq 2$ and $r_1 = 1$; where r_j is the desired ratio of t_1 to the subsequent t_j information fraction. Recursive numerical integration is then employed to compute exact solutions of the unknowns n_1 and $\{l_j, u_j\}$.

2.7.11 Impact on Statistical Inference

The use of conventional inference methods often employed for a fixed sample size design with one analysis at the scheduled end is invalid for a group sequential trial (Armitage et al., 1969; Jennison and Turnbull, 1989, 2000a; Kim and DeMets, 1987; Kim, 1989; Siegmund, 1978; Todd et al., 2001; Tsiatis et al., 1984; Whitehead and Jones, 1979). To aid early stopping decision-making, Jennison and Turnbull (2000a) state that the most pertinent questions for Clinical Trialists regarding statistical inference are on how to estimate the:

1) Point estimate of the overall treatment effect with its associated CI and p-value regardless of the timing of early stopping,

2) CIs of the interim treatment estimates to aid decision-making by the IDMC.

2.7.11.1 The Concept of Ordering the Sample Space of Interim Results

Jennison and Turnbull (2000a) highlight the nonexistence of a unique approach to ordering the sample space of group sequential results defined by sets $\{t_j, Z(t_j)\}$ for $Z(t_j) \notin \{(l_j, u_j)\}$ with early stopping at $\{\tau, Z(\tau)\}$. Proschan et al (2006) illustrate this phenomenon with an example showing conflicting hierarchy of evidence between the likelihood ratio under a number of alternatives and magnitude of the Z score at different interims. Hence, some form of rule guiding the ordering of the sample space is required in the subsequent computation of p-values, CIs, and median unbiased point estimates.

Armitage (1957) employs the idea of partial MLE ordering where the sample space is ordered solely based on the magnitude of the interim MLE of θ. Emerson and Fleming (1990) used this idea in the context of

continuous outcomes using the term 'sample mean ordering'. Armitage (1958) suggests the idea of stagewise ordering, which was subsequently adopted by a number of authors in different settings (Fairbanks and Madsen, 1982; Jennison and Turnbull, 2000d; Siegmund, 1978; Tsiatis et al., 1984). This sample space ordering depends on the boundary crossed (l_τ or u_τ), the interim stage at early stopping and its associated test statistic $\{\tau, Z(\tau)\}$, assuming interval continuation regions. In addition, the earlier stopping times provide more compelling evidence and when comparing two trials stopping early at the same interim, the larger the Z score is the more extreme the results (Proschan et al. 2006a).

The stagewise ordering is invalid for some designs with non-interval continuation regions (Emerson and Fleming, 1990; Jennison and Turnbull, 2000d). In addition, in some cases the resulting exact CIs based on the stagewise ordering may fail to include the MLE $\hat{\theta}$, hence it is difficult to interpret the results (Emerson and Fleming, 1990). Noting these drawbacks, Chang (1989) presents a likelihood ratio (LR) ordering depending on the LR test, which quantifies how extreme a particular value of the observed statistic is from the hypothesised value. Rosner and Tsiatis (1988) present a partial score test ordering approach evaluated from the score function, which gives a measure of distance of the interim results from the hypothesised value evaluated at the value under H_0.

Proschan et al (2006) underline that MLE, score test and LR orderings depend on the data observed thus far and the unobserved future data after termination, which may be problematic when the number and timing of future interims are unpredictable. In contrast, the stagewise ordering obviates this problem by only using the information prior to early stopping and that it is a linear ordering whereas others are not. In summary, for methods that utilise ordering of the sample space to compute median unbiased point estimates, adjusted CIs and adjusted p-values, reviewed literature recommends the use of the stagewise ordering approach because of its appealing properties compared to the other orderings.

2.7.11.2 Estimation of p-values Following Early Stopping

The fact that unadjusted interim monitoring causes type I error inflation implies that the unadjusted p–value tends to overestimate the evidence against H_0 (Proschan et al., 2006d). For a fixed sample size design, a p-value is estimated by computing the probability of observing at 'least extreme' results than the one observed when H_0 is true. Since data are analysed only once at the end of the trial, there is a single expected critical value to define a region of results viewed as 'least extreme'. Consequently, the ordering of this classification is unique. The same principle can be extended to calculate p-values for a GSD to test H_0 on realisation of $(\tau, Z(\tau))$ (Jennison

and Turnbull, 2000d; Proschan et al., 2006d). However, as highlighted by these authors, interim analyses of data mean that there are many critical values to consider. Hence, there are many possible (non-unique) permutations to define regions viewed as 'least extreme' on realisation of $(\tau, Z(\tau))$. This means that the classification of all the possible 'least extreme' results is complex and requires some form of ordering of the outcome sample space as described in Section 2.7.11.

Proschan et al (2006) illustrate that a two-sided p-value under any form of ordering being considered is computed by $p = 2 \times minimum(p_L, p_U)$; where $p_L$and p_U are the lower and upper p-values or cumulative crossing probabilities. For instance, for symmetric stopping boundaries using stagewise ordering, a two-sided p-value is calculated as cumulative exit probabilities using stopping boundaries $c_{i-1}; \forall\, i < \tau$ and $Z(\tau)$ where $\tau = j \leq k$ with 'acceptance' region defined by equation (2:37).

$$p = P_0\left(\bigcup_{j=1}^{i-1} |Z(t_j)| \geq c_j \bigcup |Z(\tau)| \geq z_\tau\right). \qquad 2:34$$

The authors highlight four essential desired properties of p-values: uniformly distributed, consistency with the stopping boundaries, independency on the number and timing of future interims, and stopping at the 1st interim look result in a p-value similar to that from the fixed sample size design. The authors underline that the stagewise ordering satisfies all these properties whereas the other forms of orderings satisfy only the uniformly distributed property. Hence, the stagewise ordering is the most appealing and preferred method in practice.

2.7.11.3 Estimation of the Intervention Effect

Pocock and Hughes (1989) reiterate that Clinical Trialists should place greater emphasis on estimation of the intervention effect rather than simply dwelling on statistical significance. Hughes and Pocock (1988) illustrate that multiple significance testing can introduce bias in the point estimate of the intervention effect. Emerson and Fleming (1990) state that failure to adjust for the interim analyses introduces bias in much the same way that the repeated use of single sample hypothesis testing inflates the type I error. The authors point out that trials are stopped early because extreme results have been observed, thus it is expected that the traditional MLEs be biased upwards.

Bassler et al (2010) systematically review trials that stopped early and those which did not – where both sets of trials addressed the same research questions. The authors conclude that truncated trials were associated with larger effect sizes than the latter. Several authors also highlight that trials that are stopped early for clinical benefit tend to overestimate or exaggerate the magnitude of the true intervention effect (Bassler et al., 2008;

Ferebee, 1983; Hughes et al., 1992; Jennison and Turnbull, 2000d; Liu and Hall, 1999; Pocock and Hughes, 1989; Whitehead, 1986; Zhang et al., 2012).

Proschan et al (2006) state that at early stopping, $\{\tau, Z(\tau)\}$ is a sufficient statistic – thus a search for estimators regarding θ should be confined to such a sufficient statistic, and the MLE ($\hat{\theta}$) is a subset of such estimators. Jennison and Turnbull (2000a) illustrate that the sampling density of $\hat{\theta}$ is not Normally distributed as would be the case for a fixed sample size design, but instead is multi-modal truncated Normally distributed with peaks at each interim analysis. Hence, the MLE ($\hat{\theta}$) is a biased estimator of θ. Emerson (1993) states that following a group sequential test, $\hat{\theta}$ is neither unbiased nor has minimum variance. Appendix 2.4 provides a review of estimators proposed in the literature and their properties. These include the Whitehead (1986) bias-adjusted estimator, Ferebee unbiased estimators (Ferebee, 1983; Liu and Hall, 1999), median unbiased estimators (Emerson and Fleming, 1990; Kim, 1989), conditional unbiased or bias-adjusted estimators (Emerson and Kittelson, 1997; Emerson, 1993; Fan et al., 2004; Milanzi et al., 2014; Proschan et al., 2006d; Troendle and Yu, 2010), and a bias reduction estimator using a parametric bootstrap (Wang and Leung, 1997).

In summary, reviewed literature shows that a number of estimators have been proposed and studied. There does not appear to be a unique estimator which is suitable under all conditions. Despite this, the Whitehead bias-adjusted ($\tilde{\theta}_{UWT}$) and median unbiased estimator based on the stagewise ordering appear to be preferred with desirable properties when compared to others. However, the application of the stagewise ordering is debatable for some designs with non-interval continuation regions such as the triangular test with inner wedges. In Chapter 6, a review of the application of methods used in practice to conduct inference following a group sequential test is undertaken.

2.7.11.4 Estimation of Confidence Intervals Following Early Stopping

For sequential monitoring in different settings, Armitage (1958) and Siegmund (1978) illustrate numerically and using examples that the use of naïve fixed sample size approaches to estimate CIs is invalid. Similarly, Tsiatis et al (1984) demonstrate that the exact coverage probabilities for the 90% CIs following a 5 interim GSD with Pocock or OBF boundaries using a naïve approach produce coverage probabilities range of 84.6% to 93.0%, depending on the effect sizes considered. Emerson (1993) highlights that, in a similar manner to point estimation, statistical inference based on naïve methods when applied to group sequential data produce CIs with incorrect coverage. Todd et al (1996) underline that the problem is because of the non-Normal nature of the sampling distribution of $\hat{\theta}$. The authors highlight that even the CIs constructed using the Whitehead biased-

adjusted estimator ($\tilde{\theta}_{UWT}$) with its variance may yield inaccurate coverage probabilities. Hence, the is a need for a different approach to the computation of CIs with correct coverage probabilities.

Jennison and Turnbull (2000a) underline that CIs following a group sequential test should:

a) Be guaranteed if the monotonicity assumption of the sample space holds;

b) Contain the MLE ($\hat{\theta}$);

c) Be in agreement with the original test considered and narrower CIs are preferred;

d) Be well defined irrespective of the predictability of the information fraction.

In addition to satisfying conditions a), c), and d); CIs based on the stagewise ordering do not depend on the unobserved information beyond the early stopping interim analysis, but sometimes may not satisfy condition b) (Proschan et al., 2006d; Rosner and Tsiatis, 1988).

Todd et al (1996) adopt the idea suggested by Woodroofe (1992) to modify the test statistic at early stopping $Z(\tau)$, such that its distribution is roughly standard Normal, to construct improved CIs. This approach is independent of any form of sample space ordering but the CI is computed using the MLE ($\hat{\theta}$) and the modified $Z(\tau)$ test statistic. The authors provide modified algebraic expressions to improve the computation of the modified $Z(\tau)$. Through simulation, the authors conclude that the approach achieves satisfactory coverage probabilities under the two scenarios of a triangular test and OBF boundaries. One limitation of the approach is that it uses a biased MLE ($\hat{\theta}$) and not the bias-adjusted MLE, although the authors discuss it as an alternative approach (Li and DeMets, 1999).

Some authors describe an approach to the construction of CIs for a two-sided test $H_0: \theta = \theta_0$, where $\theta_0 = 0$ (Jennison and Turnbull, 2000d; Rosner and Tsiatis, 1988; Tsiatis et al., 1984). The CI about θ is obtained by constructing 'acceptance' regions of the sample space under H_0 $\{A(\theta_0 = 0)\}$, such that $\big(\tau, Z(\tau)\big) \in A(\theta_0 = 0)$. Rosner and Tsiatis (1988) highlight the question of the appropriate choice of $A(0)$ since 'acceptance' regions cannot be formed using a uniformly powerful test because of the non-existence of a monotone LR in the sample space. Therefore, there is a need for some form of sample space ordering on realisation of $(\tau, Z(\tau))$ to define the 'acceptance' region $A(0)$ as described in Section 2.7.11.

Tsiatis et al (1984) present an approach to compute CIs using numerical integration to yield exact coverage probabilities using the stagewise ordering. The resultant CIs are not generally symmetric around the estimate of θ. The approach utilises both the stopping time and the value of its statistic. More so, the approach

depends only on the information prior to early stopping. When early stopping happens at the first interim, the exact CI generated is the same as that using the naïve approach because of unique sample space ordering. Rosner and Tsiatis (1988) adopts a similar idea, but using score test ordering based on constant and standardised distance from mean ordering. The authors evaluated the performance of these methods with respect to the probability of covering the wrong mean compared to the naïve fixed design approach and stagewise ordering. The authors recommend the stagewise ordering as appropriate when the timing and frequency of interims are unpredictable – often the case for α spending functions.

In summary, reviewed literature recommends CIs based on the stagewise ordering as a preferred method in practice citing that they have desired properties and it is the only method available for unpredicted information (Jennison and Turnbull, 2000d; Proschan et al., 2006d; Rosner and Tsiatis, 1988; Wittes, 2012). For non-interval continuation regions where the application of the stagewise ordering is questionable, modified versions of test statistics with Woodroofe's procedure suggested by Todd et al (1996) may be applicable.

2.7.11.5 Estimation of CIs Using Sample Space Ordering

A number of authors provide a framework to compute CIs following a group sequential test using the concept of sample space ordering (Jennison and Turnbull, 2000a; Proschan et al., 2006d). For a two-sided group sequential test, a $100(1-\alpha)\%$ CI of θ given by (θ_L, θ_U) is constructed by solving

$$P_{Z_L}\{A(0)\} = \frac{\alpha}{2}, \quad P_{Z_U}\{A(0)\} = 1 - \frac{\alpha}{2}, \tag{2:35}$$

where $Z_L$and Z_U are the lower and upper confidence limits on the Z score scale obtained through numerical integration or grid search of values to satisfy these equations such that

$$P_\theta(Z_L < Z_\theta < Z_U) = 1 - \alpha\,. \tag{2:36}$$

The confidence limits on the intervention effect scale are then obtained by rearranging equation (2:11). For a two-sided symmetric test using a stagewise ordering, the 'acceptance' region with stopping boundaries $c_{i-1}; \forall\, i < \tau$ and $Z(\tau)$, where $\tau = j \leq k$ is given by

$$A(0) = \left(\bigcup_{j=1}^{i-1} |Z(t_j)| \geq c_j \bigcup |Z(\tau)| \geq z_\tau \right). \tag{2:37}$$

The median unbiased estimator of θ ($\hat{\theta}_{MUE}$) is obtained by finding the solutions to

$$P_{z_\theta}\{A(0)\} = 0.5\,. \tag{2:38}$$

2.7.11.6 Computation of Repeated Confidence Intervals at Interim Analyses

Jennison and Turnbull (1989) state that it is desirable for the IDMC to have some information about the coverage probability of the true intervention effect θ at interim analyses on realisation of $\hat{\theta}(t_j)$ to aid early stopping decision-making. Jennison and Turnbull (1989, 2000d) describe the concept of repeated confidence intervals (RCIs) at interim analyses. The authors state that if $RCI(t_j); j = 1, \dots, k$ defines RCIs, then the following should hold for θ;

$$P_\theta(\theta \in \{RCI(t_1), \dots, RCI(t_k)\}) = 1 - \alpha\,. \qquad 2:39$$

The authors highlight and illustrate that when a trial is stopped early at some information fraction $\tau \in t_j < k$, then the probability that the sequence of RCIs $\{RCI(t_1), \dots, RCI(\tau)\}$ contains θ is at least $1 - \alpha$ for all values of θ indicating their conservative nature at earlier interims. In order to satisfy equation (2:39), the RCIs at interim t_j are computed using their associated critical values $c_j; \forall j$ given by

$$RCI(t_j) = \left\{\hat{\theta}(t_j) - c_j\sqrt{I(t_j)}, \hat{\theta}(t_j) + c_j\sqrt{I(t_j)}\right\}. \qquad 2:40$$

Mehta et al (2007) extend the concept of RCIs beyond a standard GSD. Whitehead (2000) illustrates a strange statistical property of RCIs which may be unattractive, although it is unlikely to occur in practice. This is realised when the maximum and minimum limits of the series of lower and upper RCI limits fail to contain the true intervention effect θ.

2.7.12 Statistical Software for Implementation

There are a number of commercial and open access statistical packages offering various functionalities in the design, interim monitoring, and analysis of group sequential trials. East (Cytel, 2015) and ADDPLAN (ICON, 2015) are commercial software offering a wide range of stopping rules options to plan a GSD. Importantly, the packages enable researchers to compare statistical properties of competing designs with respect to stopping probabilities and expected sample sizes under H_0 , H_1, and $H_{0.5}$. The estimates of intervention effects, CIs and p-values are median unbiased based on the stagewise ordering. In addition, RCIs are calculated at interim analyses to aid decision-making. One limitation is that the packages do not offer the implementation of a triangular test (Whitehead and Stratton, 1983) and bias-adjusted related estimates (Todd et al., 1996; Whitehead, 1986).

There are a number of open access user-written R statistical packages:

a) The '*RCTdesign'* package provides a comprehensive suite of functions for evaluating, monitoring, analysing, and reporting group sequential and adaptive clinical trial designs (Emerson, 2014; Gillen and Emerson, 2011). The package allows for more options to compute point estimates: Whitehead bias-adjusted MLE; median unbiased estimates based on the stagewise, and the MLE orderings; Rao-Blackwell conditional unbiased estimator. Exact CIs and p-values (Brannath et al., 2009; Gao et al., 2013) are also calculated;

b) The package '*gsDesign*' derives group sequential designs and describes their properties (Anderson, 2015);

c) The package '*AGSDest'* allows the computation of median unbiased estimates (intervention effect, CI, and p-value) and RCIs based on the stagewise ordering.

In SAS 9.2 onwards, the '*SEQDESIGN*' procedure creates GSDs by computing for a number of methods, including Whitehead triangular tests and estimates of the required sample sizes. The '*Proc SEQTEST'* procedure offers a number of options for interim monitoring for decision-making. In cases where the information fraction for the test statistics does not match the planned information fraction using '*SEQDESIGN*', the '*SEQTEST*' procedure modifies the original boundary values to adjust for the observed information fraction (Yuan, 2009). In addition, the procedure computes median unbiased estimates (intervention effect, CI, and p-value) during analysis after trial stopping based on the stagewise, LR or MLE orderings.

2.7.13 Practical, Logistical and Administrative Aspects

Practical challenges when implementing GSDs from planning to completion are discussed by many authors (Gallo, 2006; Gluud et al., 2008; Pocock, 1983; Quinlan and Krams, 2006; Vandemeulebroecke, 2008). Pocock (1983) underlines that interim analyses are liable to being of purely academic interest if a trial is relatively small. That is, value for money and patient benefits are proportional to trial size, which should be weighed against additional complexities in implementation. Pocock highlights important factors influencing the feasibility of interim analyses. These include the time lag between participant enrolment and observing the primary endpoint relative to the expected recruitment rate and treatment duration. For example, interim analyses do not have potential to reduce sample size for trials with long term endpoints relative to predicted recruitment rate. In such cases, additional participants awaiting follow-up would have been recruited by the time of the planned interim analyses. Nevertheless, interim analyses can still be conducted to expedite decision-making and reduce trial duration. The worst case scenario is when target recruitment is met by the time of interim analyses. One approach

is to control recruitment to minimise participants with delayed responses. Vandemeulebroecke (2008) questions this idea of controlling recruitment in practice. Hampson and Jennison (2013) provide a comprehensive review of statistical methods to deal with delayed responses, referred to as 'pipeline participants'. The authors present an improved method which gives a proper handling of 'pipeline' participants after the decision to stop the trial. Whitehead (2000) discusses situations when the inclusion of 'pipeline' participants is recommended. The interpretation of results when an inconsistent decision is reached after the inclusion of 'pipeline' participants is however unclear.

Implementation of interim analyses requires additional systems, procedures and processes to enhance compliance to the study protocol, processing of data and communication to expedite decision-making (Gaydos et al., 2009; Pocock, 1983; Vandemeulebroecke, 2008). The authors highlight the need for data management to enable quick and robust capturing, cleaning, processing and evaluation of the data across participating sites. In addition, the training of personnel involved in the design, conduct and analysis, especially from the statistical perspective for simulation modelling to understand the statistical and operational properties of the design under different scenarios may be needed (O'Neill, 2006; Schäfer, 2006; Vandemeulebroecke, 2008).

2.7.14 Interim Decision-Making Challenges

In practice, the IDMC is often tasked with decisions to recommend early trial stopping based on information from various sources, which can be external to the trial, rather than the statistical stopping rule alone (DeMets et al., 1984; Ellenberg et al., 2003; Lan and DeMets, 2009; Pocock, 2006). Pocock (2006) argues that stopping rules are often used as guidelines for the IDMC to aid interim decision-making. Thus, the IDMC may choose to overrule statistical stopping rules for some reasons even if the boundary is crossed.

Lan and DeMets (2009) give a comprehensive discussion on practical aspects during interim monitoring using α spending functions. The authors question the implications of the IDMC's decision to overrule statistical boundaries on the future stopping boundaries and inference. For example, when an efficacy boundary is crossed and the IDMC decides to continue the trial, does it imply that Trialists need to compute new future boundaries or stick with the existing plan? What are the consequences on the statistical properties of the design? A GSD as described so far has 'binding' stopping boundaries meaning that overruling these boundaries means that the planned design may no longer guarantee protection of the type I error (Lan and DeMets, 2009; Lan et al., 2003).

Lan et al (2003) investigated two approaches to resetting the future boundaries after overruling for the Pocock and OBF α spending functions. The authors use a fixed sample size critical value or 'buy-back' method by recalculating future boundaries based on cumulative α spent to the overruling point and assuming no interim analyses have happened. The authors investigate their performance with respect to the chance of reaching an opposite conclusion and inflation to the errors through simulations for limited scenarios on the number of looks. The authors conclude that using a fixed sample size critical value for future looks is a simple approach to handle overruling and its performance is similar to the 'buy-back' method.

Jennison and Turnbull (2011) describe a modified approach which sets boundaries such as futility as 'non-binding' and guidelines. That is, 'non-binding' futility boundaries can be overruled but with guaranteed protection on the type I error. However, as the authors point out, the impact is an increase in the efficacy boundary and a small reduction in power. The authors highlight that the approach is applicable to either efficacy or futility or both, but the later results in a slight reduction in the type I error.

In summary, designing trials with 'non-binding' boundaries, particularly for futility appears to offer protection in case of overruling of statistical boundaries (Gallo et al., 2014). The use of fixed sample size critical values for future boundaries following overruling is an alternative pragmatic and simple solution for trials designed without 'non-binding' boundaries. East and ADDPLAN software packages offer options to formulate a GSD with 'non-binding' for futility, but not for efficacy boundaries.

2.7.15 Preserving Trial Integrity

Vandemeulebroecke (2008) underlines that trial integrity relates to whether trial results are credible, interpretable, and persuasive in clinical practice. The need to preserve trial integrity and scientific validity is well acknowledged to be pivotal in the confirmatory setting in order to provide convincing trial results to a broader research community, decision-makers and policymakers to change practice (Fleming et al., 2008). Dragalin (2006) highlights some dimensions of trial integrity which encompass preplanning, adaptations as intended, consistent protocol delivery, and measures to minimise operational bias during the conduct of the trial.

The challenges and issues encountered during monitoring of group sequential trials have been articulated and some recommendations proposed (Fleming et al. 1993; Ellenberg et al. 2003a). The need for the involvement of an IDMC with their roles and duties contained in a written charter, maintenance of confidentiality of interim results and measures to minimise operational bias in the conduct of the trial due to the leaking or knowledge of

the interim results are highlighted as imperative. Herson et al (2012) detail exemplars of communication flowcharts involving key stakeholders to maintain confidentiality, minimise operational bias and preserve trial credibility. The key stakeholders include the Sponsor or Funder, data management and processing staff, investigators, and an IDMC. Gallo (2006) gives recommendations to preserve trial integrity and confidentiality.

2.7.16 Reflection

Group sequential methods described here are viewed by Regulators and researchers as well understood and methodology is well developed to control statistical properties. The approach offers ethical and financial benefits in the presence of overwhelming evidence. On the other hand, trials producing relatively small intervention effects are more likely to reach the expected end with an even larger sample size than for a fixed sample size design. In view of the poor 'success' rate of investigative interventions in confirmatory trials and a sizable proportion of trials requiring funding extensions, it would be beneficial to design most of the trials with futility stopping where possible.

Commercial and open access statistical software packages are available to facilitate the design, interim monitoring, and analysis of GSDs. A number of methods have been proposed to provide bias-adjusted or median unbiased results, although there is no unique optimum method suited for all circumstances. However, the use of these methods in practice is unclear. Chapter 6 reviews the practical application of the GSD and methods to conduct inference and highlight main findings.

Although the approach is appealing, a number of considerations should be given at the planning stage. The rationale behind the need for early trial stopping with respect to the available evidence base, impact of the results and potential benefits are important considerations. In addition, feasibility in implementing the design is paramount. In practice, realisation of the primary endpoint relative to the recruitment rate is often 'unpredictable' complicating the timing of the interim analyses and decision-making process. For instance, dealing with delayed responses of 'pipeline' participants at the time of interim analyses is not straightforward. Thus, the design is easier to execute for immediate and short to medium term primary endpoints than long term endpoints.

The decision-making process to stop an ongoing trial is a complex one involving various key stakeholders. For example, in addition to interim trial results, an IDMC often considers information from several sources when recommending early stopping decisions. Some of the information may be from external related

trials or safety signals. As a result, an IDMC may make a decision inconsistent with and overruling the planned stopping rules. The use of 'non-binding' stopping rules, especially for futility mitigates this problem.

In practice, trials are often designed with competing objectives or other important secondary outcomes. A number of authors present methods dealing with joint bivariate primary outcomes (Cook and Farewell, 1994; Jennison and Turnbull, 2000f; Todd, 2003). However, the methods are inaccessible in mainstream statistical software for implementation and methodological work is still required in this area. The problem is less pronounced when one outcome feeds into the other since inference can be extrapolated in such settings. In addition, it is unclear how to proceed with analysis of secondary outcomes, which do not depend on planned stopping boundaries when a trial is stopped early. In theory, Liu and Hall (2001) present modified estimators for continuous outcomes depending on the correlation between the primary and secondary outcomes.

The design described here had only one adaptation option – either to stop a trial for futility or efficacy. In addition, design parameters such as event rate in the control or SD of the primary endpoint were assumed to be accurate, which may not be the case. In practice, researchers may wish to safeguard against inappropriate design assumptions or to add other adaptations in a GSD such as SSR or investigating multiple arms with options to drop futile arms. Section 2.8 briefly introduces the concept of an information based GSD. The seamless and MAMS designs are introduced in Sections 2.9 and 2.10, respectively. A reflection on the theoretical framework underpinning more complex ADs is introduced in Appendix 2.5.

2.8 Design 4: Information Based Group Sequential Design

2.8.1 Motivation

The efficiency of the standard GSD described in Section 2.7 depends on the accuracy of the assumed nuisance design parameters such as event rate in the control or SD for a binary and continuous primary outcome, respectively. When these assumptions are inaccurate, a standard GSD may be underpowered or overpowered depending on whether the nuisance design parameters are underestimated or overestimated (Mehta and Tsiatis, 2001). The influence of inaccurate assumptions is demonstrated in Chapters 7 and 8 through simulation work using case studies. Thus, there is a need for a GSD that allows self-correction or SSR when assumptions about nuisance design parameters are inaccurate.

The concept of an information based GSD to improve statistical efficiency of the standard GSD when there is little information regarding nuisance parameters at the design stage has been described (Mehta and Tsiatis, 2001; Scharfstein and Tsiatis, 1998). Section 2.7.8 highlighted different ways to define information fraction. Here, a trial is monitored the using information fraction as defined by equation (2:32) rather than equation (2:10) as the fraction of the sample size. The interim information is approximated by

$$I(t_j) \approx \frac{1}{\left\{se\left(\hat{\theta}(t_j)\right)\right\}^2}. \qquad 2:41$$

Figure 2.5 is a flowchart summarising a simplified testing procedure for the proposed information based GSD for evaluating an investigative intervention against a comparator. Mehta and Tsiatis (2001) highlight that the information fraction is a difficult construct to conceptualise – thus it should be converted to a sample size scale, which is understood by the research team to enhance the timing of interim analyses. The authors provide a mathematical framework to calculate the maximum information fraction (I_{max}) which depends on the chosen stopping boundary, planned number of interim analyses, type I and II errors, type of primary outcome, and clinically relevant effect sought. The authors also describe how the information is converted to the sample size scale. It should be noted that the stopping boundary values are recalculated to mirror the observed interim information. Furthermore, at some interim point, a decision has to be made on whether to increase the sample size if the interim information fraction is lagging far behind the expected, that is required to maintain the desired power.

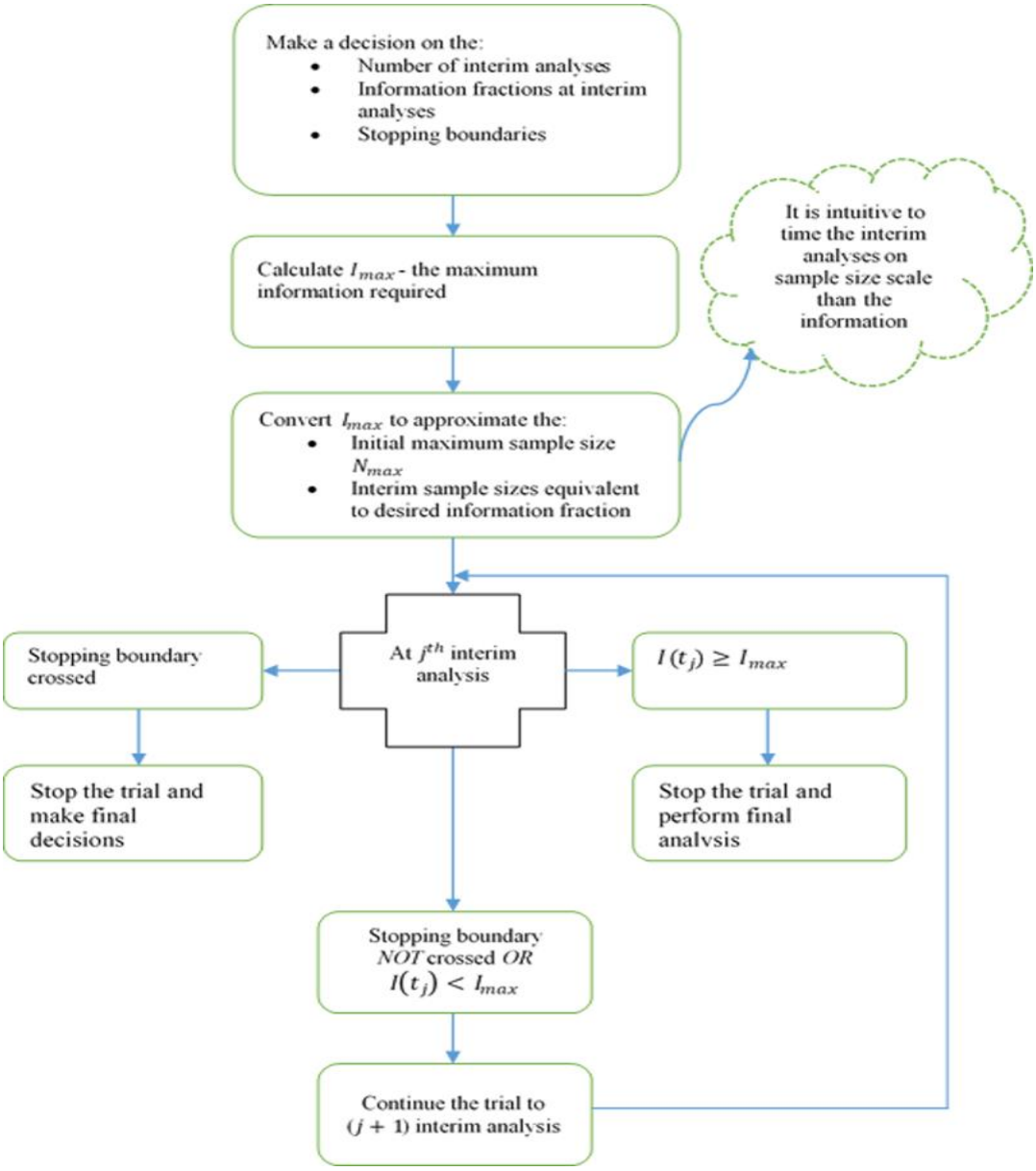

Figure 2.5. Flowchart for an information based group sequential design.

2.8.2 Reflection

An information based GSD is an alternative to the standard GSD described in Section 2.7 when there is little information to inform the design as the design self corrects to preserve the desired type I error and power. Its implementation requires additional statistical knowledge. East software (Cytel, 2015) supports the implementation of the design for a number of primary outcomes.

Although the design is appealing from a statistical perspective, it raises a number of operational challenges. The first hurdle is convincing the Public Funders about statistical efficiency of the design when

assumptions about the nuisance design parameters are inaccurate. Furthermore, the possibility of increasing the sample size should be communicated with the Funder beforehand for contingency planning. This could be done under a number of plausible scenarios regarding the least power desired to be preserved or tolerated uncertainty threshold. The second hurdle is the need to convince Reviewers in view of the prominence of the standard GSD. Some Reviewers may argue against the conduct of a confirmatory trial when there is huge uncertainty around nuisance design parameters in favour of a pilot trial. On the other hand, it could be argued that there are cases where research appears to suggest there is enough information to inform the design which turns out to be inaccurate, as illustrated in Chapter 7.

In practice, as shall be highlighted in Chapter 5, the PRIMO trial which was funded by the private sector implemented the design (Pritchett et al., 2011; Thadhani et al., 2012). However, it appears the design is not commonly applied.

2.9 Design 5: Seamless Design

2.9.1 Motivation

Traditionally, the evaluation of investigative interventions is conducted in a series of independent phases with time gaps between them and each phase has distinct key objective(s) (Bretz et al., 2006; Temple, 2000). The progression to the next phase depends on an earlier phase meeting the progression criteria set depending on the objective(s) of the phase. There are deficiencies in the conduct of standalone trial phases emanating from the fact that a lot of time is underutilised between phases (Bretz et al., 2006; Gould, 2006).

A number of approvals are required before the conduct of a trial depending on the nature of the study (Haynes et al., 2010). Such approvals are gained from RECs (Research Ethics Committees), Regulators such as MHRA, EMA, or FDA depending on the nature of the intervention, and local research governance bodies. Considerable variation in time spent setting up trials and seeking ethics and regulatory approvals is well acknowledged (Hackshaw et al., 2008; Haynes et al., 2010; Mallick and O'Callaghan, 2009). This research governance process is often time consuming and burdensome (Smajdor et al., 2009). Additional time is also spent on administrative aspects and protocol development (Chen, Gesser, et al., 2015). The seamless design has been proposed to eliminate the considerable time-gap between the end of one trial phase and the beginning of the next phase (Gould, 2006; Kairalla et al., 2012). Multiple phases are integrated into one trial and conducted under one

protocol and the main research governance approval – hence enabling uninterrupted continuation between phases (Bretz et al., 2006; Kairalla et al., 2012).

The seamless design can be classified into two broad categories: operational and inferential seamless (Berry, 2012; Cuffe et al., 2014; Maca et al., 2014; Zang and Lee, 2014). For operational seamless design, data generated from different phases are treated and analysed independently to address distinct objective(s) of individual trial phases. Appropriate statistical methods are then used to analyse distinct trial phase data. Whereas, for inferential seamless design, a fraction of the data used from earlier phase(s) contributes to the data used to answer objectives of proceeding phase. Consequently, this approach requires advanced statistical methods to weight and combine portions of data from participants whose data contribute to earlier phase(s) and those that are enrolled thereafter (Kairalla et al., 2012). For example, some authors present hypotheses test frameworks using combination test methods to be introduced in Appendix 2.5 (Bretz et al., 2006; Hommel, 2001) and other methods proposed elsewhere (Kimani et al., 2013; Stallard and Todd, 2003, 2011).

There are variations of the seamless design which are trial dependent. For instance, a seamless 2/3 design combines conventional phase 2 and phase 3 into one trial with the phase 2 addressing exploratory and the phase 3 confirmatory objectives. Yardley et al (2015) applied an operational seamless phase 2/3 design allowing for treatment selection. The authors compared two investigative interventions versus the control in phase 2 with an option to 'pick-the-winner' and test it against the control in phase 3 as illustrated in Figure 2.6. The sample sizes of phase 2 and 3, and respective analyses are district. Furthermore, the analysis of phase 2 should account for the multiple pairwise comparisons considered.

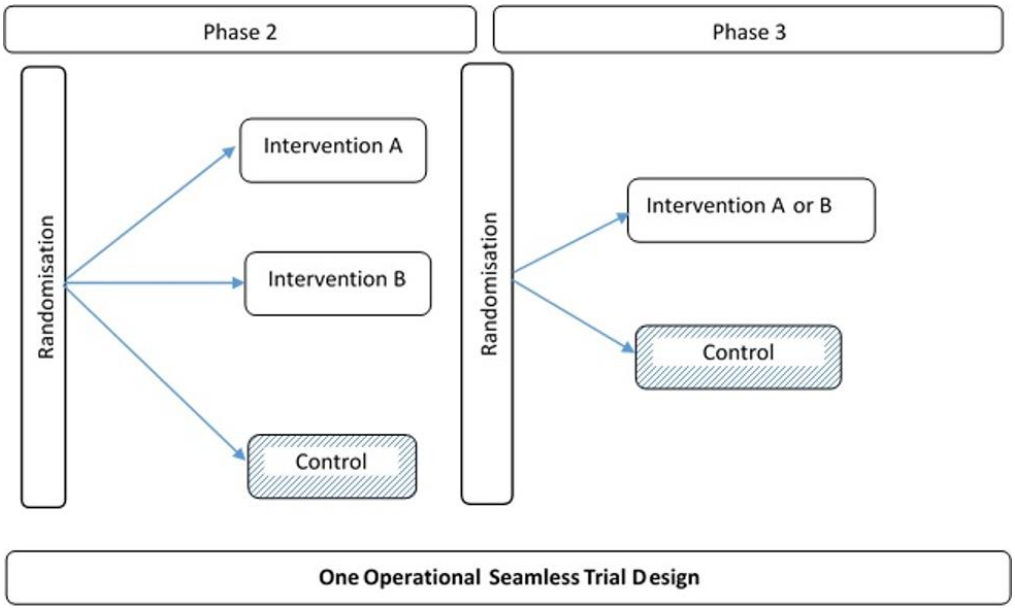

Figure 2.6. Example of an operationally seamless phase 2/3 design.

In contrast, INHANCE study employed an inferential seamless phase 2/3 design with multiple drug doses compared to a placebo and active control as displayed in Figure 2.7 (Cuffe et al., 2014). STAMEDE trial is a seamless design which went further to add new intervention arms during the course of the trial (Sydes et al., 2009a). However, the addition of new intervention arms to an ongoing confirmatory trial has methodological issues highlighted elsewhere and is beyond the scope of this thesis (Cohen et al., 2015).

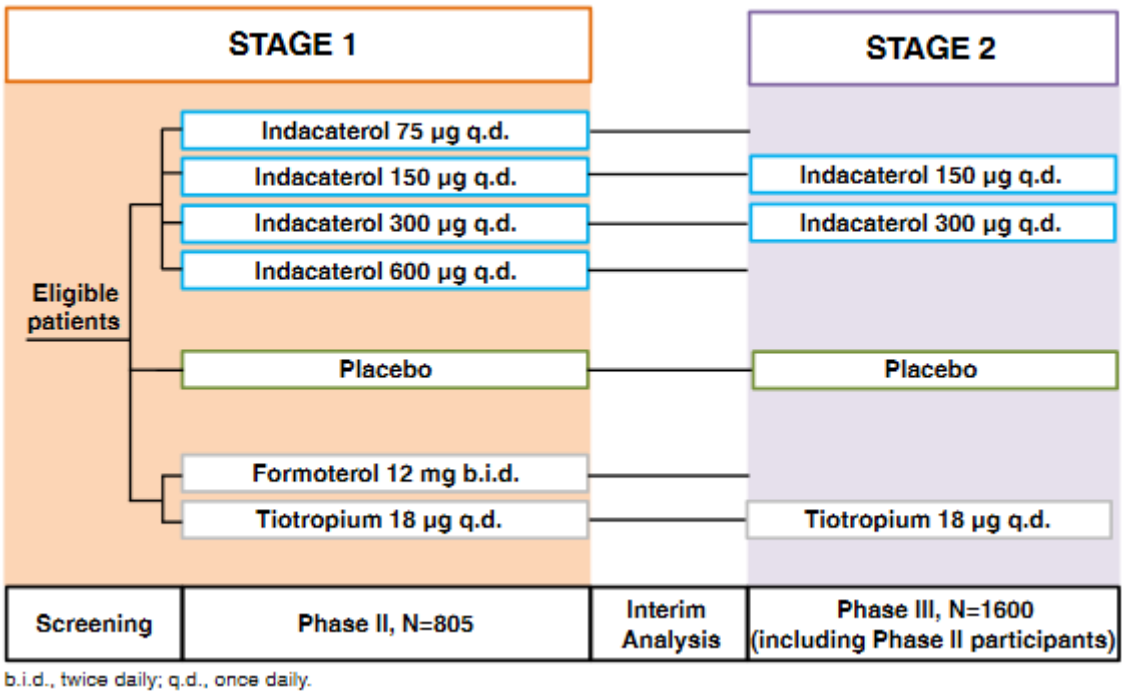

Figure 2.7. Inferential seamless phase 2/3 design.

Source: (Cuffe et al., 2014).

In Chapter 3, Section 3.4.4.3 highlights a degree of regulatory conservatism towards the use inferential seamless design in confirmatory trials. From a private sector drug development perspective, Cuffe et al (2014) discuss lessons learned from the application of operational and inferential seamless designs based on practical experience using real case studies. A review in Chapter 5 highlights the prevalence of the application of the seamless design in routine practice.

2.9.2 Reflection

Seamless designs require the understanding and cooperation of research governance bodies for the conditional approval of the integrated trial phases. Timely and efficient decision-making process between phases is needed. The decision-makers include the TSC, IDMC, Regulators, and Funders. Otherwise, the expected efficiencies may be lost. Upfront financial commitment to the whole integrated trial phases is needed, hence, the buy-in and commitment of Public Funders is very important. The designs also require pre-specification of the design properties of all phases at the design stage to enable planning, although proceeding phase depends on the results from the prior phase. Lastly, the application of the inferential seamless design requires a great deal of statistical expertise and its use in confirmatory trials is viewed with a degree of regulatory conservatism, particularly in pivotal trials, to be highlighted in Chapter 3 under Section 3.4.4.3. East (Cytel, 2015) and

ADDPLAN (ICON, 2015) commercial software, and open access R '*asd'* package (Parsons et al., 2011) support the implementation of certain inferential seamless designs.

2.10 Design 6: Multi-Arm Multi-Stage Design

This section introduces a MAMS design strictly conducted to address confirmatory objectives. Some key statistical and practical considerations are highlighted and important references provided. Available implementation software or code resources are highlighted.

2.10.1 Motivation

As highlighted in Chapter 1, trials are commonly conducted with two arms; one investigative intervention compared to a control. This approach has deficiencies in the presence of competing investigative interventions (Bratton et al., 2013; Bratton et al., 2015; Freidlin et al., 2008; Parmar et al., 2008, 2014). A MAMS design enables the evaluation of multiple competing interventions against a shared control as opposed to the conduct of a series of independent two arm trials and speeds up the evaluation process. Wason et al (2013) articulate the potential benefits of the MAMS design. Figure 2.8 displays a variant of the MAMS design with two interim analyses and final analysis. Three investigative interventions are compared to the control and at each interim analysis, only futile interventions are stopped ('drop-the-loser') based on some defined criterion.

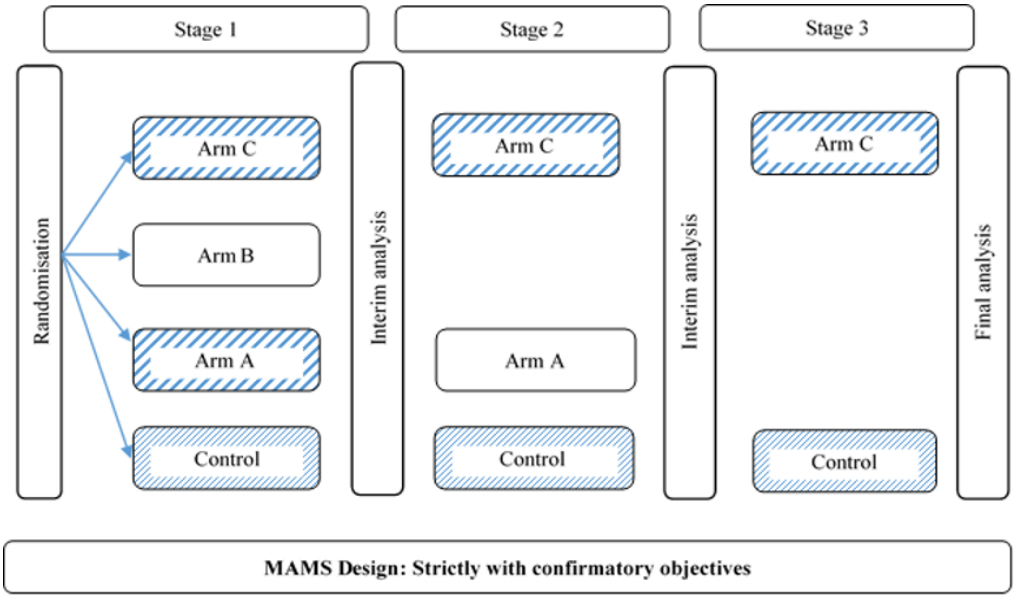

Figure 2.8. An example of a multi-arm multi-stage design.

2.10.2 Statistical and Practical Considerations

The MAMS design poses a number of statistical challenges. There is simultaneous multiple hypothesis testing of many pairwise comparisons at each interim analysis and multiple analyses of data across interims (Follmann et al., 1994). This should be adjusted to control for the type I error and power. The principle of weak and strong control of the familywise error rate has been articulated elsewhere (Dmitrienko et al., 2010; Follmann et al., 1994; Proschan et al., 1994). Various multiple testing principles and procedures suitable for different situations which are outside the scope of this thesis have been described (Dmitrienko et al., 2010).

One approach to control the type I error is to use multiple testing procedures within the group sequential methods. Thall et al (1988) describe a two stage approach only to select the 'best' promising of the multiple interventions compared to a shared control, which is limiting in the presence of competing promising interventions. Superiority test is only conducted at the final analysis if the trial proceeded beyond the 1st interim analysis. This design avoids multiple pairwise comparisons at the interim analysis, however, it is limiting. Follmann et al (1994) describe a hybrid group sequential approach with multiple pairwise comparisons to a shared control only allowing for dropping of futile interventions. Magirr et al (2012) extend this design to allow early stopping for futility and efficacy using the Dunnet (1955) procedure for strong control of the familywise type I error. For continuous outcomes, the design is implemented in R using the '*MAMS*' package (Jaki and Magirr, 2014). The method does not use flexible spending functions and another limitation is that it does not lessen the level of significance testing of future tests when some investigative interventions are dropped at proceeding interim analyses. Thus, future hypothesis tests are penalised. To address this problem, Proschan and Dodd (2014) devised a simple approach to drop arms with an interim intervention effect below a pre-specified threshold and nominal significance level adjusted for by excluding dropped arms using the Bonferonni (Bland and Altman, 1995), Dunnet (1955), Hochberg (1988), and Holm (1979) procedures. However, there is no proof of strong control of the type I error.

Wason and Jaki (2012) investigate the 'optimality' of a class of MAMS designs with respect to expected sample size, control of the type I error, and power through simulations. Some approaches to the MAMS design have been described in the context of survival outcomes (Royston et al., 2011) with a related '*nstage*' Stata implementation package (Bratton, Choodari-Oskooei and Royston, 2015). Wason et al (2013) provide statistical recommendations for the implementation of the MAMS design in confirmatory trials. These include:

1. The need for strong control of the familywise type I error,

2. Avoiding adding new intervention arms to an ongoing trial – otherwise future stopping boundaries would need to be adjusted to control the familywise error,

3. Powering of the trial should encompass a clinically relevant effect required to declare superiority and a minimum clinical effect of less importance below.

2.10.3 Reflection

The MAMS design has a potential to improve efficiency in the conduct of clinical trials in the presence of competing multiple investigative interventions. The results in Chapters 3 shows that the design has attracted attention across sector among multidisciplinary key stakeholders. However, there are statistical and practical challenges associated with the design. More resources and effort, particularly towards recruitment and implementation are required. Importantly, availability of statistical software or code is currently a barrier for implementation and statistical inference and is an area requiring further research. Maintaining blinding of the interim results is difficult. Those involved in the trial conduct can easily make indirect inference about the direction of intervention effects following decisions to stop recruitment in certain futile arm(s).

Chapter 3. Interviews Exploring Roadblocks to the Use of Adaptive Designs

3.1 Introduction and Rationale

In Chapter 1, it was highlighted how the use of ADs in clinical trials research has recently gained much attention because of their potential advantages in certain clinical scenarios compared to traditional RCTs (Kairalla et al., 2012). However, despite the promising benefits to Clinical Trials, patients, and Funders, the application of confirmatory ADs, especially in the public sector, has been viewed as disappointing in relation to the prominence given in the related statistical literature (Bauer and Einfalt, 2006; Morgan et al., 2014).

Citing the slow uptake of ADs, the pharmaceutical industry established a Pharmaceutical Research and Manufactures of America (*PhRMA*) Adaptive Design Working Group (ADWG) to facilitate dialogue among key stakeholders in drug development, and to develop a consensus position on the use of ADs (Gallo et al., 2006). Focusing on drug development across trial phases, the working group investigated barriers and opportunities to the use of ADs. Related research and discussions have subsequently been undertaken (Burnham et al., 2015; Coffey and Kairalla, 2008; Krams et al., 2007; Millard, 2012; Morgan et al., 2014; Quinlan and Krams, 2006; Quinlan et al., 2010). However, much of this research focused on the pharmaceutical industry perspective, particularly in the USA, with little attention on the publicly funded setting.

A number of authors highlight that the public sector may have its own unique multifaceted barriers, which need to be investigated, and addressed to improve the application of ADs in this setting (Coffey et al., 2012; Kairalla et al., 2012; SAACTD Workshop Committee, 2009). With this in mind, the NIH in the USA and associates funded and facilitated a two day workshop to establish cross-industry discussions with representatives from the NIH, FDA, EMA, the pharmaceutical industry, non-profit foundations, patient groups, and academia (Coffey et al., 2012; SAACTD Workshop Committee, 2009). Even though this workshop marked a significant milestone, it did not explore the perceptions and attitudes of key stakeholders directly involved in the day-to-day conduct of clinical trials and the decision-making process. Moreover, generalisability of some of the NIH findings to the UK publicly funded setting may be questionable.

To bridge this gap, but focusing on early phase trials, Jaki (2013) investigated the application of adaptive and Bayesian methods using a cross-sectional survey of registered UK CTUs. Jaki surveyed Statisticians and summarised five key perceived barriers hampering the use of these methods in early phase trials. Morgan et al

(2014) surveyed the use of ADs in the private sector and academia using different sources, and explored limited perceptions on barriers to use. This chapter endeavours to fill the gap in cited research by exploring perceptions on barriers, concerns, and potential facilitators to the application of ADs, among key stakeholders in trials research. Importantly, the research focuses on confirmatory RCTs in the publicly funded setting.

This chapter acknowledges the external support and advice of Dr Jonathan Boote, Prof Alicia O'Cathain, Dr Daniel Hind, and Prof Cindy Cooper. In addition, Kylie Cross, Helen Wakefield, and Lauren O'Hara for interviews transcription support. This work has already been published in Trials journal (Dimairo, Boote, Julious, Nicholl, et al., 2015).

3.2 Aims and Objectives

In order to bridge the cited gap in the literature, the main aim of this chapter is to undertake an in-depth investigation of underlying roadblocks impeding the appropriate use of confirmatory ADs, with a focus on the UK publicly funded setting. The specific objectives are to:

1) Explore barriers and concerns to the use of confirmatory ADs based on perceptions and experiences of key stakeholders in clinical trials research,
2) Explore the opinions on potential facilitators to mitigate some of the perceived roadblocks,
3) Use the findings to guide the design of follow-up surveys to gauge wider perceptions for generalisability of the findings.

3.3 Methods

It is paramount to understand the perceptions towards the use of ADs from the perspectives of those key stakeholders (Clinical Trialists and Decision-makers) directly involved in the conduct and funding of clinical trials in order to unlock the potential benefits of confirmatory ADs. In this regard, this research can be viewed within the phenomenological paradigm as it aims to investigate key stakeholders' experiences and perceptions and how they influence the understanding of obstacles to the use of confirmatory ADs (Englander, 2012).

3.3.1 Study Design

The research employed cross-sectional, in-depth, semi-structured, and one-to-one interviews of key stakeholders involved in clinical trials research (Legard et al., 2003). This approach encourages interviewees to

express their perceptions towards the use of ADs willingly by asking open-ended questions. The use of focus groups was deemed impractical. In addition, the focus group approach is inefficient for this research as dominant key stakeholders may stifle the freedom of expression of experiences and perceptions of others, which are important to address the research question.

The questions were *a priori*-designed based on topics from previous literature (Chang et al., 2006; Chow and Corey, 2011; Coffey and Kairalla, 2008; Coffey et al., 2012; Quinlan et al., 2010) and some were researcher-driven. Section 3.3.6 describes the interview process and the template of the interview guide used.

3.3.2 The Choice of the Sample Size

Most qualitative interview research uses the concept of reaching the data saturation point to justify the number of interviewees required. However, the point of data saturation is often unknown in advance, as it depends on a number of factors such as the scope and nature of the research subject, study design, and available resources (Creswell, 2007; Mason, 2010; Morse, 2000; O'Reilly and Parker, 2013). In addition to practical considerations, Creswell (2007) suggests the need to conduct up to 10 homogeneous interviews to address phenomenological research. In the context of this research, homogeneous groups relate to key stakeholders with similar expertise (roles and responsibilities) in clinical trials research. Hence, this research aimed to recruit 6 to 8 interviewees per expertise category to yield at least 20 interviewees, depending on the observed degree of overlap in expertise. An overlap in expertise of interviewees affords an opportunity to explore wider perceptions and experiences with a smaller sample. The need for further sampling in some expertise categories was adapted to reach saturation. This was guided by the richness of information gathered from previous interviews to explore certain phenomena.

3.3.3 Selection of Interview Participants

Interviewees were purposively sampled in sequence based on their primary roles and responsibilities in trials research until the desired sample was reached. A cross-disciplinary approach was adopted to optimise maximum variation to capture diverse perceptions and experiences (Coyne, 1997). These key stakeholders in clinical trials research with diverse expertise were sought to guide purposive sampling:

a) UK CTU leaders (Directors or Deputy Directors),

b) Members of public funding boards and advisory panels (Chairs or Vice Chairs and Ordinary members),

c) IDMC members,

d) Trial Statisticians,

e) Health Economists,

f) Chief Investigators and,

g) Regulators.

The selection of interview participants targeted key stakeholders, predominantly in the UK publicly funded sector. However, a cross-sector approach adopted sought some interviewees with private sector experiences, to explore diverse perceptions, experiences and attitudes. In particular, statistical expertise was purposively sought among private sector interviewees due to their perceived substantial experience on ADs (Kairalla et al., 2012). In addition, a small number of interviewees outside the UK were included heeding advice given by two interviewees during the interviews.

3.3.4 The Process of Approaching Target Participants

An invitation letter (Appendix 3.1) supported with an information sheet (Appendix 3.2) was sent to target participants using a number of platforms:

a) Mass emailing to key stakeholders within the specialised network groups of CTU leaders and Trial Statisticians through the UK CRC (Clinical Research Collaboration) registered CTU Network (UK CRC, 2014). This was based on a 2013 sampling frame of 45 registered UK CTUs;

b) Mass emailing to key stakeholders within the MRC Network of Hubs for Trial Methodology Research (NHTMR) via a periodic newsletter (MRC, 2014);

c) Personalised emailing to referred contacts and hard to reach groups such as regulators, public funding advisory panel or board members, and the private sector.

An invitation letter emphasised that targeted participants were eligible to take part regardless of their underlying perceptions, attitudes, and experiences. This was intended to minimise the potential of responder bias due to oversampling of participants likely to express positive perceptions towards ADs use. UK CTU leaders and lead Trial Statisticians were asked to circulate the invitation letter within their units.

Responders who expressed an interest to participate in the research were asked to complete a short questionnaire detailing their demographics and key expertise. In addition, responders were offered an opportunity

to ask questions on any research related issues they had. Responders who agreed to take part were requested to return a completed baseline questionnaire and signed informed consent form – either electronic or hard copy.

3.3.5 Research Ethics and Consenting Respondents

The REC of the School of Health and Related Research (ScHARR) at the University of Sheffield granted ethics approval (0676) for this research. Appendix 3.3 is the research governance and ethical approval letter granted. Appendix 3.4 is the signed ethics declaration for the research of Chapters 3 and 4. All interviewees signed the informed consent form prior to interview in accordance with the ethics approval requirements (Appendix 3.5).

3.3.6 The Interview Process

Between March and August 2014, interviews were conducted with informed consent by telephone or skype or through face-to-face conversations depending on feasibility and the preferences of the interviewee. This approach facilitates the inclusion of interviewees across a wider geographical area of interest.

Interview guides tailored for the expertise of interviewees to prompt questions guided the interview process. Despite the use of interview guides, attention was paid to prompts based on important markers mentioned by interviewees, which were relevant to the subject. The completion of the informed consent and a short baseline questionnaire was checked prior to all interviews. At the beginning of the interview, the following steps were undertaken:

- The interviewee was thanked for their willingness to contribute to the research and for their time;
- The overall aims and objectives of the research, its scope, what has been done, the future direction, and the expectations of the interviewee during the interview were explained to the interviewees;
- Interviewees were offered an opportunity to ask questions related to the research.

The interview guide covered the following specific topics or aspects during the interviews:

a) Interviewee's primary roles and responsibilities in clinical trials research;

b) Level of awareness, training, and understanding of ADs;

c) Familiarity with opportunities or benefits associated with the use of confirmatory ADs;

d) Awareness and knowledge of ADs, which are applicable in confirmatory trials;

e) Personal views and attitudes regarding the use of ADs and future prospects, particularly in confirmatory trials;

f) Perceptions regarding the use of ADs by members of the research community;

g) Views and attitudes towards the use of *ad hoc* versus planned ADs in confirmatory trials;

h) Perceptions about the accessibility of adaptive methods and implementation resources by key stakeholders in clinical trials research;

i) Perceptions about general challenges or obstacles hampering the use of ADs where appropriate at the design, implementation and reporting stages (*prompting experience and solutions to barriers where possible*);

j) Perceptions about the role specific challenges posed by the use of ADs (*prompting experiences and solutions to barriers where possible*);

k) Perceptions about challenges specific to the public funded setting (*prompting experiences and solutions to barriers where possible)*;

l) General concerns raised by the use of ADs or specific types of AD;

m) Experiences in the design, implementation, and reporting of adaptive clinical trials (*prompting examples and lessons learned which could be shared with other Clinical Trialists where possible*);

n) Perceptions about credibility, validity and acceptability of the findings from an adaptive trial (*prompting specific adaptations*);

o) What should be done to improve the uptake of ADs where appropriate in confirmatory setting.

Interviewees were asked at the end of the interview if they wished to verify their interview transcript or talk about something relevant on the subject, which they felt was not covered but worth contributing. Closing remarks thanked the interviewee for their contribution and promised a follow-up summary of the research findings in due course.

Five internal pilot interviews, of which 4 were face-to-face, were conducted to test the appropriateness of interview guide questions, prompts and interview duration. The interview process was found fit for purpose so no changes were made after internal pilot interviews. All interviews were audio recorded and verbatim transcribed with the help of three experienced transcribers within the Sheffield CTRU.

3.3.7 Analysis of Interviews and Reporting

Transcribed interview data were imported into NVivo10 software (QRS International, 2014), for the management and organisation of the data analysis process. The structure of the analytical process employed the

framework approach (Gale et al., 2013; Smith and Firth, 2011). The process involved: familiarisation and annotation of transcripts; identifying a thematic framework (Braun and Clarke, 2006); indexing, charting, mapping, and interpretation (Pope et al., 2000, 2006; Ritchie and Lewis, 2003). Mapping helps to establish associations between emerging themes and subthemes. Thus facilitating understanding of the subject, communication, and interpretation. Themes and subthemes captured the most important aspects of the data on perceptions, attitudes and experiences.

The classification of barriers to use into micro- and macro-level domains pertinent to key stakeholders at individual and organisational levels adapted a framework applied in the field of evidence-based practice (Cabana et al., 1999; Funk et al., 1991; Gifford et al., 2013). This approach helped the indexing of uncovered themes and interpretation of findings.

The conduct, analysis and reporting of this research was guided by the consolidated criteria for reporting of qualitative research (COREQ) checklist, where appropriate (Tong et al., 2007). Supplementary interview data and reported case studies of applied ADs are provided to support uncovered themes.

3.3.8 Quality Control Process

A subsample of interview transcripts was validated for consistency in the annotations and indexing process in identifying themes with the support of an experienced qualitative researcher (Dr Jonathan Boote). The number of interview transcripts for the validation process was not fixed in advance, but was influenced by a subjective measure of agreement. Satisfactory agreement was reached after cross validation of the first 26% (7/27) of the interview transcripts.

The mapping of themes and subthemes was discussed independently with two experienced qualitative researchers who were members of the advisory panel of this research (Dr Jonathan Boote and Prof Alicia O'Cathain). All interviewees were offered the option to verify their interview transcripts if they wished, although none considered the opportunity. Attention was given to deviant cases to explore possible explanations of the views expressed on the use of ADs (Pope et al., 2000, 2006), highlighted in Section 3.4.3.6. As part of external validation, the findings from interviews and quantitative surveys are compared in Section 4.5.3 of Chapter 4.

3.4 Results of In-depth Qualitative Interviews

This section begins by presenting the demographics, characteristics, and experiences of interviewees. Uncovered themes on perceived opportunities and barriers to and concerns about the use of ADs are presented and supported with selected interview data. Furthermore, some perceived potential facilitators to alleviate roadblocks are reported. These results have been reported elsewhere (Dimairo et al., 2015).

3.4.1 Demographics and Characteristics of Interviewees

Between March and August 2014, 27 participants were interviewed by telephone (n=17), skype video call (n=2), skype telephone call (n=1), and face-to-face (n=7). The median (IQR) duration of interviews was 31 (26 to 38) minutes, with a maximum of 51 minutes. Seventeen (63%) interviewees were male. Interviewees had diverse overlapping responsibilities in clinical trials research and the representation of the interplay between their roles is shown in Figure 3.1. Table 2 of (Dimairo et al., 2015) provides complementary summaries of the number of participants with overlapping roles.

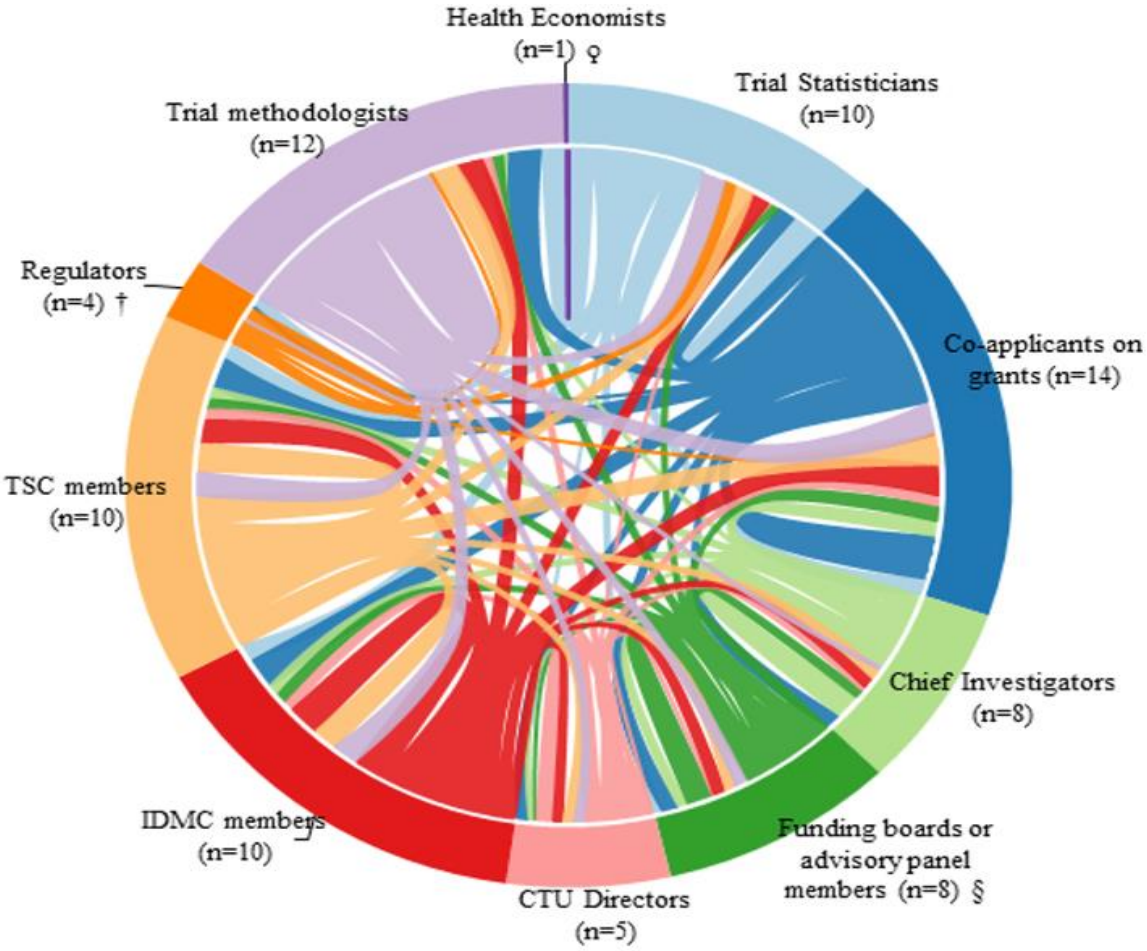

Figure 3.1. Overlap of roles and responsibilities of 27 interviewees in clinical trials research.

† Expertise during the interview fit in with statistical regulatory assessments although not stated as Statisticians on the baseline form. ϙ Member of a national health economics appraisal board. § 3 were panel chairs of public funding bodies.

Table 3.1 summarises the characteristics and demographics of interviewees. Interviewees were geographically from the UK 23(85%), and 4(15%) from Switzerland, Australia, German, and the USA. The majority (78%) of interviewees were holders of a PhD, DPhil or DSc or equivalent academic qualification. Only one health economist agreed and consented to take part in the interviews. The reasons for non-participation among 17 health economists directly invited to participate, were unfamiliarity with ADs (n=5), non-response (n=10), busy schedule (n=1), and expressed willingness but incompatible schedule (n=1).

Table 3.1. Characteristics and demographics of interviewed participants.

Variable	Scoring	Total (N=27)
Sex	Male	17(63%)
	Female	10(37%)
Age group (years)	>30-35	4(15%)
	>35-40	2(7%)
	>40-45	8(30%)
	>45-50	4(15%)
	>50-55	4(15%)
	>55	5(19%)
Academic qualifications	MSc/MA or equivalent	3(11%)
	PhD/DPhil/DSc or equivalent	21(78%)
	Other	3(11%)
Trials experience (years)	0 to 2	1(4%)
	>2 to 5	1(4%)
	>5 to 10	2(7%)
	>10 to 15	6(22%)
	>15	17(63%)
Current employment sector	Private	4(15%)
	Public [a]	22(81%)
	Both private and public	1(4%)
Mode of interview	Face-to-face	7(26%)
	Telephone	17(63%)
	Skype telephone call	1(4%)
	Skype video call	2(7%)
Location	UK	23(85%)
	International	4(15%)

[a] 6 participants had previous private sector experiences.

3.4.2 Clinical Trials Research and Adaptive Designs Experiences of Interviewees

Twenty-three (85%) interviewees had more than 10 years of clinical trials research experience. The publicly funded setting employed the majority of interviewees: public sector 22(81%), private sector 4(15%), and both public and private sector 1(4%). Six of the 22 interviewees employed in the public sector had employment history in the private sector. Interviewees had diverse degrees of previous experience of the application of ADs:

a) No experience at all (n=9), of which 6 expressed interest in ADs;

b) Only design experiences (n=9);

c) Experiences in the design and conduct, either in early phase or/and confirmatory trials (n=8);

d) Experiences of the statistical regulatory assessment process (n=4).

3.4.3 General Perceptions of Adaptive Designs in Confirmatory Trials

This subsection presents results of themes relating to the perceived benefits of confirmatory ADs, therapeutic areas of opportunity, and general attitudes.

3.4.3.1 Value of Adaptive Designs

Interviewees mentioned potential benefits of ADs that fall into three broad categories depending on the type of AD considered: ethical benefits to patients, value for money in clinical trials research, and improving design efficiency to address the research question(s).

3.4.3.1.1 Ethical Benefits to Patients

The option for early stopping of trials as soon as there is sufficient evidence to address the research question(s) offers opportunities to:

- Minimise over recruitment of research participants and exposure to potentially ineffective and/or unsafe interventions;
- Accelerate the evaluation, approval, and commissioning of interventions into clinical practice allowing patients to benefit from effective interventions quicker;
- Enable patients to be allocated to trial interventions, which they are more likely to respond better to – which is critical in serious health conditions;
- Identify subgroups of patients who are more likely to benefit from an intervention.

"It really depends on the type of AD, so if you have a GSD then of course you can stop early for futility or overwhelming effect and this clearly has many ethical and financial advantages. So for futility stopping – if it doesn't work you can stop early on and the patients don't get exposed to a drug which doesn't work or if you have overwhelming effect that is also very positive you can move on with the development of your drug and you don't have to finish the whole trial." (QL11 Statistician, design and conduct experience)

"…from a patient point of view, the sooner that if there is a new intervention that is really effective then we want to get that into NHS practice. Equally if it is dangerous or if there is anything that we shouldn't be using then we would want to get that out and into guidelines and NHS practice as much as possible." (QL01 CTU Deputy Director, Proposal Developer, design experience)

3.4.3.1.2 Value for Money in Clinical Trials Research

Interviewees stated that options for early stopping of trials might offer opportunities to avoid the pursuit of 'lost causes', such as paying for non-essential data. Therefore, it enables Funders to reallocate limited available resources efficiently to other promising or priority areas.

"The main advantages of ADs accrue to Funders because Funders are not paying for essentially redundant data and ethically I think there is a benefit to patients because clearly we don't want to be recruiting patients to trials when there is no significant potential of that trial and additional data giving you any new information." (QL04 Chief Investigator, Vice Chair – Public Funder, design experience)

" … ADs will make you address the objectives of interest enabling you to make the right decisions earlier rather than later. For example, the biggest opportunity is stopping poor drugs early. Most of our drugs fail, 90% of the drugs that we start developing in phase 1 never get to the full registration, we should be killing those drugs as early as possible and ADs allow you to do that, whether it is in phase 2 or 3 there is always that opportunity to stop early for futility." (QL15 Statistician, design and conduct experience)

3.4.3.1.3 Efficiency in the Design and Evaluation of Investigative Interventions

Depending on the type of AD considered, interviewees pointed out that ADs may be of value in improving efficiency by:

- Mitigating the risks of inaccurate design assumptions;
- Enabling simultaneous testing of multiple competing interventions in a single trial instead of multiple series of two arm trials;
- Enhancing the swift addressing of research questions to expedite decision-making;
- Making efficient use of limited patient populations, particularly important in rare or orphan diseases;
- Making efficient use of a finite pool of investigators.

"… when you certainly have a limited patient pool, like orphan disease implications where you know you are not going to be able to actually recruit sufficient patients for a full Phase 2/3 traditional development programme. And we accept and understand that, as Regulators and industry – you are offered appropriate incentives under orphan designation in the EU (European Union) and US. So, there is undoubtedly a challenge – an opportunity to maximise the best use of patients. I have once actually seen a combined Phase 1/2/3 study, all in one go." (QL16 Regulator, assessment experience)

3.4.3.2 Therapeutic Areas of Opportunity

Most interviewees highlighted that ADs are applicable across a wide spectrum of therapeutic conditions. However, some interviewees felt that ADs may be more appropriate or appealing for certain health conditions or populations. This may be influenced by factors such as severity of the health condition raising the importance of ethical consideration, standard-of-care options available to patients, and limitations of standard methods when addressing some research questions.

> "Yes we do run adaptive designs at all stages of development phase 1, 2 and 3 … we run a lot of group sequential designs, they are very common these days. It depends a little bit on the therapeutic area but I think group sequential designs nowadays are very common and very popular and we do have some trials on sample size re-estimation, blinded but also unblinded sample size re-estimation and we have a few examples on more complex adaptation like treatment selection, population selection and things like that…. Maybe some areas are also easier than other areas like oncology might be a bit easier to perform an adaptive design because it is a life threatening disease. " (QL11 Statistician, design and conduct experience)

The areas of opportunity mentioned include, but are not limited to, oncology; emergency medicine; respiratory, cardiovascular, infectious, and rare diseases.

As highlighted throughout Chapter 2, interviewees stated that the nature of the clinical primary endpoint(s) considered is important in influencing the relevance of the proposed ADs and practicality of implementation. An experienced investigator shared a case study demonstrating how an AD could be valuable in evaluating interventions during outbreaks of rapid evolving and fatal pandemics such as influenza or Ebola (Case Study A, Appendix 3.6). The interviewee cited the severity of the conditions, coupled with the need for urgent policy decision-making as driving factors.

Importantly, most interviewees emphasised the imperative need for Clinical Trialists to provide a clear rationale for the proposed AD and its fitness for purpose to address research question(s).

3.4.3.3 Shift in Attitudes of Public Funders

Views expressed demonstrate a paradigm change in attitudes towards the application of ADs by Public Funders motivated by the value for money and desire to use the public purse more efficiently. Public Funders interviewed expressed desire and receptiveness to fund adaptive trials and encouraged Trialists to consider ADs, as long as they are appropriate to address the research question(s). Some Clinical Trialists interviewed acknowledged this change in attitudes by Public Funders.

"I think generally speaking we are receptive to those ideas (of ADs) and in fact we, at [organisation] have held our own workshops on ADs last year or the year before in order to try and promote more use of ADs providing they are appropriate of course. So I think ten years ago our attitudes were more towards traditional parallel group. It was a sort of traditional well-known pathway but I think now our modern thinking is that we welcome ADs when appropriate and it is very much for the applicants to make the case for why they want maybe 4 arms with interim analyses for dropping arms." (QL35 Chair – Public Funder)

3.4.3.4 Regulators' Receptiveness and Improving Awareness and Experiences

There appeared to be regulatory receptiveness towards the application of ADs in principle. However, the receptiveness seems conditional on strong caveats relating to aspects such as measures to minimise operational bias to preserve trial credibility and integrity, control of type I error, and use of appropriate statistical inference. Regulators interviewed highlighted that the need for an audit trail with tangible evidence to show that such caveats are met is paramount.

The views of interviewees highlighted growing regulatory awareness and experience on AD-related aspects, particularly among statistical assessors. This is a result of the increasing numbers of AD-related scientific advice consultations and applications by Clinical Trialists; especially on SSR, futility analysis, and GSD trials.

"I haven't got the figures in front of me and I wouldn't know how to get them but you see a lot more of them at the moment in the scientific advice arena, when people are coming saying "this is what we are going to do, what do you think?" … I get a lot of them starting and not so many of them have finished yet." (QL19 Regulator, regulatory assessment experience)

Importantly, Regulators advised Clinical Trialists to continuously engage them through scientific advice meetings, and to adhere to their guidance on appropriate application of ADs from planning to trial completion.

3.4.3.5 Cross-disciplinary Interest

The majority of interviewees conveyed widespread growing interest towards the use of ADs, despite the existence of the acknowledged issues highlighted in Section 3.4.4.

"I guess there are a lot of concerns about them and so that's perhaps why they're not taken up so much. But it is interesting to see that there's a lot more interest in the past few years and so maybe that is changing." (QL26 Statistician, design and conduct experience)

"… influential bodies like the FDA are now embracing ADs and there is probably an increasing number of ADs that are being utilised and will come through and report over the next 2/3/4 years …" (QL21 Chief Investigator, design experience)

The desire to improve design efficiency to address research questions and ethical issues, and maximise value for money in research appears to influence the expressed cross-disciplinary interest. This is because of the acknowledgement of the shortcomings of the traditional approach to the conduct of clinical trials, which requires improvements in certain situations.

"… (ADs) makes a lot of sense from my point of view and in terms of optimising the design and feasibility of the study to address the particular research question. I think it is important that Statisticians and Clinicians discuss thoroughly the options that are available in clinical trial design to agree the best proposal because each will have particular insights with regard to how to address a research question and so communication is really essential." (QL24 Chief Investigator, design experience)

"…there is a lot of interest in them from a Funder's point of view, in that particularly difficulties in recruitment. When it has taken a long time to recruit for trials when recruitment is not up to its expected levels, it is very helpful to be able to have a design that allows you to have multiple looks at the data and to potentially stop early." (QL04 Chief Investigator, Vice Chair –Public Funder, design experience)

3.4.3.6 Positive Clinical Will

The interviewed clinical investigators communicated a positive will and receptiveness towards use of ADs to exploit opportunities, whenever appropriate to address research questions efficiently.

"We definitely have an interest in advancing new methods in the field of sepsis and in particular there is probably room for improving clinical trial design and that is the focus of our group (ADs methods)." (QL22 Chief Investigator, design experience)

Nonetheless, the conveyed positive desire by clinical investigators appears to depend mostly on how Trialists market ADs to them and the availability of technical and practical support for their design and implementation.

"Sometimes you need to sell it to them to get them to see its positives and advantages and in terms of the extra complication it takes to implement them". (QL07 Statistician, design and conduct experience)

"I think generally once you have explained it (AD), and said that it will be a very big expensive trial if we did it fully powered for as long as it would take, but that it can be broken down to give different options to the Funder for shorter time periods, and less cost – then they can see the advantages to it. … If we are happy to do it and design it, and write that section up for them they will take it on." (QL01 CTU Deputy Director, Proposal Developer, design experience)

Although the majority of interviewees expressed widespread interest and receptiveness towards appropriate use of ADs, one clinical investigator interviewed insistently expressed negative views towards the notion of ADs. The interviewee had more than 25 years of experience leading the conduct of large multicentre trials involving thousands of patients. On further examination, the interviewee disclosed the frequent use of unplanned *ad hoc* adaptations using unblinded data, for example, to review the sample size based on the observed intervention effect. This revealed a clear lack of understanding of the impact of such *ad hoc* changes based on unblinded data on statistical aspects of the design and introduction of operational bias in trial conduct.

"In other words, we would unblind the lipid differences (treatment effect) during the trial, and as a result of that information, we decided to extend the duration of follow-up for the trial, to give ourselves, you know, a better chance of detecting an effect, because the difference in the cholesterol was less than we had anticipated when we designed the trial. So you could argue that we adapted on the basis of that, and some people would consider that, therefore, an adaptive design. I mean, we would just consider that to be good monitoring and part of what the oversight of an ongoing clinical trial should include ..." (QL06 Chief Investigator, no design and conduct experience)

3.4.4 Perception of Themes on Barriers in Confirmatory Trials

This subsection reports results relating to barriers to and concerns about the use of confirmatory ADs based on perceptions and experiences of interviewees.

3.4.4.1 Cross-disciplinary Lack of Awareness

Some interviewees communicated the widespread lack of awareness of the types of ADs and scope, circumstances when ADs are appropriate, and implementation resources, as barriers to appropriate use. Consequently, some interviewees conveyed missed opportunities and underutilisation of ADs when appropriate in some trials. Due to the growing prominence of ADs, some conveyed concern that the misunderstanding of when they are appropriate may lead to misuse including in certain circumstances when ADs are not superior to traditional fixed sample size designs.

3.4.4.2 Misunderstanding of the Meaning of an Adaptive Design

One resonant message inferred is the potential for confusion concerning the meaning of an AD, and acceptable scope in confirmatory setting. Interviewees acknowledged a broadening in scope of what is considered

as an AD in recent years. Hence the term is often loosely defined, but with broad contextual meaning prone to misinterpretation leading to confusion among Clinical Trialists.

"I would say, over the last three years, I've become aware of (the) detail of ADs. Prior to that, it was a sort of loosely bandied term … I could be in a room and everybody thinks they're talking about the same thing and they're talking about very different things." (QL08 CTU Director, no experience)

"So I am generally in favour (of ADs), however convincing the community of that takes some work. So a big threat for ADs is just that it's a cutesy word which means different things to different people, there's misinformation about it and there are some existing biases in the community and so there really needs to be a lot of education." (QL22 Chief Investigator, design experience)

Some interviewees pointed out that the confusion highlighted has recently been partly addressed from a regulatory and industry perspective, through some guidance (FDA, 2015). However, they still believed that there is a current problem in the public sector, where many investigative interventions do not require regulatory approval beyond standard ethics.

3.4.4.3 Cross-disciplinary Degree of Conservatism

The majority of interviewees viewed cross-disciplinary conservatism as one of the major barriers to the usage of ADs in the confirmatory phase. Essentially, this complex multifaceted degree of conservatism appears influenced by many factors such as:

- Trial phase and nature of research objective(s);
- Therapeutic area, study population, and nature of intervention under investigation;
- Rationale put forward and completeness in description of the proposed AD(s);
- Type and scope of the proposed AD, the availability of well-established methods for statistical inference, and perceptions towards that AD by policymakers;
- Perceived complexities associated with the AD and impact on implementation, potential introduction of operational bias during trial conduct, and interpretation of the findings;
- Underlying familiarity and understanding of the proposed AD.

Details of inferred factors influencing conservatism and negative attitudes towards the use of confirmatory ADs are summarised in Table 3.2.

Table 3.2. Inductive themes perceived to influence conservatism to confirmatory adaptive designs use.

Stakeholder	Secondary theme associated with conservatism	Contributors linked to secondary theme
Cross-disciplinary	Unfamiliarity and lack of understanding Fear of introducing operational bias during conduct and compromising the credibility of the trial	
	Concern about the robustness of ADs in decision-making	Fear of making wrong decisions Concerns about premature early stopping of trials Concern that the research community may struggle or be reluctant to accept the findings from an adaptive trial Contrived general perception by Journal Editors and Reviewers that early trial stopping is a failure Impact of early trial stopping on other secondary but important objectives
	Research teams being more comfortable with traditional fixed sample size designs than ADs	Sticking to what we know best and fear of venturing into the unknown Lack of knowledge and experience Generation effect – more senior Trialists being sceptical of change from what they know best and perceive as standard Perceived operational and statistical complexities during planning and implementation
Regulators	Buy-in reluctance in the confirmatory setting	Lack of understanding of the inferential and regulatory price to pay by using an AD Fear of lowering the level of evidence Fear of making wrong decisions which may taint their reputation in the future (for instance, approving a drug which will subsequently be proved to be unsafe or ineffective) Limited experiences in the assessment and approval of ADs
Statisticians	Negative attitude towards ADs among some influential statistical community members	Generation effect – more senior researchers being sceptical of change from what they know best and perceive as standard
Private and Public Funders	Reluctant to fund potential high risk high value research projects with huge uncertainty	Uncertainty around the actual sample size, duration and actual cost of the trial Inadequate description of variable costs, decision-making criteria and time frames on grant applications (Public Funders)
	Limited commissioning and funding experiences, especially among Public Funders	Difficulties in drawing up flexible employment contracts (Public Funders) Limited number of AD grant proposals being submitted by researchers for consideration (Public Funders)
	Negative attitudes towards ADs among some public funding panel members	Lack of familiarity
IDMC and TSC members	Perceived negative attitudes towards multiple examinations of the trial data Reluctant to stop trials early unless for safety reasons	Lack of familiarity and understanding

IDMC: Independent Data Monitoring Committee; TSC: Trial Steering Committee.

The majority of interviews highlighted the limited scope for ADs in confirmatory trials because of the definitive nature of research objectives, with direct influence on policy decision-making in approving interventions into clinical practice. As a result, some interviewees strongly advised against undertaking too many adaptations in confirmatory trials, citing challenges in the interpretation of results.

"… people should be cautious I guess in trying to do too much and having too many adaptations … We must still make sure we have that body of confirmatory evidence, so I think there might be a place in phase 3 for ADs, but only sort of minimal adaptations. We should sort of keep things under control in that particular setting …" (QL19 Statistician, regulatory assessment experience)

The insufficient description of the scope of the proposed AD, and related statistical and operational properties supported with tangible evidence, such as from simulations or established references, was viewed to influence conservatism.

Some ADs such as the MAMS design attracted cross-disciplinary attention, particularly from policy and decision-makers. Interviewees cited potential efficiency and value for money in testing multiple competing interventions in a single trial, as opposed to conducting multiple series of independent two arms trials.

"In terms of the multi-arm trials I'm much more comfortable now with the idea of maybe setting out, even on a phase 3 trial, with 4 or 5 potential interventions and dropping the ones that look least promising." (QL14 Statistician, no experience)

In contrast, some Regulators and Statisticians, expressed reservations towards the seamless 2/3 AD introduced in Section 2.9 of Chapter 2. They cited questionable efficiency due to the lack of adequate 'thinking time' between phases and the need to pre-specify the design properties of phase 3 at the onset of phase 2. In addition, some Regulators expressed concerns specific to inferential seamless 2/3 AD, citing the lack of understanding of the regulatory and inferential price to pay by using this design, and regulatory dilemma regarding where the design lies in the hierarchy of confirmatory evidence. Consequently, Regulators feared lowering the level of confirmatory evidence.

Confusion in terminology between operational and inferential seamless ADs, perhaps due to poor communication of the methodology was inferred. Both seamless ADs aim to reduce the time in testing interventions by combining phase 2 and 3 objectives in one trial under a single protocol; conventionally addressed in separate trials. However, in addressing confirmatory objectives, operationally seamless AD only uses phase 3 outcome data whereas inferentially seamless AD combine phase 2 and 3 outcome data (before and after

adaptation). Some Regulators and Statisticians expressed concern that combining phase 2 and 3 outcome data for confirmatory inference may shift the intervention effect towards clinical benefit, driven primarily by phase 2 data. Hence, inferentially seamless ADs require more complex statistical methods to account for potential bias in inference.

Concerns were also raised regarding potential population drift when using a response adaptive randomisation design where the chance of patients being allocated to an investigative intervention is modified depending on the clinical outcomes of those already in the trial (Case Study C, Appendix 3.8).

3.4.4.4 Lack of Knowledge and Experience

Most interviewees communicated the lack of knowledge and experience of ADs as a major roadblock preventing their appropriate use. This was viewed as intertwined with insufficient access to case studies to facilitate practical training, raising awareness of benefits and when ADs are appropriate, and learning about barriers and facilitators to successful implementation. Certain interviewees raised concerns about deficiencies in the current training approaches, viewing these as more oriented towards statistical methodology rather than translational practical learning. More so, the weaknesses in some current academic graduate training curricula, which do not tend to incorporate ADs as alternative designs, were articulated.

"… the main challenge … I think it is a bit broader - is the lack of experience and knowledge within the bio-statistics community. There is a lack of understanding of adaptive methods, a lack of understanding of the opportunities, you know and a lack of familiarity." (QL12 Clinical Research Leader, Trial Methodologist, design and conduct experience)

A number of interviewees conveyed a lack of familiarity and knowledge of alternative ethical and efficient designs among ethics and scientific review board members, which may obstruct their ability to adequately review grant proposals.

3.4.4.5 Statistical and Practical Complexity

3.4.4.5.1 Amount of Extra Time and Effort Required

Most interviewees stated that in general, ADs require additional time, work and effort from a statistical and operational perspective, compared to traditional fixed sample size designs during planning and implementation.

3.4.4.5.2 Implementation Practicalities

The majority of interviewees highlighted the importance of how the implementation of the proposed AD is going to work in practice. This operational feasibility includes aspects such as logistics and administration, resources, accrual of the primary endpoint data relative to the expected recruitment rate, implications of trial governance processes and collaborating sites, and intervention delivery. The level of operational challenge tends to grow with increase in the complexity of the proposed AD.

3.4.4.5.3 Simulation Design Work

Depending on the complexity of the proposed AD, interviewees mentioned that ADs require more effort and time to undertake adequate simulation work under various scenarios. This helps to understand the statistical properties of the design and implications of decision-making scenarios. Some of the interviewees voiced concerns about inadequate simulation work and its consequences on statistical properties and decision-making. Some interviewed Regulators raised similar concerns about a response adaptive randomisation case study, on whether the simulations were adequate to cover the entire domain of the desired sample space to guarantee control of the type I error (Case Study C, Appendix 3.8). Interviewees identified the need for applied training of Trial Statisticians on performing adequate simulation work of ADs.

3.4.4.5.4 Robust Data Management Infrastructure

Interviewees expressed that data management and related logistical challenges may hinder the application of ADs. This is because of the need to provide robust data to inform the adaptation process and to minimise introduction of operational bias. Interviewees highlighted some important considerations:

- Compatibility of data management infrastructure with collaborators;
- Real time data capturing, cleaning, and processing. An interviewee shared an example of a successful multicentre case study using tablet computers for real time electronic data capturing in an African-based trial setting (Case Study B, Appendix 3.7);
- Turnaround time of data management processes to inform the adaptation;
- Systems, processes, and procedures supported with audit trails to minimise potential operational bias encompassing what information should be disclosed and to whom, how the information should be transferred, and firewalls and clarity on who is doing what.

3.4.4.5.5 Additional Statistical Considerations

In-house capacity of statistical expertise supported with quality control, validated software or user-written statistical codes to execute the AD, and delivery time of results to inform interim decision-making were among some of the perceived statistical obstacles. However, these depend on the complexity of the proposed AD. An experienced Statistician shared a case study, in which they adapted methods from another clinical area using a different endpoint, but with additional statistical work and time commitment (Case Study B, Appendix 3.7).

3.4.4.5.6 Maintaining Trial Credibility and IDMC Duties

Interviewees highlighted the importance of maintaining confidentiality by the IDMC during communication and execution of their duties supported with documentation. It was advised that the training of, and discussions with, IDMC members prior to trial commencement regarding the proposed AD; related decision-making criteria; execution of their duties as guided by formalised documents; communication protocol; and clarification on related issues are essential. Depending on the complexity of the AD, certain interviewees highlighted that the IDMC members may require more effort, time and expertise to understand the design, its decision rules and execution.

3.4.4.6 Challenges in Marketing ADs to Key Stakeholders

As highlighted in Section 3.4.3.6, some interviewees perceived that more time and effort is needed to market and communicate the rationale and practical aspects for the proposed AD to key stakeholders during planning. The target key stakeholders include Funders, Regulators, Clinical Collaborators, and patients. The amount of time and effort depend on the type and scope of the proposed AD.

3.4.4.7 Concerns about Trial Credibility and Integrity

3.4.4.7.1 Preference for Prospectively Planned Adaptive Designs

The majority of interviewees expressed strong preference for prospectively planned ADs, with decision rules clearly pre-specified at the design stage. This facilitates adequate understanding of the design's statistical properties through simulation and enhances proper planning. Regulators interviewed emphasised that pre-planning of ADs is a regulatory necessity to safeguard trial credibility, integrity, and validity. A resonant view was that ADs are not a fix for poor planning. Hence, interviewees voiced concern about unplanned *ad hoc* adaptations, which they view with great suspicion for cherry-picking and potentially hiding negative findings to advance hidden personal agendas of some researchers.

"I think it (ADs) will always raise an element of suspicion if there have been some decisions made along the way that have been data driven. And the key thing is just to have all the documentation in place; it has to be set out precisely in the protocol how it will be done and you need the right mechanisms in terms of the monitoring committee or steering committee makes the decision and make sure you comply with all the mechanisms. I mean it's like GCP (Good Clinical Practice); it's not enough to do the right thing, you've actually got to be able to prove you've done the right thing… with adaptive trials it's that much harder to prove that you've done it legitimately. So you've got to be very careful about the process and got to be able to demonstrate through documentation that you have followed true process." (QL14 Statistician, no experience)

Even though most interviewees acknowledged routine monitoring as part of every trial, some communicated a lack of understanding of the impact of *ad hoc* changes on the statistical properties of the design, introduction of bias, interpretation and credibility of the trial results. Some interviewees highlighted the need for some minimal flexibility in case of unexpected circumstances within the planned AD framework.

3.4.4.7.2 Fear of Introducing Operational Bias

The fear of compromising the trial credibility through introduction of operational bias during conduct and the potential population drift during adaptation due to knowledge of interim results were major perceived concerns.

3.4.4.8 Concerns about Validity of Trial Findings

Some Statisticians and Regulators interviewed expressed anxiety regarding appropriate use of statistical inference following an AD. They argued that Clinical Trialists pay little attention to the impact on trial results (point estimates, CIs and p-values). However, there was acknowledgement of improvement in awareness regarding control of the type I error. In addition, some interviewees highlighted the need for adequate transparency in the conduct and reporting of ADs. Opinions appear divided on whether the current CONSORT guidance for fixed sample size designs is fit for purpose for ADs.

3.4.4.9 Public Sector Perspective

3.4.4.9.1 Worry about Impact of ADs on Research Staff Employment Contracts

Some interviewees mentioned that the existing public funding models for trials designed with fixed sample size create uncertainty for research staff contracts when trials are stopped early. As a result, some UK CTU leaders are anxious about supporting certain ADs with options for early stopping. Nevertheless, it was highlighted that this problem is not unique to ADs, since some fixed designed trials are stopped early, mainly because of poor recruitment.

Some interviewees mentioned that design flexibility is inevitable because of the UK Public Funders' preferences towards risk assessment in trials using internal pilots with staggered research contracts. In addition to the reputation and experience of the CTU, interviewees highlighted that concerns about the impact of the funding model on research contracts depends on factors such as the:

- Type of AD proposed – ADs such as SSR and MAMS are less likely to be affected,
- Size of the research group and trial portfolio – large CTUs can more easily reassign staff to other trials in the pipeline when a trial is stopped early,
- Remit of the Public Funder and flexibility in their funding models.

"… Because of the size of the trials unit, there are many trials taking place so we look very closely at people's contracts and what studies are taking place. It is not just based on one study. We have a lot of different trials at the trials unit so the infrastructure allows for –if the trial stops early then they would be able to work on another trial. So it is not driven by the fact that the contracts or by whether or not it would stop early on this particular trial because of the other trials taking place requiring statistical, trial management, data management support." (QL27 Statistician, design and conduct experience)

Interviewees advised Public Funders to draw up standardised, flexible funding agreements compatible with key research partners including CTUs, universities, research sites and UK CRN (NIHR CRN, 2015). Interviewed Public Funders acknowledged the need to produce such contracts. Some suggested modification to the current staggered research contracts for trials with internal pilots or those for programme grants.

3.4.4.9.2 Lack of Capacity and Time within UK CTUs

Interviewees mentioned the lack of expertise and capacity, particularly a dearth of Statisticians and proposal developers to support complex ADs. However, they acknowledged that the level of capacity and expertise varies across CTUs. A resonant roadblock mentioned was the time limitation and consequent inability to support the design of complex ADs. They cited the extra work required against existing pressure to deliver on competing priorities based on conventional fixed sample size designs.

"One is just the lack of expertise within the unit, so it is easier when you are very busy to put forward a design you know rather than one you don't. It is also easier because if you put forward a design that does not look the same to Clinicians who expect straightforward designs you have to be very confident in that design to be able to convince them to some extent." (QL9 CTU Director, Statistician, design experience)

3.4.4.9.3 Lack of Bridge Funding for UK CTUs to Support Planning

Some CTU leaders raised concerns about the lack of a business case to support the design of complex ADs because of the time required, which is unpaid for, given the uncertain future success of grant applications. CTU leaders called for funding opportunities in the form of design development grants to support adequate design work of complex ADs. Such grants could be conditional on proposals meeting research and funding priorities of the Public Funders.

"I think for some of the really complex ADs, it would be good if there was availability to go for some small trial development grants so that you could say "look this is a convincing clinical question, we think it should be approximately this sort of design but actually we need £20,000 or whatever to properly work it up and design it." And that type of trial development grant I think would help unlock some of that." (QL09 CTU Director, Statistician, design experience)

Even though a Public Funder admitted the need for such grants, the interviewee argued that bridge funding is partly addressed through the NIHR infrastructure support funding accessible to over 25 accredited UK CTUs on a rolling contract basis (NIHR, 2014a). However, the views of CTU leaders seem to suggest that this funding is insufficient given the high risk associated with supporting the design work of complex ADs. Public Funders suggested that researchers might consider applying for small grants within the remit of other NIHR funding streams to support developmental work of ADs.

"Typically for complex ADs then you have to do quite a lot of modelling –that could take 12 or 18 months. Ideally, there should be grants to cover that early development work. Yes, I have sympathy to the idea that there needs to be additional funding but on the other hand I suppose all work that CTUs do prior to a trial application is done at risk. When I was CTU director, typically you are talking about 2 years work before you applied to do a definitive trial. I could say there ought to be more grants to help with all of that and the reality is that we in [organisation] in a sense do pay that upfront because we have a scheme whereby we support CTUs. We give them £250,000 per year if you like, like a front loaded loan, which they use to buy core staff in order to develop new projects. So in a way I think we are doing it already." (QL25 Chair – Public Funder)

3.4.4.9.4 Constraints of the Current Grant Application Process

Some interviewees highlighted the need to increase the proposal development time prior to submission deadlines to give researchers adequate planning time, especially for commissioned calls. This is more relevant for particular types of ADs, which are time consuming to design and require extensive statistical simulation work.

"From a practical point of view when you are designing adaptive trials there is more work involved for the application in planning the trial and working out the timelines … you have to do it for a number of different scenarios. So the work involved in that is more from the Trialist and Statistician's point of view, the Statistician

has to do various modelling and look at different scenarios and we have to do all of the different planning and you are usually on a fairly tight deadline for applications because of the way that NIHR funding works. So if you only have 6 weeks to work with the team, trying to fit in time to do lots of different scenarios can be quite tricky and can make it more difficult." (QL01 CTU Deputy Director, Proposal Developer, design experience)

Some interviewees also argued for a slight modification to the grant application form to give enough space for researchers to describe the rationale, design and its properties, decision scenarios, and variable costs adequately. This may also allow the inclusion of relevant appendices.

"There is not an existing section in grant submissions that says "if you are doing an adaptive trial design please provide the following information", so I just don't know that it's well organised yet and that could be a good thing or a bad thing …" (QL22 Chief Investigator, design experience)

3.5 Discussion

This chapter laid the foundation exploring the perceptions of a diverse group of key stakeholders on the use of ADs based on their opinions and experiences. These results informed the design of quantitative surveys described in the next chapter aimed to draw generalisable inference on perceptions on barriers and some potential solutions. There are numerous cross-sector barriers uncovered in confirmatory trials that include:

- Lack of practical knowledge and experience,
- Lack of applied training coupled with insufficient access to case studies of undertaken ADs to facilitate practical learning,
- Lack of awareness of opportunities associated with AD use,
- Lack of understanding of the acceptable scope of ADs,
- Statistical and practical complexities associated with the planning and implementation of ADs,
- Limited time to support adequate planning relative to the competing priorities based on traditional designs,
- Challenges in marketing ADs to other key stakeholders,
- Additional demands for data management infrastructure,
- Multifaceted degree of conservatism influenced by various factors, which have been described in detail.

Barriers specific to the public sector encompass:

- Lack of bridge funding accessible to UK CTUs to support the design work of complex ADs,
- Anxiety about the impact of early stopping of trials on staff research contracts,
- Lack of capacity and expertise within CTUs to support ADs.

Most importantly, some of the factors inferred to influence a complex degree of conservatism in confirmatory trials include:

- Fear of introducing operational bias,
- Concerns about robustness and credibility of ADs in decision-making,
- A degree of regulatory buy-in reluctance,
- Research teams being more comfortable with traditional fixed designs.

Despite the numerous roadblocks found, the positive clinical will, cross-sector and cross-disciplinary interest among interviewees in appropriate use of ADs to explore opportunities is encouraging. The positive clinical will found appears to contradict previous findings in early phase trials that suggested that clinical investigators insist on the application of certain preferred traditional designs (Jaki, 2013). Recent literature based on surveys supports inferred improvement in regulatory awareness and experiences of certain ADs in the EU and USA (Elsäßer et al., 2014; Gaydos et al., 2012; Lin et al., 2015; Quinlan et al., 2010).

The overwhelming preference for 'prospectively-planned adaptation' or the 'adaptation by design' concept, particularly in the confirmatory phase conveyed by interviewees, is reinforced in the literature and regulatory guidance documents (CHMP, 2007; Chow and Chang, 2008; Chow and Corey, 2011; Coffey and Kairalla, 2008; FDA, 2010, 2015). The shared views on the need for safeguards and firewalls to minimise leaking of interim results, with clear processes, procedures and documentation with audit trails are reinforced elsewhere (Gallo, 2006; Quinlan et al., 2010). Detailed regulatory considerations and caveats during the planning and implementation of ADs are highlighted in guidance and reflection documents (CHMP, 2007; Elsäßer et al., 2014; FDA, 2010, 2015).

Some of the key facilitators to enhance appropriate use of ADs in confirmatory trials for further investigation through quantitative surveys include the need for:

- A CONSORT guidance document tailored for ADs to enhance their conduct and reporting;
- A troubleshooting 'toolkit' of general and design specific statistical and practical questions or issues that Clinical Trialists need to consider when contemplating ADs;

- An AD guidance document tailored for publicly funded trials similar to the one for the development of complex interventions (Craig et al., 2008, 2013).

A detailed discussion of some potential facilitators to overcome barriers, sector differences on perceptions, impact of the results on future research, and comparison of the findings with extant literature is specifically conducted in Section 4.5 of Chapter 4.

In summary, the findings presented are based upon views and experiences of key stakeholders with diverse expertise. This maximised the capturing of diverse perceptions hindering the use of ADs, thus enhancing robust exploration of barriers, concerns and some potential facilitators to improve the appropriate uptake of ADs. In addition, cross validation to check for consistency during the analytical process was undertaken with the support of experienced qualitative researchers. However, the main limitation is the poor participation of health economists that limited the exploration of ADs-related issues among this stakeholder group. The non-participation was likely due to a lack of basic understanding of ADs and their implications for health economic evaluation, and to some extent, Health Economists may feel on the periphery of clinical trial design. Therefore, there is scope for further research to understand the implications of ADs on health economic evaluation.

Chapter 4. Surveys on Perceptions of the Use of Confirmatory Adaptive Designs

4.1 Introduction

In Chapter 3, cross-sector roadblocks to the appropriate use of ADs based upon in-depth interviews of cross-disciplinary key stakeholders in clinical trials research were explored. Furthermore, the chapter probed opinions on potential facilitators to mitigate some barriers and concerns. Importantly, the findings yielded rich information enabling better understanding of underlying obstacles hampering the use of confirmatory ADs. However, despite this strength, the nature and size of the sample limited generalisability of the findings and the design used was not tailored for this purpose. As a result, it is unclear whether the findings apply to a wider audience beyond the interviewed sample. Nevertheless, the findings provided a solid platform to guide further exploration of perceived obstacles.

This chapter therefore, extends the work of Chapter 3 to gauge wider opinions on barriers, concerns, and potential facilitators using follow-up quantitative surveys aimed to generalise the results. These surveys were tailored for UK CTUs, Public Funders, and the private sector to investigate wider perceptions. The results from this chapter provide a strong foundation for the future research of this thesis and beyond. The chapter findings were presented at the 3rd International Clinical Trials Methodology Conference (ICTMC), Statistics for the Pharmaceutical Industry (PSI), and UK CTU Bi-annual Statistics Operations Meeting (Dimairo, Julious, Todd and Nicholl, 2015; Dimairo, Todd, Julious and Nicholl, 2015). The work of this chapter has already been published in Trials open access journal (Dimairo, Julious, Todd, Nicholl, et al., 2015).

This chapter acknowledges the external advice of Dr Tracey Young on the item response modelling approach adopted in Section 4.3.7. In addition to my supervisors, the collaborative contribution and advice of Dr Jonathan Boote is acknowledged.

4.2 Aims

Building on Chapter 3, this Chapter aims to further explore and gauge the wider perceptions on barriers, concerns, and potential facilitators to the appropriate use of confirmatory ADs. The specific objectives are guided by the desire to address a number of research questions based on quantitative surveys.

1) What are the general perceptions of key stakeholders involved in clinical trials research regarding the use of ADs?
2) How are the perceived barriers and concerns ranked in order of importance for prioritisation?
3) How do the perceptions differ between the private and public sectors?
4) What are the types, scope, and prevalence of different confirmatory ADs implemented in practice?
5) What are the potential ways to address perceived obstacles in order to improve the appropriate uptake of confirmatory ADs?

4.3 Methods

Chapter 3 uncovered multifaceted barriers to and concerns about the use of confirmatory ADs, some of which appear to exist across sector and among policymakers and decision-makers. In order to gauge wider opinions, it is therefore imperative to target cross-disciplinary and cross-sector key stakeholders in clinical trials research. This influenced the decision to conduct multiple parallel surveys to address the research objectives stated in Section 4.2.

4.3.1 Study Design and Sampling Frame

Three cross-sectional, parallel, online surveys were undertaken targeting registered UK CTUs, selected UK Public Funders, and the private sector predominantly in the UK. Online surveys were considered only for pragmatic reasons in order to reach out to a wider geographical audience. The selection of sampling frames was influenced by the need to cover the mainstream routes supporting the funding and conduct of confirmatory trials in the UK. It was deemed unnecessary to conduct a separate survey targeting Regulators because of the small number of regulatory stakeholders in the UK. Furthermore, some of these Regulators had already given their perceptions during in-depth interviews.

The questions on surveys were tailored for key stakeholders under consideration as guided by the preliminary results of Chapter 3. For example, some questions were common and phrased consistently across surveys as they pertained to all stakeholders. However, other questions were unique to certain stakeholders – hence, they were only included in relevant surveys.

4.3.1.1 Survey A: Targeted at Registered UK CTUs

The UK CRC (2014) comprises a network of accredited CTUs with adequate expertise to coordinate high quality conduct of clinical trials. Major UK Public Funders, for example NIHR and MRC, require collaboration with accredited CTUs as part of their grant funding policy aimed to enhance the high quality design and conduct of publicly funded confirmatory trials. Hence, these accredited CTUs are fundamental in the UK clinical trials research network. From 2013 to 2014, there were 55 accredited CTUs scattered across the UK (UK CRC, 2014), which were all targeted for the survey.

4.3.1.2 Survey B: Targeted at Selected UK Public Funders

The NIHR plays a key role in publicly funded medical research, contributing around a third of the total UK research funding estimated at over £1 billion between 2013 and 2014 (NIHR, 2015). Within the NIHR, the HTA programme is the largest funding stream, for the commissioning of independent research, primarily confirmatory trials, to investigate clinical and cost effectiveness, and the wider impact of medical interventions tailored for the NHS (NIHR HTA, 2014a). In 2015, the HTA had 4 boards, supported by 5 advisory panels and a priority group. Some of the members of boards and advisory panels and priority groups are publicly contactable (NIHR HTA, 2014b).

There are also other funding streams within the NIHR supporting a smaller proportion of RCTs, for example, the Efficacy and Mechanism Evaluation (EME), and the Research for Patient Benefit (RfPB) streams. Charity organisations such as the Cancer Research UK (CRUK) also play an important role in funding public clinical trials research. The CRUK was selected because it is one of the largest UK charities funding confirmatory trials. In addition, huge opportunities to use adaptive designs are perceived in therapeutic areas such as oncology, as reflected in Section 3.4.3.2 of Chapter 3. A second large UK Charity Funder was approached to take part, but unfortunately the coordinating team were uncomfortable for their boards and panel members to be contacted. Therefore, a survey tailored for Public Funders targeted members of boards and advisory panels for the HTA, EME, RfPB, and CR UK. The diversity in clinical trials expertise of board and advisory panel members, highlighted under Section 4.4.2, offered an opportunity to capture diverse cross-disciplinary perceptions.

4.3.1.3 Survey C: Targeting Selected Private Sector

Although the overall focus of this thesis in on the publicly funded setting, a private sector tailored survey, was viewed important for exploratory comparison with the public sector as a platform for cross sector

collaboration on challenges affecting both sectors. Pharmaceutical and biotech companies and Contract Research Organisations (CROs) are the mainstream platforms supporting clinical trials research in the private sector. Unlike in the public sector, the private sector participants are 'hard to reach' due to non-existence of contact details accessible in the public domain. As a result, only private organisation with direct contacts were approached for pragmatic reasons to complete the survey as described in Section 4.3.5; pharmaceutical or biotech organisations (n=13), and CROs (n=12).

4.3.2 The Rationale for Sample Size Approach

As highlighted in Section 4.2, the key objectives of the surveys are multidimensional in nature without a single 'primary' survey question to base the sample size on. In addition, the surveys were not to assess comparability in perceptions based on hypotheses testing. As a result, a decision was taken to target the reachable 'population' of all target participants. Hence, all target participants for the three parallel surveys were approached and the sample size depended only on the response rates observed.

4.3.3 The Design of Online Survey Instruments

A list of themes on barriers and concerns expressed was compiled based upon the results from in-depth interviews reported in Chapter 4. These themes were grouped depending on whether they pertained to CTUs, private sector, Public Funders or across sector. Key perceived facilitators were included to gauge wider opinions and guide the future direction of this research. Most of the questions were designed in a closed form. Some open-ended questions were included for respondents to add detailed responses where applicable.

Widely accepted Likert Scales (Vagias, 2006) were used to assess the perceptions of respondents on the importance of barriers, concerns towards, and usefulness of, potential facilitators to the appropriate use of ADs in confirmatory trials. The phrasing of questions was consistent across the UK CTU and private sector surveys, with exceptions on occasions when specific questions were unique to a certain sector. The surveys also captured demographics and characteristics of respondents, experiences of ADs, and the historical use of ADs undertaken within UK CTUs and the private sector. The finalised questionnaires for the three surveys designed using SurveyMonkey (2014) are displayed in Appendix 4.1 to 4.3.

4.3.4 Use of Quality Control Measures

Help was sought from individuals with sector experiences of interest in reviewing the draft designs of the survey questionnaires; CTUs (2), private sector (3), Public Funders (3). These individuals were known contacts or referred contacts due to their experiences and ability to provide relevant opinions. The individuals helped to troubleshoot appropriateness, rephrasing, and interpretation of questions. In addition, before the launch of the surveys, all questionnaires were dummy piloted within ScHARR Medical Statistics Group for further checking on easiness and time taken to complete, logical flow, and errors.

4.3.5 Approaching Target Participants

For the survey tailored for UK CTUs, the unit Directors or designated Senior Statisticians were approached to complete an online questionnaire. These were selected because of their in-depth understanding of the organisational, practical and scientific issues within and beyond their CTUs, which may influence perceptions to barriers to and concerns about the use of ADs. The survey permitted only one response per CTU, by either the Director or designated Senior Statistician. Two rounds of invitation emails with a supporting information sheet were sent through the UK CRC registered CTU network of Directors and Senior Statisticians unit representatives. A third round of personalised invitation emails were directly sent to 21 contactable non-responders to the previous two invitation rounds.

The survey tailored for Public Funders targeted chairs, ordinary and lay members of boards and advisory panels. One round of email invitations with a supporting information sheet was sent with the support of some programme coordinators: 4 HTA boards and EME, CRUK and RfPB advisory panel members. Personalised emails were sent to contactable members of the HTA advisory panel members. Programme coordinators strongly advised against multiple rounds of emails. Overall, 212 contactable members were invited to complete an online survey: HTA (n=110), EME (n=20), RfPB (n=40), and CRUK (n=42).

The unavailability of a public accessible domain to contact members and confidentiality restrictions constrained the outreach process for the survey tailored for the private sector. As a result, only organisations with known contacts with the support of my supervisors (Prof Susan Todd and Prof Steven Julious) were approached. Two rounds of direct email invitations were sent to 25 organisations targeting trials Research Leaders or designated Principal or Senior Statisticians. Multiple responses from organisations with more than one trials research groups were permitted where applicable.

All invited participants for the three surveys were given a period of 3 to 8 weeks to complete the relevant tailored online questionnaire.

4.3.6 Research Ethics and Consenting Participants

Ethics approval (0676) described in Section 3.3.5 of Chapter 3 covers this part of quantitative surveys research. However, survey respondents gave their informed consent agreement by responding to a consent question on the leading page made available on all the survey questionnaires (Appendix 4.1 to 4.3). Only responses of those who had agreed to the informed consent statement were included for analysis.

4.3.7 Outline of Statistical Analysis and Reporting

Descriptive statistics with the aid of forest plots and clustered bar charts were produced using Stata (StataCorp, 2014). Rasch modelling for ordered response items using a Rating Scale Model was used to rank the perceptions of respondents in order of importance (or concern) as characterised by the 'difficulty' parameter (Andrich, 1978). The 'difficulty' parameter summarised the direction of the weight of the distribution of perceptions of barriers (or concerns) on an importance (or concern) scale. This analysis was performed using RUMM2030 (RUMM Laboratory Pty Ltd, 2014) and graphs enhanced in Stata 14.1.

The log odds of a respondent selecting a higher category of an item on an importance (or concern) scale over the previous adjacent category was modelled as a function of responder's ability, attached perceived importance (or concern) of an item, and threshold parameters of item categories as (Andrich, 1978)

$$ln\left(\frac{p_{nik}}{p_{ni(k-1)}}\right) = \theta_n - (\delta_i + \tau_k)\,, \qquad 4:1$$

where: p_{nik} is the probability of a respondent n with ability θ_n selecting category k for an item i; $k = 0, 1, 2, \ldots, m$ are the ordered choices and m is the number of item steps; δ_i is the 'difficulty' of item i, which is the importance (or concern) location parameter of interest; τ_k is the threshold parameter corresponding to choice k in item i; θ_n and τ_k are nuisance parameters.

Goodness of model fit to the data was assessed using a conservative Bonferroni correction depending on the number of questions (items) in the model (Bland and Altman, 1995). For example, with 27 items on barriers and 5% nominal level, the model was deemed inadequate to model the data if the type I error for individual items

was less than 0.0019. Differences in perceptions across sector were explored descriptively without any statistical significance testing.

The findings of a review of existing guidance on the reporting of survey research guided the design considerations and reporting of this research (Bennett et al., 2011). Only relevant guidance items were considered to enhance reporting. Presentation focused on numbers and proportions of item responses, estimates of the perceived importance with associated 95% CIs and item ranks.

4.4 Results

In this section, the response rates, demographics, characteristics, and experiences of respondents are described. To address the questions highlighted in Section 4.2, the presentation of the main survey results shall:

1) Rank the perceptions on barriers and concerns in order of importance for prioritisation,
2) Descriptively explore the differences in perceptions between the private and public sector,
3) Gauge the opinions on potential solutions to improve appropriate use of ADs in confirmatory trials.

4.4.1 Response Rates

Of the 55 accredited UK CTUs invited, 30(55%) consented to take part in the survey. A total of 25(46%) CTUs responded to key questions regarding barriers, concerns, and possible facilitators. For the private sector, 25 private organisations were approached to complete the survey. The crude response rate was approximately 68% (17/25). A total of 13(52%) responded to all key survey questions of interest.

For Public Funders survey, a total of 212 members of the public boards and advisory panels were invited to complete the survey: HTA (n=110), EME (n=20), RfPB (n=40), and CRUK (n=42). Of these, 86(41%) members responded to the survey tailored for Public Funders. However, the response rates to questions were variable and just 64(30%) responded to all key questions.

Just 2 CTUs and 6 members of public funding bodies sent feedback through email explaining reasons for their non-participation. The main reason cited was the lack of basic understanding of ADs to contribute meaningfully to the surveys. The implications of this to the generalisability of the results are discussed in Section 4.5.3.

4.4.2 Demographics, Characteristics and Experiences of Respondents

4.4.2.1 UK CTUs Respondents

CTU responders represented a wide geographical area across the UK and covered diverse therapeutic areas of clinical trials research. Overlapping and diverse therapeutic areas of research include oncology (n=13), mental health (n=11), primary care (n=9), public health (n=9), musculoskeletal (n=8), respiratory (n=8), cardiovascular (n=7), diabetes (7), health services (n=7), emergency medicine (n=6), infectious (n=2), rare or orphan diseases (n=2), perinatal medicine (n=1), surgical interventions (n=1), and other (n=3). Figure 4.1 displays the approximate distribution of the study interventions as a percentage of the total number of trials based on complete data reported. In addition, the distribution of the trials requiring regulatory approval (such as from the MHRA, EMA or FDA) beyond standard ethics had a median (IQR) of 50% (16% to 80%) based upon 23 complete responses. Responders were CTU directors 10(33%), designated Senior Statisticians 18(60%), and 2 did not state their role.

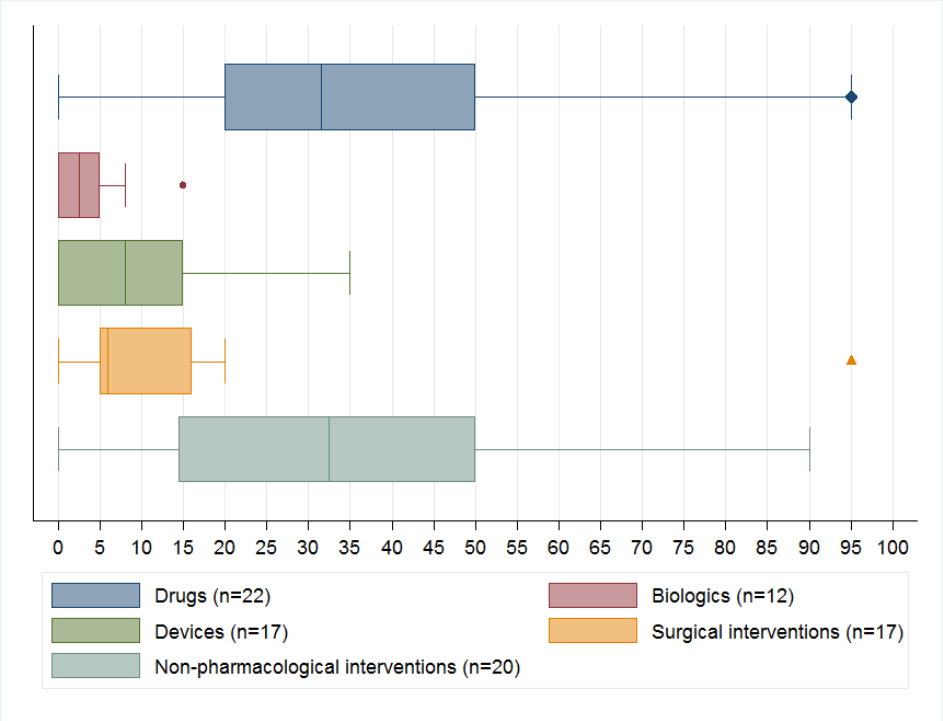

Figure 4.1. The distribution of the nature of interventions investigated by UK CTUs respondents.

Twenty-three (77%) respondents had more than 10 years' experience in clinical trials research. Most respondents (63%) reported moderate personal level of familiarity with the concept of ADs, with just 5(17%) selecting 'very' or 'extreme' familiarity. Appendix 4.4 summarises the perceived level of awareness of types of

confirmatory ADs among CTU research groups along with personal and CTU experiences in the design and conduct of ADs in the confirmatory setting.

4.4.2.2 Private Sector Organisations

The majority of private sector organisations respondents were based in the UK (76%) and just over half (56%) were representative of pharmaceutical companies. Respondents and their representative organisations had varying degrees of experience in the design and conduct of confirmatory ADs as shown in Appendix 4.5.

4.4.2.3 Public Funders Respondents

The majority of respondents (86%) were members of government funded research boards or advisory panels. The respondents' expertise in clinical trials research was diverse and overlapping, representing the typical composition of public funding boards and advisory panels. These included Trial Statisticians 11(13%), Chief Investigators 40(47%), Trial Methodologists 20(23%), Trial Management Experts 6(7%), Clinical Experts 23(27%), Health Economists 9(10%), IDMC 33(38%) and TSC members 24(28%), CTU directors 12(14%), Patient Representatives 7(8%), and other 6(7%). Ordinary members and Chairs or Vice Chairs of funding boards and panels constituted 63(73%) and 10(12%), respectively. Appendix 4.6 details the diverse levels of familiarity with ADs, awareness of types of confirmatory ADs, and reviewing and commissioning experience of AD research proposals, of respondents to the survey tailored for Public Funders.

4.4.3 Perceptions on Barriers to the Use of Confirmatory Adaptive Designs

4.4.3.1 Pertaining to UK CTUs

Figure 4.2 displays ranked perceptions of CTU respondents on important barriers. Here, the estimates of the perceived relative importance parameter with associated 95% CIs estimated using a Rating Scale Model for ordered response outcome using equation (4:1) are presented. The estimates indicate the direction of the weight of the distributions of respondents' perceptions on an importance scale. The smaller or more negative the relative importance parameter the larger the proportion of respondents who viewed an item as an important barrier compared to other items. Appendix 4.7 provides detailed supplementary summary data of ordered item responses on perceptions.

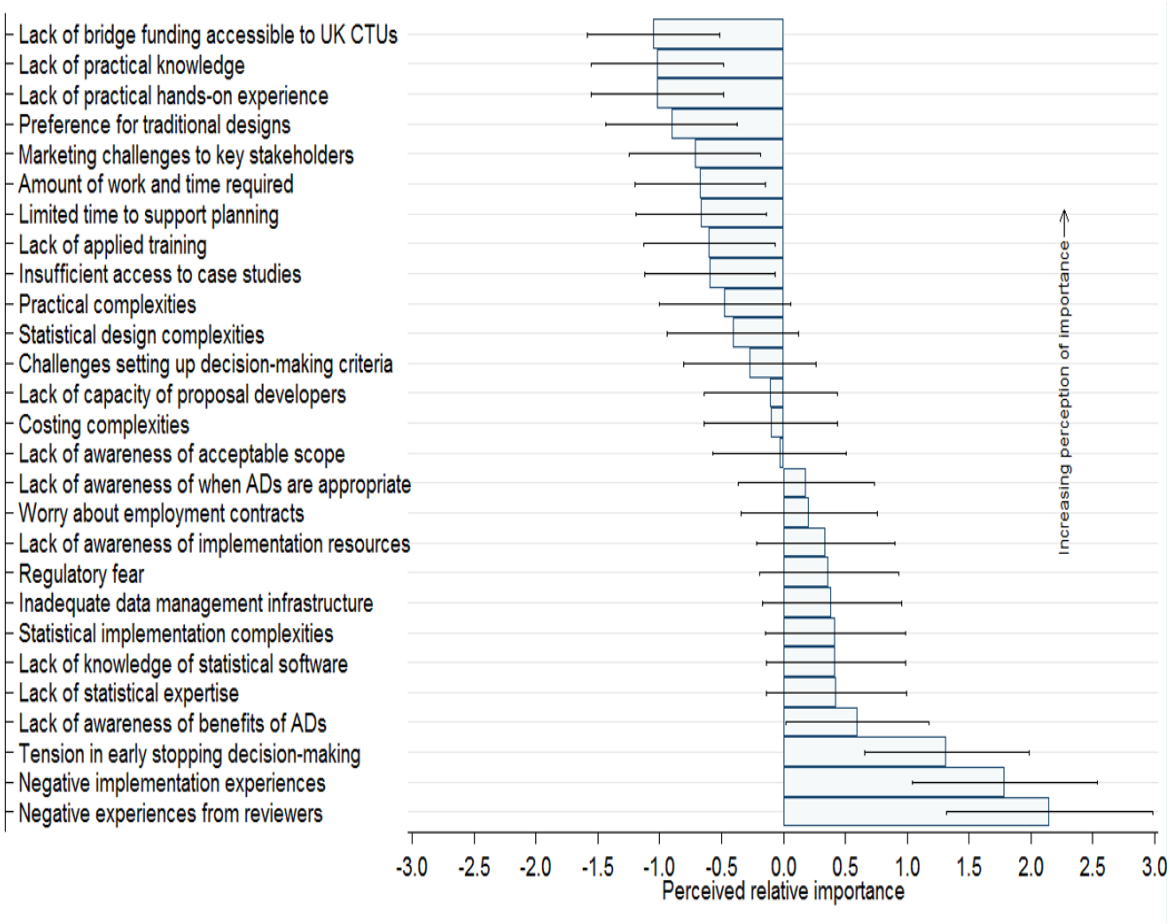

Figure 4.2. Ranked perceptions of UK CTUs respondents on important barriers.

The top-ranked obstacle reported by CTUs was the lack of funding support accessible to them to aid the design development work of complex and time-consuming ADs. This was reported by 8(32%) and 12(48%) respondents as an 'extremely' and 'at least moderately' important barrier, respectively. The lack of practical implementation knowledge and hands-on experience were jointly the second leading barriers with 6(24%) and 15(60%) of respondents reporting them as 'extremely' and 'at least moderately' important barriers. The opinions of respondents suggested that research teams within CTUs have a strong preference for traditional mainstream designs, which they know well, and feel uncomfortable supporting ADs, even when appropriate. Just 3(12%) respondents did not view preference for traditional designs an important barrier.

Thirteen (52%) respondents reported difficulties faced by researchers in marketing ADs to key stakeholders in clinical trials research (such as Clinical Collaborators, Funders, and Regulators) as an 'at least moderately' important barrier. The amount of time and effort required to support the design of ADs, and time constraints relative to competing priorities of traditional mainstream designs was reported as 'at least an

important' barrier by 12(48%) respondents. Among the leading barriers reported was the lack of applied training coupled with insufficient access to case studies on previously undertaken ADs to facilitate practical learning and successful implementation.

The middle ranked reported barriers were; associated statistical complexities during design such as simulations work; practical complexities during implementation; difficulties faced by Clinical Trialists in setting up acceptable planned decision-making criteria to guide the adaptation process; and the dearth of proposal developers with knowledge to support ADs.

The lack of statistical knowledge of ADs and knowledge of statistical software for implementation were reported amongst the least important obstacles. Barriers reported as 'not at all' important by many respondents were: negative experiences based on Funders' or Reviewers' comments (76%); negative implementation experiences (76%); early stopping decision-making tensions among key decision-makers (60%); and lack of awareness of benefits of ADs (48%).

4.4.3.2 Pertaining to Private Sector Organisations

Figure 4.3 shows the ranked perceptions of private sector organisations on barriers. Appendix 4.8 provides supplementary summary data on ordered perceptions with detailed description of the meaning of barrier items presented in Figure 4.3. Marked with red diamonds are barriers more prominent in the private sector compared to CTUs (Figure 4.2 versus Figure 4.3). On the other hand, marked with blue squares are barriers reflected by CTUs as more prominent compared to the private sector.

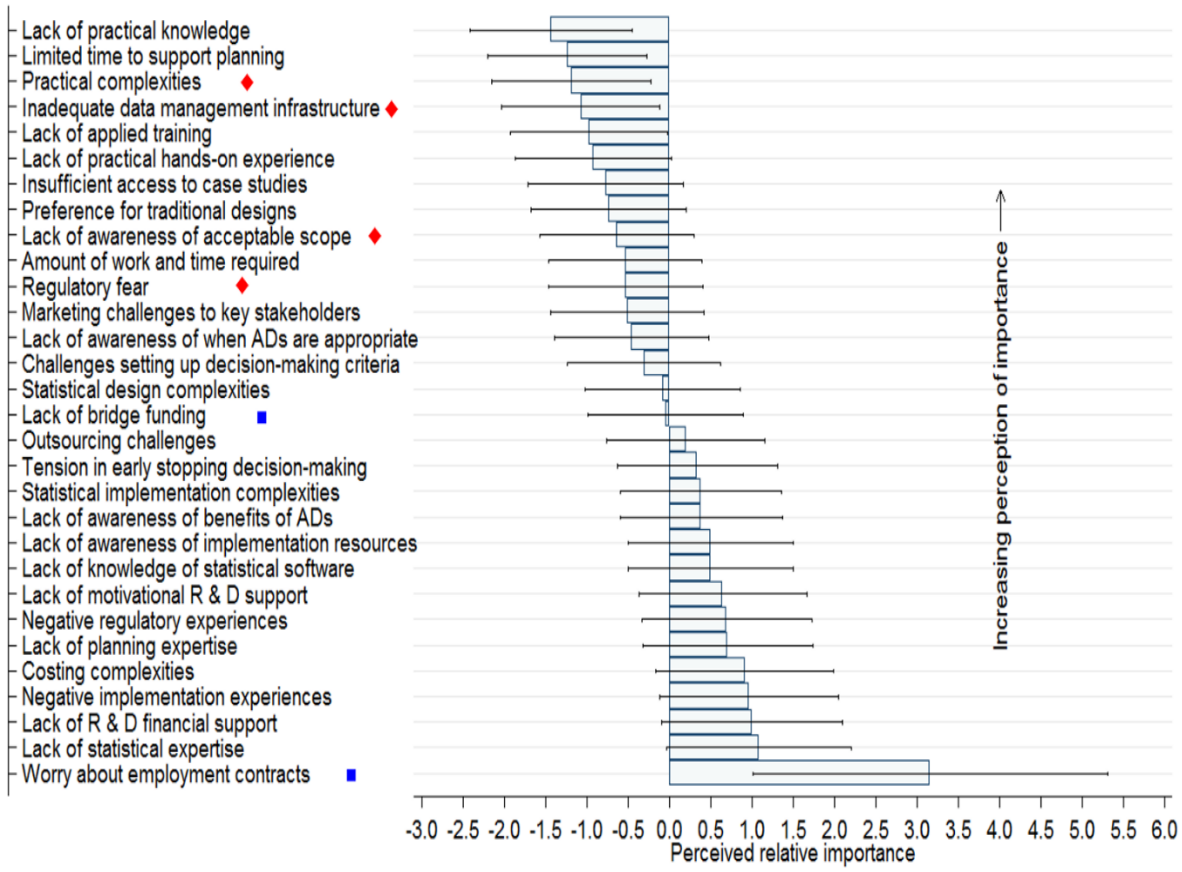

Figure 4.3. Ranked perceptions of private sector organisations on important barriers.

In general, the perceptions of CTUs and the private sector on the importance of a number of barriers were consistent, however, there are a few exceptions that are marked in Figure 4.3. For example, perceived complexities during practical implementation, inadequate data management infrastructure, and fear of risking regulatory approval appeared very prominent in the private sector. In contrast, the lack of bridge funding to support developmental design work and worry about research staff employment contracts when trials are stopped early were highly and middle rated by CTUs, respectively.

The leading ranked barriers reported as ‘at least moderately’ important were the dearth of practical implementation knowledge 9(69%), time limitations to support the planning of complex ADs relative to competing priorities of mainstream designs 6(46%), and associated practical complexities during implementation of ADs 9(69%). In addition, inadequate data management infrastructure to support execution 5(42%), the dearth of applied training to facilitate practical implementation 9(69%), the lack of hands-on practical experience

8(62%), limited access to case studies of the few undertaken ADs to facilitate practical learning 6(46%), and research teams being more comfortable with traditional mainstream designs 8(62%) were among leading barriers.

Barriers reported as unimportant by a sizable number of respondents were:

- The lack of awareness of implementation resources 6(46%),
- The lack of knowledge of existing AD-related statistical software 6(46%),
- The lack of motivational support from Research and Development (R & D) 6(46%),
- Negative regulatory experiences 5(38%),
- The lack of expertise to support planning 5(38%),
- Costing complexities during planning 6(46%),
- Negative experiences during implementation 7(54%),
- Insufficient R & D financial support to invest in AD infrastructure 8(62%),
- The dearth of statistical expertise to support ADs 8(62%),
- Worry about staff employment contracts when trials are stopped early 10(77%).

4.4.3.3 Pertaining to Public Funders

As evident in Figure 4.4, respondents ranked the importance of most barriers considered with a small degree of differentiation between them. Appendix 4.9 provides supplementary summary data of respondents' perceptions and detailed description of barrier items presented in Figure 4.4.

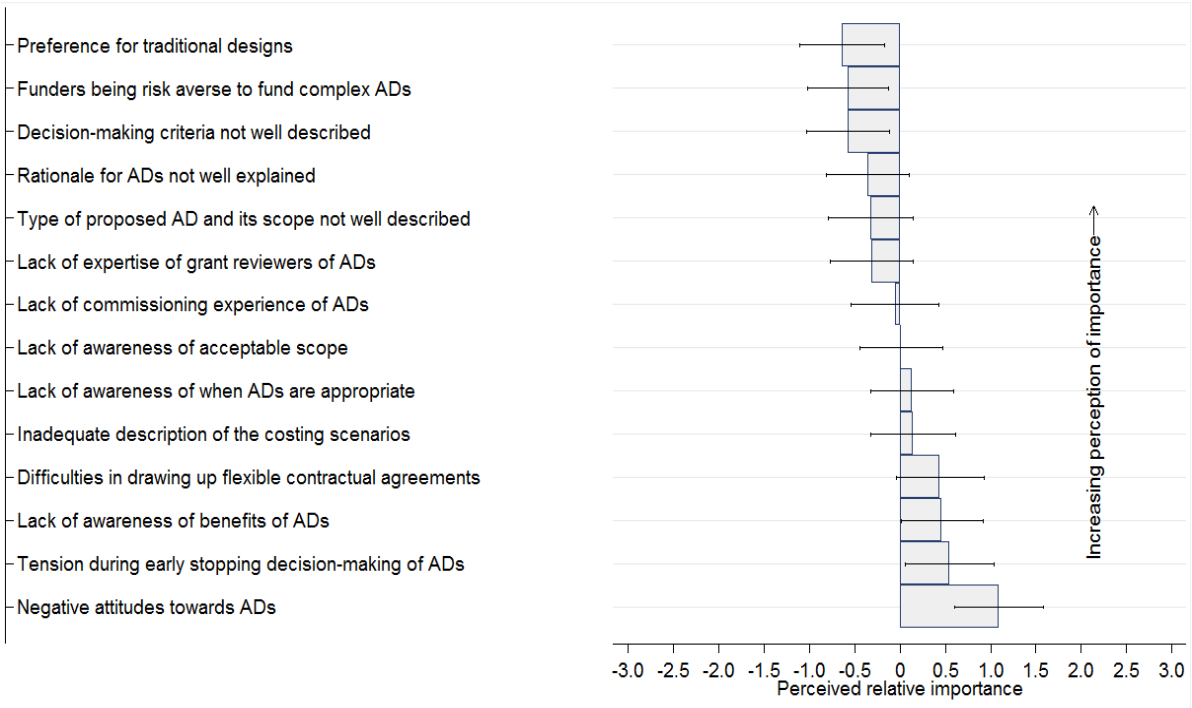

Figure 4.4. Ranked perceptions of Public Funders respondents on important barriers.

The preference of Public Funders for traditional mainstream designs over ADs and their risk-averse attitude to fund projects associated with a high degree of financial uncertainty were among the leading barriers reported. The preference for traditional designs was selected by 40(63%) and 19(30%) respondents as 'at least moderately' and 'extremely' important barriers. The inadequate description of the rationale for the proposed AD and decision criteria to guide the adaptation process in grant proposals by Clinical Trialists were reported as 'at least moderately' important impediments during the review process by 33(53%) and 38(60%) respondents, respectively. The lack of expert proposal Reviewers to provide advisory support to Funders during the grant review and commissioning process was selected by 32(55%) and 11(19%) respondents, as a 'moderately' or 'extremely' important barrier, respectively.

Thirty-four (55%) respondents selected the inadequate description of proposed ADs and their scope in grant applications by researchers as at least a 'moderately' important barrier. Among middle ranked barriers were the lack of; commissioning experience of ADs-related research, awareness of acceptable scope of ADs, and when they are appropriate in confirmatory trials. The challenges faced by Public Funders in drawing up contractual agreements which are suitable to support ADs was viewed by 38% and 10% respondents as a 'moderately' or an 'extremely' important barrier, respectively.

The least ranked barriers reported were tensions during early trial stopping decisions among key decision-makers (32%) and negative attitudes towards ADs by some of the funding boards and panel members (26%). Although least ranked, these were selected by 28(32%) and 16(26%) respondents as at least 'moderately' important barriers, respectively.

4.4.4 Cross-sector Relating to Concerns on the Use of Confirmatory Adaptive Designs

In general, respondents were least concerned about early stopping of trials because of futility. Significant proportions of respondents were 'not at all' or 'slightly' concerned about futility early stopping; 14(56%) CTUs, 8(62%) private sector organisations, and 39(57%) Public Funders. Appendix 4.10 provides detailed summary data of cross-sector perceptions on concerns about the use of confirmatory ADs.

Concerns about the robustness of ADs in decision-making and acceptability of findings to change practice when trials are stopped early were slightly more apparent among Public Funders respondents. These were reported as 'moderate' and 'extreme' concerns by 35% and 38% Public Funders, 28% and 20% CTUs, and 31% and 23% private sector organisations, respectively.

In the private sector, top leading issues selected as 'moderate' or 'extreme' concerns were early stopping for non-inferiority (39%), impact of stopping trials early on secondary trial objectives (39%), fear of introducing operational bias (30%), and early stopping for efficacy (38%). Twenty-two percent of Public Funders and 28% CTU respondents reported the fear of introducing operational bias as at least a 'moderate' concern. Respondents among Public Funders (19%), CTUs (20%), and the private sector (31%) expressed the potential change in the population during implementation of an AD and its implications on the interpretation of findings as at least a 'moderate' concern.

4.4.5 Cross-sector Perceptions of Possible Facilitators

In general, there was consistency in perceptions of cross-sector respondents on the usefulness of proposed key facilitators to enhance the appropriate use of confirmatory ADs. Figure 4.5 shows opinions of cross-sector respondents: Public Funders (n=64), CTUs (n=25), and private sector organisations (n=13).

There was overwhelming support for a troubleshooting toolkit of specific questions which Clinical Trialists need to ask themselves when considering different types of ADs; 23(92%) CTUs, 61(95%) Public Funders, and 12(92%) private sector organisations viewed it as 'somewhat' useful. Furthermore, there was

compelling support for the accessible publication of case studies of implemented ADs focusing on aspects such as design and rationale, implementation, regulatory and statistical challenges, lessons learned, and facilitators as 'very useful' to Clinical Trialists. Twenty-three (92%) CTU and 57(89%) Public Funder respondents reported the need for a consensus guidance document on acceptable scope of ADs, addressing issues tailored for publicly funded confirmatory trials, would be 'at least somewhat' useful.

The development of a CONSORT statement tailored for ADs was selected as 'at least somewhat' useful to enhance transparency and completeness in the conduct and reporting of ADs by: 56(88%) Public Funders, 23(92%) CTUs and 13(100%) private sector organisations. Forty-six (72%) Public Funder respondents held the view that refresher training of funding boards and panel members to improve familiarity with AD-related issues could help them in the reviewing and commissioning process.

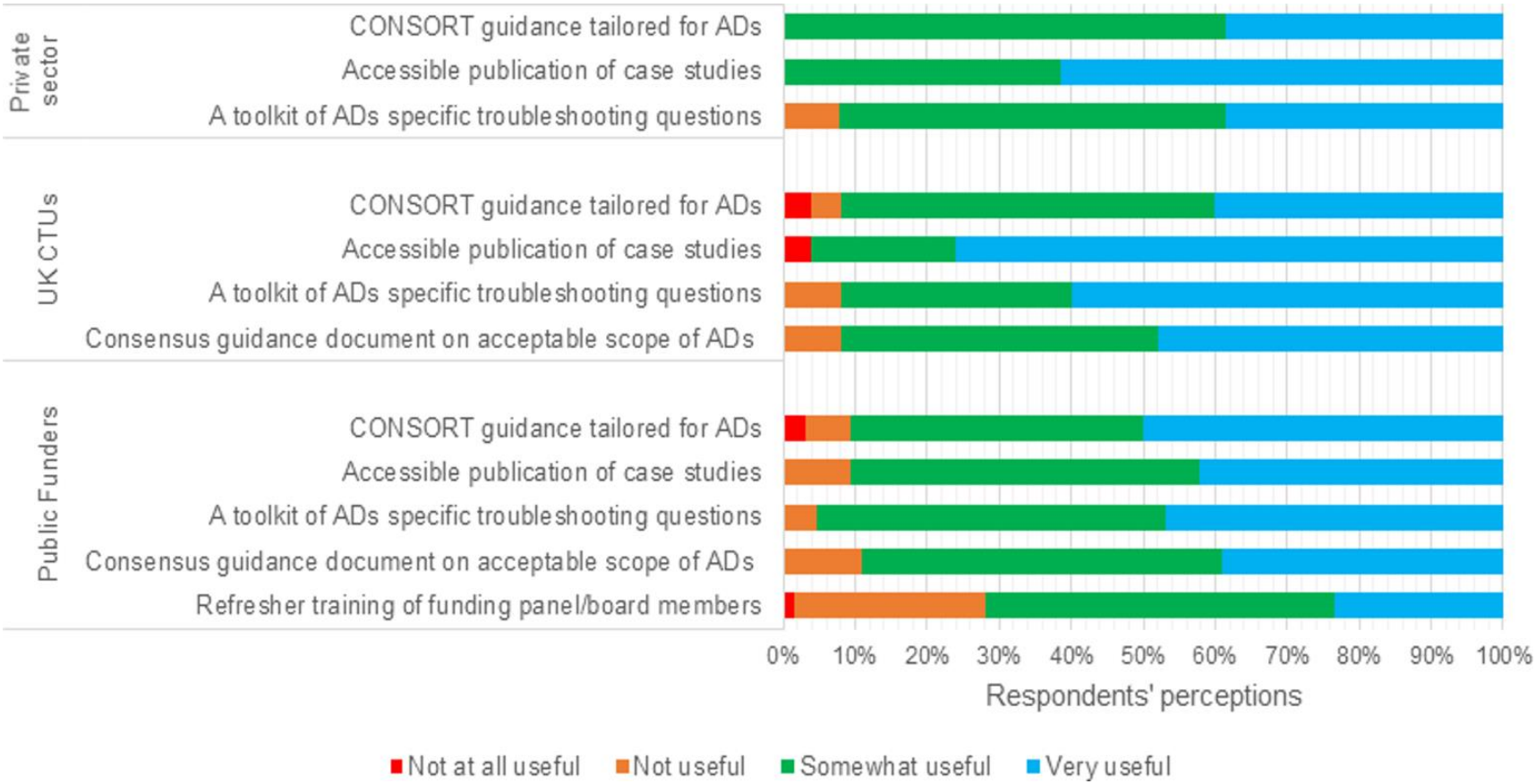

Figure 4.5. Cross-sector perceptions on proposed key facilitators.

4.4.6 Organisational Priorities on Adaptive Design-Related Aspects

When respondents were asked to rate the level of organisational priority they give to the use of confirmatory ADs and/or research on related methods within the next 5 to 10 years, 15(50%) CTUs selected it as a 'medium priority', and just 3(10%) as a 'high priority'. In contrast; 5(29%), 4(24%) and 4(24%) private sector organisations selected it as a 'medium priority', 'high priority', and an 'essential priority', respectively.

Only 2(7%) CTUs and 3(18%) private sector organisations reported having an AD-related Working Group within their organisation. The desire to 'definitely consider' the use of confirmatory ADs in the future, when appropriate, was expressed among 16(53%) CTUs and 11(65%) private sector organisations.

Forty-five (55%) Public Funder respondents rated their boards or panels priorities on funding themes on confirmatory ADs in the next 5 to 10 years as at least a 'medium priority'; 19(22%) as 'high priority' and only 4(5%) as an 'essential priority'. As a funding board or panel, only 26(30%) reported that they had previously recommended funding a confirmatory AD-related grant proposal, however, 26(30%) did not respond to the question. When asked whether they would consider recommending a confirmatory AD-related grant proposal for funding in the future when appropriate to address research question(s); 42(49%) indicated that they 'would definitely consider', 21(24%) 'might or might not consider', 1(1%) 'would not consider', and 22(26%) did not respond to the question.

4.4.7 Survey Results on the Application of Confirmatory Adaptive Designs in the UK

The submission of historical AD-related grant proposals for funding considerations was reported by 13(43%) CTU respondents. Historical application of at least some type of confirmatory AD was reported by 27% (8/30) of CTUs and 47% (8/17) of private sector respondents.

Table 4.1 summarises detailed types and frequency of reported confirmatory ADs stratified by sector of application. It appears the application of ADs is not widespread across sector organisations. However, there appeared to be a small number of leading organisations that reported frequent use of particular types of ADs. The most frequently used ADs were (CTUs; private sector): SSR (23%; 41%); standard two arm GSD (23%; 47%); futility analyses using stochastic curtailment methods (27%; 29%); and operational seamless 2/3 design (23%; 35%). Moreover, ADs with some form of futility stopping, such as dropping futile treatment arms or stopping

trials, appeared most popular, consistent with cross-sector and multidisciplinary receptiveness uncovered in Chapter 3.

The scope of the SSR employed appeared to be varied and the use of operational seamless 2/3 design seems to be gaining traction across sectors. Some ADs such as information based GSD and standard GSD with SSR appeared rarely used, particularly in the UK public sector. Some respondents who reported the use of particular types of ADs did not disclose the approximate number of such ADs undertaken. Thus, the total number of applied ADs may be a slight underestimate.

Table 4.1. Distribution of the type of confirmatory adaptive designs stratified by sector of application.

Type of AD and its description	UK CTUs			Private sector		
	Number of CTUs	Number of trials	Missing responses†	Number of Organisations	Number of trials	Missing responses†
SSR	7(23%)			7(41%)		
Blinded SSR allowing for increase only	4	4	1	2	11	-
Blinded SSR allowing for increase or decrease	2	1	1	2	3	-
Unblinded SSR allowing for increase only	2	5	1	2	10	-
Unblinded SSR allowing for increase or decrease	2	5	1	-	-	-
Unblinded SSR based on promising zone concept	2	-	2	3	10	-
Standard two arm GSD	7(23%)			8(47%)		
Stopping early for futility only	2	7	1	3	26	-
Stopping early for efficacy only	1	-	1	-	-	-
Stopping early for efficacy or futility	4	6	0	3	8	1
Stopping early for safety only	4	2	2	2	5	1
Stopping early for safety or futility	2	2	0	1	5	-
Stopping early for non-inferiority only	-	-	-	1	-	1
Futility analysis based on stochastic curtailment	8(27%)			5(29%)		
Based on CP	5	7	1	3	3	2
Based on PP	2	1	1	1	-	1
Based on CI of the interim effect	3	3	1	-	-	-
Operational seamless 2/3 design	7(23%)			6(35%)		
Dropping futile treatment arms in phase 2 only	5	5	1	3	3	1
Selecting only one promising treatment in phase 2 only	1	-	1	3	2	1
Selecting multiple promising treatments in phase 2 only	-	-	-	2	3	1
Other	2	-	2	-	-	-
Inferential seamless 2/3 design	2(7%)			3(18%)		
Dropping futile treatment arms in phase 2 only	2	1	1	1	2	-
Adding or dropping futile treatment arms in phase 2 only	1	.	1	-	-	-
Strictly phase 3 MAMS design	2(7%)			2(12%)		

(Table continued)						
Stopping trial for efficacy or futility or dropping futile intervention arms	1	1	-	-	-	-
Information based GSD	-	-	-	4(24%)	3	2
Standard GSD with SSR	-	-	-	1(6%)	-	1
Patient enrichment or subgroup selection	2(7%)	-	2	2(12%)	-	2
Response adaptive randomisation	2(7%)	2	-	2(12%)	2	1

CP: Conditional Power; CTU: Clinical Trials Unit; CI: Confidence Interval; GSD: Group Sequential Design; MAMS: Multi-Arm Multi-Stage; SSR: Sample Size Re-estimation; PP: Predictive Power. † Some responses on the approximate number of trials undertaken were missing.

4.5 Discussion

It is important to emphasise that ADs are not appropriate for every trial. When contemplating the use of ADs, logistical as well as statistical considerations should be made on a trial-to-trial basis. These considerations include the accrual of the primary endpoint data in relation to expected recruitment rate, the rationale for choosing the design, feasibility or practicalities of implementing the design, and potential benefits versus additional complexities in implementation. Some of the considerations were highlighted in Chapters 2 and 3.

4.5.1 Main Findings and Interpretation

This chapter details existing barriers and concerns about the appropriate use of ADs in the confirmatory setting perceived by cross-sector key stakeholders. Stakeholders' perceptions about barriers are largely consistent across sectors, with some exceptions that reflect differences in organisations' funding structures, experiences, capacity, and nature of investigative interventions. This highlights the need for cross-sector collaboration to address some of the roadblocks. There is cross-sector and multidisciplinary interest in the use of ADs when appropriate to answer research question(s). However, the potential benefits of ADs can only be realised when key obstacles to their use are adequately addressed.

The most important cross-sector barriers are connected to the lack of practical implementation knowledge and hands-on experience. This appears intertwined with the lack of applied training and paucity of implemented case studies to facilitate practical learning and problem-solving. This is also linked to the amount of time and effort required for adequate planning. Moreover, both the private sector and UK CTUs voiced concerns that they are under immense pressure to deliver on other competing priorities based on simpler mainstream designs. As a result, researchers have limited time to support complex ADs, even when appropriate. Importantly, the lack of funding support accessible to UK CTUs in the form of small grants to aid design developmental work of time-consuming and complex ADs is the major stumbling block to the use of ADs.

There is a strong need on the part of Public Funders to address sources of funding accessible to UK CTUs wishing to support the use of relevant, complex and time consuming ADs. For example, even though the MAMS design appears to be efficient in evaluating multiple interventions in one trial, with options to drop futile arms, it requires in-depth statistical simulations and time commitment at the design stage. This developmental stage is often unfunded, with researchers taking risks due to the uncertain future success of research grant applications.

Given the high risk involved, CTUs may be reluctant to support such ADs even when they are more relevant to answer research question(s) efficiently. Although the NIHR provides infrastructure funding accessible to accredited CTUs (NIHR, 2014a), it is often used for other purposes, such as meeting contractual obligations of staff who may not receive funding between studies. There is an opportunity for the NIHR and MRC to create a small funding stream to support the planning of time-consuming designs provided the research questions meet their priority needs, and there is a strong design rationale. The funding should be conditional on open access publication of design-related material such as software programs to enhance the planning of future related trials. Another recommendation is to encourage the use of ADs which are simple to implement within the existing scope of public funding models for fixed sample size designs, such as SSR and futility analyses.

As highlighted, the most important barriers reported are associated with the lack of practical knowledge and experience among key stakeholders. It is well acknowledged that there have been numerous theoretical developments in ADs and more are needed to address unknowns. However, what is lacking is a translational framework to enhance the use of ADs in practice. Accessible publication of case studies of 'successful' and 'unsuccessful' ADs with related materials is encouraged. These publications should encompass aspects such as: rationale and design; statistical and practical challenges, and how they are resolved; implementation resources; lessons learned; regulatory, data management and communication hurdles, and how these are resolved; and other facilitators to successful implementation. Learning from researchers or organisations who are routinely implementing ADs is paramount. Importantly, there is a need for a focal group of practical experts publicly funded to support CTUs with little practical expertise wishing to implement ADs. Such experts should provide practical training on ADs accessible to UK CTUs. Although an initiative exists through the ADWG of the MRC NHTMR AD (MRC, 2014), some Trialists interviewed in Chapter 3 viewed it as being more theoretically oriented.

Researchers who receive public funding for AD-related methodological research are strongly encouraged to produce open access resources such as free-to-use software or codes to implement the methods developed. Thus facilitating the application of the methods. An additional recommendation is that CTUs receiving AD-related bridge or research funding should form a compendium of case studies for publication. Open access publication of research outputs and resources such as in monographs is important. This could be a useful resource aimed at reducing research waste and improving the appropriate conduct of adaptive trials. Such knowledge-sharing would be helpful for applied knowledge transfer. Collaboration between CTUs on ADs is strongly encouraged.

Concerns regarding the robustness of ADs in decision-making and their credibility to change clinical practice when trials are stopped early are real and should be addressed. Even though there are multi-dimensional aspects to these concerns, transparency and adequate reporting of trial conduct may mitigate some of the concerns. For example, suboptimal reporting of measures to minimise operational bias during the trial conduct, and inappropriate use of statistical methods to obtain unbiased trial results may influence consumers of research findings to view results from ADs with suspicion. The need for a CONSORT statement tailored for ADs to enhance their reporting and conduct has been overwhelmingly supported across sector by key stakeholders. Case studies of ADs investigating a wide range of interventions published in 'high impact' journals and their influence on clinical practice may also help to convince sceptical research consumers.

Like any new methods, the use of ADs in confirmatory trials is bound to raise anxiety among some researchers. Some of this anxiety could be alleviated by a cross-disciplinary, consensus guidance document on ADs well-crafted to address pertinent issues in confirmatory trials. For example, research on complex interventions has gone through a similar phase, but the emergence of related guidance documents (Craig et al., 2008) improved researchers' receptiveness towards their conduct. There is an equally important need for a troubleshooting toolkit with pertinent design-specific questions which Trialists need to ask themselves when considering ADs. This would facilitate the appropriate use and adequate planning of adaptive trials.

It is fundamental for Clinical Trialists to provide adequate explanation or description of aspects related to the proposed AD to key stakeholders such as Reviewers, Funders, Collaborators, and Regulators. This encompasses the rationale in relation to research objectives, potential benefits compared to mainstream designs, scope, decision criteria to guide the adaptation and decision-making process, variable costs and trial durations, measures to minimise operational bias and control of statistical properties (type I error rate, power and unbiased results), among others. As highlighted in Chapter 3, Public Funders and regulators are receptive to the appropriate use of ADs. The fear of risking regulatory approval does not necessarily reflect regulatory perspective, but is mostly an artefact of inadequate description of the proposed AD and its suitability to address the research question(s), among other aspects. There are encouraging indications that regulatory receptiveness to appropriate use of ADs is positive, with improving awareness and experiences, particularly with respect to scientific advice and review of AD proposals (Elsäßer et al., 2014; FDA, 2015; Lin et al., 2015). Public funders are also encouraged to modify their grant application forms to facilitate adequate description of AD-related aspects. This could be achieved by allowing Clinical Trialists to add specific relevant AD material as appendices.

The preference for traditional designs over ADs expressed by Public Funders appears to be connected to the lack of knowledge and understanding reflected in Chapter 3. Periodic "refresher training" of public funding boards and panel members prior to their reviewing and commissioning meetings may help alleviate the lack of awareness of the acceptable scope of ADs, when they are appropriate, their benefits in confirmatory trials, and other AD-related issues. Furthermore, the experience of funding boards and advisory panels can only be improved when researchers put forward more appropriate AD-related grant proposals for consideration. A positive change in attitudes and receptiveness towards appropriate use of ADs by Public Funders highlighted in Chapter 3 is an encouraging opportunity, which should be communicated to and exploited by researchers.

The challenges faced by researchers in developing widely-acceptable decision-making criteria at the design stage to inform the adaptation process can be alleviated through multidisciplinary engagement and discussions during planning. This process should include close discussions among key stakeholders such as Trial Statisticians, Clinicians, patient representatives, clinical peer advocate groups, and Regulators.

4.5.2 Relating Findings to Existing Literature

Authors investigated the use of ADs and perceptions on barriers regardless of trial phase in the USA through surveys of; 17 private sector organisations (Quinlan et al., 2010) and 17 private sector organisations and one academic institution (Morgan et al., 2014). The major barriers reported by these surveys include the preference for mainstream designs, fear of risking regulatory acceptance, lack of education or lack of knowledge about adaptive methods, and extra time and resources required for planning as major perceived barriers. Some reported practical or technical barriers were design specific. Jaki (2013) also explored the use of ADs and Bayesian methods in early phase trials through a cross-sectional survey of registered UK CTUs, predominantly surveying Statisticians. Jaki attributes the poor uptake of these methods to five key barriers: the lack of software, clinical investigators insisting on preferred methods, lack of expertise, inadequate funding structure, and time required for trial design.

In summary, the preference for mainstream designs, additional time and effort required during planning, and inadequate funding structure to support design work reported in the most recent related research are consistent with the results of this chapter. The meaning of the lack of expertise or lack of education reported in the related literature is unclear. However, this thesis found that this relates to the lack of practical knowledge and applied training and less to statistical theory. The fear of risking regulatory acceptance appears to be consistent with

previous research, however, it is more pronounced in the private sector than the public sector. Some previous findings on less pronounced barriers or concerns such as the lack of statistical software and fear of introducing operational bias contradict this chapter's results. This may be partly explained by the differences in considered trial phases and related nature of adaptive methods.

4.5.3 Strengths and Limitations

The design of the surveys was built upon robust, in-depth interviews of cross-sector and cross-disciplinary key stakeholders reported in Chapter 3. Furthermore, this research appears to be the first to formally explore the perceptions of public funding boards and advisory panels on the use of confirmatory ADs. Therefore, the findings add to the available knowledge and have also uncovered concerns which need addressing to unlock potential benefits of confirmatory ADs.

The key limitation of this research is that the results are based upon moderate response rates across sectors. In other recent surveys of registered UK CTUs observed response rates of 38% (Bower et al., 2014) and ranging from 25% to 67% (Tudur Smith et al., 2014) were found. The number of responders to the private sector survey was similar to a previous survey by the ADWG (USA), although they achieved a 100% questionnaire response rate over a one-year period (Quinlan et al., 2010). Morgan et al (2014) reported a response rate of just 20% for a related survey dominated by the private sector (~94.4%). So the response rate of the surveys described here are consistent and even better than some of the cited surveys.

The private sector results are based only on organisations with known direct contacts. However, the private sector survey was for complementary purposes. Similarly, the response rates to the Public Funder survey and some related questions were low to moderate. Furthermore, the most uncontactable members of boards and advisory panels were Physicians. There seems to be little literature on response rates using this sampling frame for comparability. The decision-makers and policymakers seem to be a difficult group to reach and achieve satisfactory response rates due to their busy schedules.

In view of the aforementioned limitations and the few reasons for non-participation reported in Section 4.4.1, it is likely that non-responders are different from responders. For example, non-responders seem more likely to be unfamiliar with ADs or to not view ADs as a priority. Hence, some of our findings on barriers such as on the lack of awareness of; ADs, opportunities and acceptable scope in the confirmatory setting, and lack of statistical expertise pronounced during in-depth interviews reported in Chapter 3 are most likely to be

underestimated. Most respondents to CTU and the private sector surveys were designated Senior Statisticians, who may be more familiar with statistical aspects of ADs. The survey results on historical application of ADs were limited by the moderate response rates observed both in the private and public sector. Finally, the survey results are not immune to recall bias and had some missing responses regarding the number of certain applied ADs. Thus, the findings may provide a conservative picture on barriers and concerns, and the application of ADs in practice.

4.5.4 Implications for the Work Described in the Remainder of the Thesis

The remaining research described in this thesis focuses partly on addressing the lack of implementation knowledge and degree of conservatism influenced by concerns about robustness of ADs in decision-making and acceptability to change clinical practice. Chapter 5 reviews and presents the case studies of undertaken confirmatory ADs. Chapter 6 investigates the appropriateness of the current CONSORT guidance in enhancing transparency and adequate reporting of ADs, and proposes recommendations. Chapter 7 uses retrospective planned case studies to illustrate the design of, related considerations to, and opportunities to use ADs. Robustness of ADs and any limitations are demonstrated given that the results of completed trials are already known. Key lessons learned are reflected to help the planning of future related trials. Building on Chapter 7, the planning of certain ADs is demonstrated using actual grant applications submitted to the Public Funders and lessons learned are discussed in Chapter 8. Chapter 9 concludes the thesis with an overall discussion, recommendations for best practice, and areas of future research beyond this thesis.

Chapter 5. Review of Case Studies of Confirmatory Adaptive Designs

5.1 Introduction

In Chapter 4, wide perceptions of key stakeholders on roadblocks to the use of confirmatory ADs were investigated thorough multiple cross-sector quantitative surveys. The chapter also surveyed wide opinions on key potential facilitators to some barriers and concerns uncovered in Chapter 3. One of the themes from stakeholders' perceptions highlighted in Chapter 4 is the need for case studies of ADs. This is because case studies would be an important resource to bridge the gap in applied knowledge and raise awareness of opportunities and scope of confirmatory ADs. In addition, the case studies may help to demystify ADs and reduce the fear associated with their use among some key stakeholders.

This chapter reviews and describes the examples of applied confirmatory ADs reported in the literature. Furthermore, the chapter also provides foundation for subsequent work to be reported in Chapter 6. In addition to my supervisors, the chapter acknowledges the collaborative support of three other researchers: Isabella Hatfield, an NIHR Research Methods Intern; Laura Flight, NIHR Research Methods Fellow; and Annabel Allison, Sheffield CTRU Statistician. They contributed during the review process for quality control and to ensure that the work met peer review publication standards. The work was lead and supervised by myself and has been published in Trials journal (Hatfield et al., 2016).

5.2 Aims and Objectives

In an endeavour to improve the appropriate use of confirmatory ADs, this chapter aims to identify case studies of registered ADs applied in the confirmatory setting and highlight their type and scope. Specifically, the objectives of the chapter to fulfil these aims are to explore the prevalence and characteristics of confirmatory ADs applied both in the public and private sectors. Furthermore, to examine the most common therapeutic areas, where particular types of ADs are being used and trends in their usage. Importantly, to explore the adequacy of ClinicalTrials.gov (1997) in identifying registered ADs.

5.3 Methods

This section describes the approach undertaken to conduct a cross-sectional audit study to address the highlighted aims and objectives. The section commences with an explanation of the reasons for identifying case studies via the clinical trials registers. The screening process of eligible trials and characteristics of interest are described as well as the description of the process adopted to examine the adequacy of ClinicalTrials.gov in identifying ADs.

5.3.1 The Rationale for the Literature Search

The chapter focuses on identifying ADs registered on clinical trials registers and databases for a number of reasons. Reporting or publication bias is a well acknowledged phenomenon in clinical trials research (Dickersin et al., 1987; von Elm et al., 2008; McGauran et al., 2010). Prospective registration of trials using a number of platforms has been highlighted as a solution to minimise publication bias (Song et al., 2010). For example, publication bias found in peer reviewed journals where positive findings of 'successful' trials are more likely to be published compared to those with 'negative' findings is minimised because trial registration is now mandatory (Hopewell et al., 2009). Furthermore, clinical trials registers offer opportunity to reduce the time-lag between trial commencement and publication, which takes many years for published trials after completion (Hopewell et al., 2007). Importantly, the identification of case studies via clinical trials registers offers an opportunity to explore the efficiency of clinical trial registers in capturing ADs and propose appropriate recommendations.

There are a number of international, regional, and national clinical trials registry platforms such as the World Health Organisation International Clinical Trials Registry Platform (WHO ICTRP) (2004), ClinicalTrials.gov (1997), and EU Clinical Trials Register (EU CTR, 2004) that have been launched. The next section describes the approach used to guide the choice of which data sources to use for the review.

5.3.2 Scoping Exercise to Troubleshoot the Literature Search

A comprehensive scoping exercise was undertaken to test the practicalities of the review and efficiency of the search algorithm using the WHO ICTRP (2004). Initially, this registration platform was chosen because it is international and linked to several national and regional registers, including ClinicalTrials.gov. In addition, it meets the International Committee of Medical Journal Editors (ICMJE, 2004) registration requirements to enable

the extraction of trial information of interest. The scoping exercise identified trials registered on a randomly selected date (25 June) over a more recent 5-year period (2009 to 2013). The period (2009 to 2013) was chosen because it was anticipated that the proportion of ADs would be greater in more recent years.

In total, 414 trials were randomly identified, assessed for eligibility based on predefined inclusion criteria given in Section 5.3.5, and data completeness examined. Data completeness was unsatisfactory and problematic on key variables. For instance, trial phase was unreported in at least 14% of trials. Of the 414 trials, 71(17%) meeting the inclusion criteria were manually investigated to determine if they were AD or not, using accessible trial-related material such as protocols and publications to aid classification decision-making where necessary. Only 3 of the 71 eligible trials were ADs.

The search via the WHO ICTRP (2004), which is international and linked to ClinicalTrials.gov was deemed problematic due to poor data completeness and the restrictive nature of the searching algorithm, which is limited to lay and scientific titles. Thus, any trial that did not state the adaptive nature of the trial on the search terms list in these titles would not be identified by the electronic search. This problem was less apparent when ClinicalTrials.gov was used during the scoping exercise. As a result of major limitations in using the WHO ICTRP (2004), the main review was restricted to ClinicalTrials.gov, because it offered better data completeness, search options, and flexibility and improvement in filtering records.

5.3.3 Data Sources

For the reasons highlighted in Sections 5.3.1 and 5.3.2, the ClinicalTrials.gov (1997) register was the primary source of the audit review and it is one of the largest clinical trials registers. However, as highlighted later in Section 5.4.7, this register has its own limitations. As a result, the ClinicalTrials.gov (1997) register was supplemented with trials identified through the NIHR project portfolio database (NIHR, 2014b); which contains more information on individual trials through accessible protocols, and 'known' ADs from Clinical Trialist contacts both in the public and private sector. Furthermore, the choice of supplementary sources was influenced by the funding source of this research aimed to make recommendations applicable to the Funder.

Clinical Trialists were contacted via a number of platforms requesting information regarding applied confirmatory ADs that they know of. These platforms included:

1) LinkedIn posts on the PSI and AD working groups, which had a combined total of 1061 members as of the 10th July 2014;

2) Emails to the 'Med Stats Google' group. The exact number of Statisticians on the mailing list is unclear;

3) Emails to the network of Senior Statisticians representing 55 registered UK CTUs.

Contacts were given 4 weeks to respond. Identified ADs meeting the inclusion criteria from the two supplementary sources (the NIHR project portfolio database and from contacts) were then linked back to the ClinicalTrials.gov (1997) register in order to extract additional trial information of interest.

5.3.4 Search Strategy

Figure 5.1 shows an iterative process adopted to identify the final search terms.

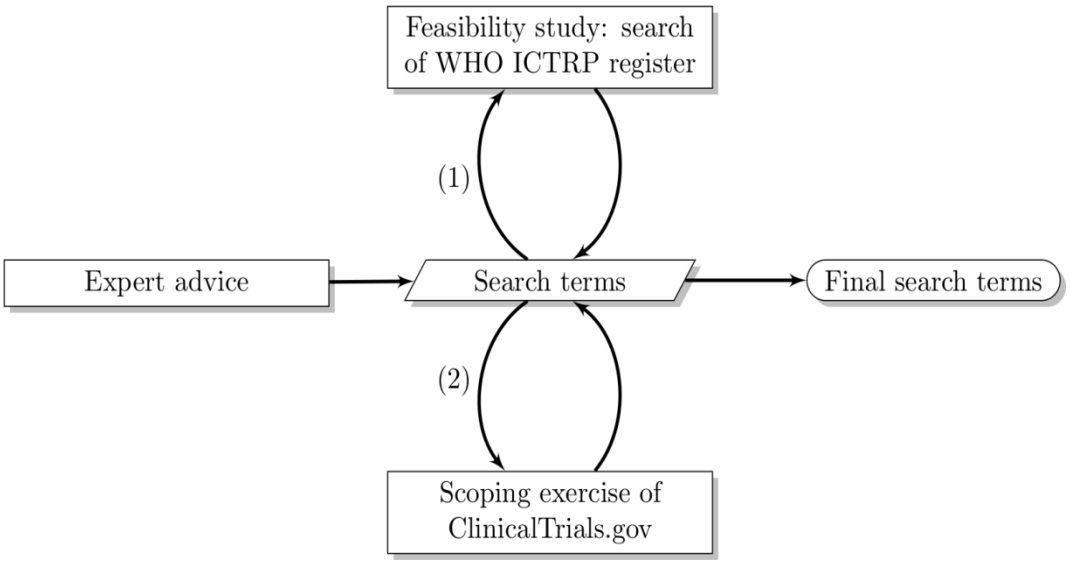

Figure 5.1. A flow diagram of the decision-making process used to determine the final search terms.

Appendix 5.1 contains a list of collated AD-related terms used by Clinical Trialists and experts in the literature, and explored during the scoping study. The individual search terms were applied to the ClinicalTrials.gov register during the scoping exercise to eliminate redundant and undesirable search terms. Appendix 5.1 also summaries the number of trials identified using individual search terms and inclusion decision comments.

On the 1st June 2014, the final search terms were applied to trials meeting the eligibility inclusion criteria (Section 5.3.5) in the ClinicalTrials.gov register, the NIHR project portfolio database (NIHR, 2014b), and 'known'

trials. A list of final search terms was used combined with Boolean operator '*OR*' based on scoping results given in Appendix 5.1:

> *"Adaptive", "Interim", "Dose selection", "Bayesian", "Futility", "Enrichment", "Stopping rule", "Seamless", "Group sequential", "Go/no go", "Preplanned", "MAMS", "Multi-stage", "Multiple stage", "Multiple arm", "Active learning", "Accumulating data", "Continuous reassessment", "Reanalysis", "Pick the winner", "Internal pilot", "Drop the loser", "Dose escalation", "Sample size adjustment", "Sample size re-estimation".*

Extracted trials were manually screened to identify whether they were ADs or not, with the aid of any accessible trial-related publications for the classification decision-making process. In order to minimise classification errors, all identified ADs were verified with the support of two other independent Reviewers.

5.3.5 Eligibility Criteria

A number of eligibility criteria of trials were predefined. Clinical trials were eligible for inclusion into the review if they met the following criteria:

a) Randomised trials investigating any number of interventions on humans including a comparator arm;
b) Phase 2 or 2/3 or 3 trials. However, consistent with the scope of this thesis, focus was given to phase 2/3 and phase 3 RCTs;
c) Trials registered between the 29th February 2000 and the 1st June 2014;
d) Trial documents written in English language.

5.3.6 Data Extraction, Main Outcomes, and Statistical Analysis

The main outcome measures were the type and frequency of identified ADs applied in confirmatory RCTs. The following data were recorded for identified ADs meeting the eligibility criteria stated in Section 5.3.5:

- The year of registration and completion;
- The nature and duration of the primary outcome;
- The expected total sample size;
- The scope of the study (national/international) and the country of the lead Chief Investigator;
- The nature of the experimental intervention and the comparator;
- The number of intervention arms under investigation;

- The Funder or Sponsor of the study;
- The current state of the trial, for example, terminated, ongoing or completed;
- The therapeutic area and population under investigation;
- The nature of the stopping rules when appropriate such as efficacy or futility;
- Whether or not the trial is published;
- Reason for termination of those trials that terminated early;
- Nature of the study design, for example, parallel group;
- Nature of the primary hypothesis of interest, for example, superiority, non-inferiority or equivalence.

Descriptive statistics depending on the type of variables were used, with the aid of graphs for data presentation. The distribution of the type of ADs was explored stratified by the main source of funding (private or public funded).

5.3.7 Examination of the Adequacy of ClinicalTrials.gov in Capturing Adaptive Designs

In order to investigate the adequacy of ClinicalTrials.gov in capturing AD trials, a restricted search of published trials via Ovid MEDLINE was performed. Ovid MEDLINE was chosen as it offers comprehensive search strategy and filtering options. However, further to some limitations highlighted in Section 5.3.1, Ovid MEDLINE does not contain a lot of information on ongoing trials, unless for example, for trials with published protocols.

On the 1st June 2014, an Ovid MEDLINE search was performed using the terms ("*clinical trial, all*" OR "*controlled clinical trials*" OR "*pragmatic clinical trial*" OR "*randomised clinical trial*"). The search terms were combined with Boolean operator 'AND' together with filter limits ("*English*" AND "*humans*" AND "*full text*"). Included were publications from the 29th February 2000 to the 1st June 2014, consistent with the earlier main review via ClinicalTrials.gov and NIHR project portfolio database and meeting the inclusion criteria.

The search retrieved 2079 trials. It was deemed impractical to manually review all the retrieved trials in view of the time limitations. A decision was made to randomly select a feasible fraction of retrieved trials for further examination. In this regard, 300(14.4%) were randomly selected using R software. Random selection was for pragmatic reasons due to time constraints. Randomly selected trials were screened for eligibility and it was established whether they were ADs or not. The type of AD was captured for those identified as ADs. The

registration details of trials classified as ADs were checked and mapped onto the ClinicalTrials.gov register. A manual screening was performed to determine if ADs identified via Ovid MEDLINE were also identified via the ClinicalTrials.gov audit. The number and proportion of missed ADs ('false negative rate') by searching the ClinicalTrials.gov register were estimated.

5.4 Results

This section presents the results of the screening process and characteristics of identified eligible ADs. The distribution of ADs and trend in usage are presented. The proportion of AD trials missed by reviewing the register and additional sources ('false negative rate') is reported. The results exploring the adequacy of ClinicalTrials.gov in capturing ADs are also given.

5.4.1 Trials Eligibility Screening

Figure 5.2 is a flow diagram of the screening process. There were 159,645 registered trials on ClinicalTrials.gov and about 2300 on the NIHR register as of the 1st June 2014. Of these 161,945 combined trials, 554 were assessed for eligibility together with 19 'known' ADs from UK CTUs (n=6) and LinkedIn contacts (n=13). There was no information of known AD trials received from the Med Stats Google group. The reasons for ineligibility are stated in Figure 5.2. In total, 143 trials were eligible ADs in phase 2, 2/3 or 3. Of these, 68 (48%) eligible ADs with some confirmatory objectives (phase 2/3 or 3) are focused on here for further analysis. Of the 19 'known' trials from contacts, 15 had been captured already by the search strategy.

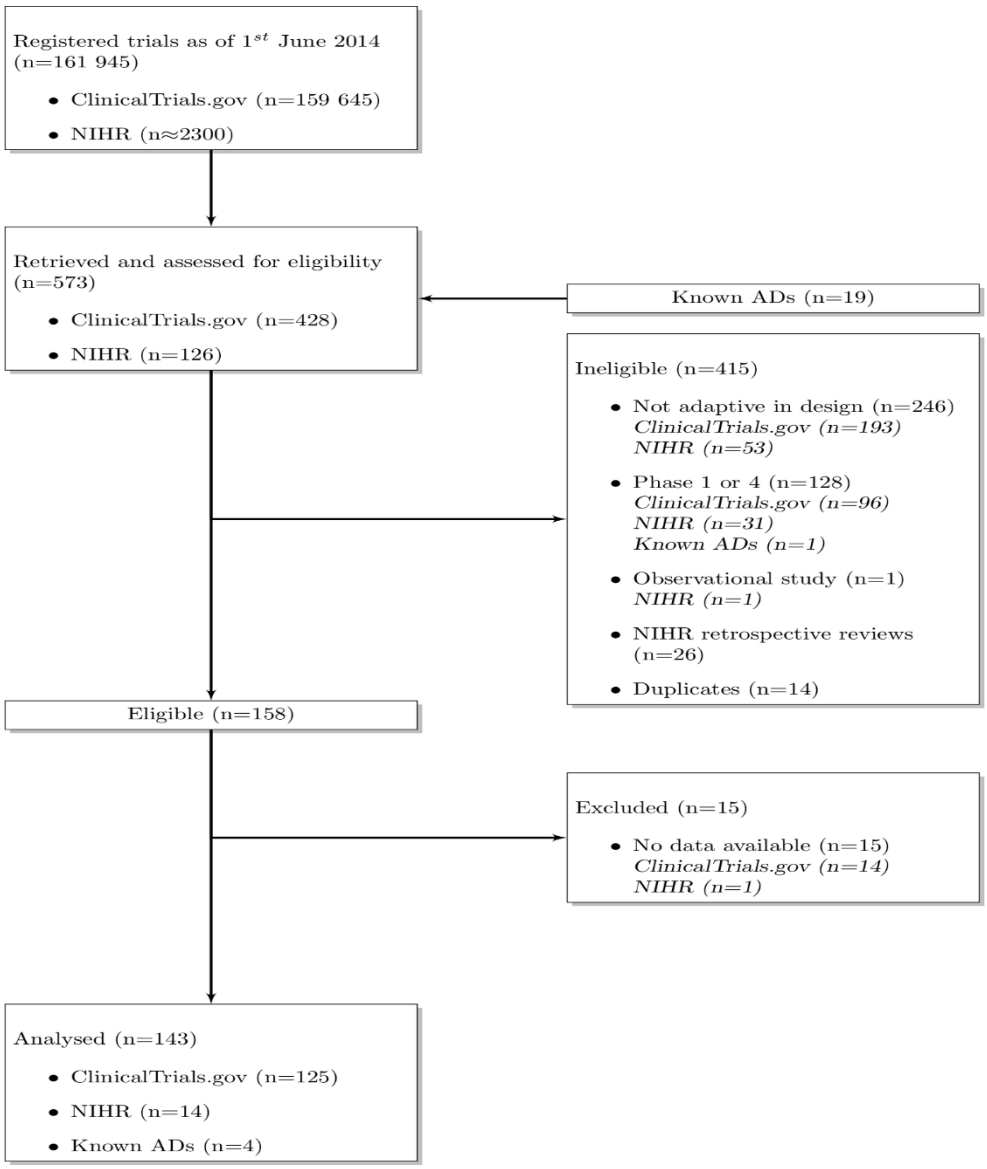

Figure 5.2. A flow diagram showing the review process including reasons for exclusion of trials.

5.4.2 Characteristics of Identified Confirmatory Adaptive Designs

Table 5.1 summarises the characteristics of the 68 identified confirmatory ADs. Of these 68 trials; 36(53%) had standalone confirmatory objectives (phase 3), and the remaining 47% had combined exploratory and confirmatory objectives (phase 2/3). The use of ADs appears to be most common in oncology (31%), although there are diverse areas of application, supporting results in Section 3.4.3.2 of Chapter 3. Trials investigating some form of drug intervention were most common (84%). Thirty (44%) trials were investigating multiple intervention

arms against a comparator. Just 5(7%) were investigating interventions in populations aged below 16. The three main types of primary outcomes were continuous (49%), time-to-event (21%), and binary (19%).

The primary endpoint was well defined in 51(75%) trials. Of these, 8(16%), 13(26%), 5(10%) had a primary endpoint of less than 1 month, between 1 and 3 months, and between 3 and 6 months, respectively. In addition, 20(39%) had primary endpoints between 6 and 12 months. Here, for trials with multiple endpoints, the duration of the longest endpoint was considered focusing only on phase 3 primary endpoints. The distribution of the duration of primary endpoint in days is shown in Table 5.1.

Table 5.1. Characteristics of identified confirmatory adaptive designs stratified by Funder or Sponsor.

Variable	Scoring	Phase 2/3		Phase 3		All phases
		Private	Public	Private	Public	Total
		(n=19)	(n=13)	(n=22)	(n=14)	(N=68)
Therapeutic area	Oncology	4(21%)	6(46%)	7(32%)	4(29%)	21(31%)
	Stroke	-	-	1(5%)	1(7%)	2(3%)
	Mental health	3(16%)	1(8%)	1(5%)	1(7%)	6(9%)
	Dementia	1(5%)	-	-	-	1(2%)
	Cardiology	-	-	2(9%)	2(14%)	4(6%)
	Musculoskeletal	3(16%)	1(8%)	1(5%)	-	5(7%)
	Respiratory	3(16%)	-	1(5%)	1(7%)	5(7%)
	Immunodeficiency	-	1(8%)	-	1(7%)	2(3%)
	Primary care	-	-	-	1(7%)	1(2%)
	Diabetes	1(5%)	-	3(14%)	-	4(6%)
	Oral and gastroenterology	1(5%)	1(8%)	1(5%)	2(14%)	5(7%)
	HIV	1(5%)	1(8%)	3(14%)	1(7%)	6(9%)
	Learning difficulties	-	-	1(5%)	-	1(2%)
	Other	2(11%)	2(15%)	1(5%)	-	5(7%)
Stopping decision criteria	Efficacy	2(11%)	1(8%)	3(14%)	2(14%)	8(12%)
	Safety	1(5%)	-	-	-	1(2%)
	Futility	-	1(8%)	-	-	1(2%)
	Efficacy/safety	12(63%)	4(31%)	13(59%)	4(29%)	33(49%)
	Efficacy/futility	-	2(15%)	1(5%)	1(7%)	4(6%)
	Safety/futility	-	-	-	1(7%)	1(2%)
	Efficacy/safety/futility	4(21%)	5(38%)	5(23%)	6(43%)	20(29%)
Geographical scope	National	4(21%)	11(85%)	6(27%)	12(86%)	33(49%)
	International	15(79%)	2(15%)	16(73%)	2(14%)	35(51%)
State of the study ŧ	Active, not recruiting	-	2(15%)	-	-	2(3%)
	Recruiting	6(32%)	8(62%)	3(14%)	5(36%)	22(32%)
	Ongoing after recruitment	1(5%)	-	4(18%)	1(7%)	6(9%)
	Completed	7(37%)	3(23%)	9(41%)	5(36%)	24(35%)
	Terminated after recruitment	5(26%)	-	6(27%)	3(21%)	14(21%)
Study population	<16	3(16%)	1(8%)	1(5%)	-	5(7%)

(Table continued)

	≥ 16	12(63%)	9(69%)	19(86%)	13(93%)	53(78%)
	>50 only ǫ	4(21%)	1(8%)	1(5%)	1(7%)	7(10%)
	All ages	-	2(15%)	1(5%)	-	3(4%)
Primary outcome(s)	Binary	-	4(31%)	5(23%)	4(29%)	13(19%)
	Continuous	16(84%)	4(31%)	9(41%)	4(29%)	33(49%)
	Time to event	2(11%)	3(23%)	6(27%)	3(21%)	14(21%)
	Ordinal	-	2(15%)	-	1(7%)	3(4%)
	Categorical	-	-	1(5%)	-	1(2%)
	Continuous and time to event	1(5%)	-	1(5%)	-	2(3%)
	Continuous and binary	-	-	-	1(7%)	1(2%)
	Continuous, binary and time to event	-	-	-	1(7%)	1(2%)
Study intervention	Drug	19(100%)	7(54%)	19(86%)	9(64%)	54(79%)
	Device	-	2(15%)	2(9%)	4(29%)	8(12%)
	Drug and device	-	1(8%)	1(5%)	-	2(3%)
	Drug and diet	-	-	-	1(7%)	1(2%)
	Other	-	3(24%)	-	-	3(4%)
Number of intervention arms †	1	5(26%)	9(69%)	13(59%)	11(79%)	38(56%)
	2	7(37%)	2(15%)	6(27%)	1(7%)	16(24%)
	3	4(21%)	-	2(9%)	-	6(9%)
	4	2(11%)	1(8%)	1(5%)	1(7%)	5(7%)
	≥5	1(5%)	1(8%)	-	1(7%)	3(4%)
Duration of primary outcome (days) §	Median (IQR)	84(84 to 176)	365(336 to 365)	336(126 to 365)	88(28 to 365)	182(84 to 365)
	Min to Max	15 to 730	0.010 to 730	70 to 730	0.003 to 365	0.003 to 730
Sample size	Median (IQR)	306(177 to 920)	300(151 to 1000)	451(227 to 810)	1417(500 to 3020)	473(206 to 1210)
	Min to Max	100 to 8031	30 to 8100	150 to 8381	152 to 5418	30 to 8381

IQR: Interquartile Range; Min: Minimum; Max: Maximum. ŧ State of the study as of September 2014. ǫ Only trials not included in ≥16 category. † Number of arms planned at the beginning, excluding the comparator (control) arm. § Only for 51 trials with well-defined primary endpoints of the confirmatory phase.

5.4.3 Description of the Reasons for Early Stopping

Fourteen of the 68 trials (21%) were stopped earlier than planned for a variety of reasons. Table 5.2 details these reasons stratified by main Funder or Sponsor. Of the trials that stopped early, 8(57%) were stopped for futility reasons; 7 in the private sector and 1 in the public sector. Nine (64%) trials that were terminated early had strictly standalone confirmatory objectives. In addition, 9(64%) were conducted in the private sector. Only one private sector trial stopped early for efficacy.

Table 5.2. Reasons for early stopping of adaptive designs.

Reasons for early stopping	Phase 2/3		Phase 3		All Phases
	Private	Public	Private	Public	Total
Futility	3	-	4	1	8
Poor recruitment	1	-	-	1	2
Efficacy	-	-	1	-	1
Safety	-	-	-	1	1
Financial	1	-	-	-	1
Business	-	-	1	-	1
	5	-	6	3	14

5.4.4 Exemplars and Classification of Identified Confirmatory Adaptive Designs

Table 5.3 summarises the number of broadly classified ADs identified by the review. At least 40(59%) were identified as ADs that used GSDs. Six trials with standalone confirmatory objectives employed SSR, although the scope of SSR was unclear. Available information was inadequate to classify whether the identified seamless ADs were operationally or inferentially seamless in nature, as described in Section 2.9 of Chapter 2. Similarly, the scope of the SSR employed was challenging to ascertain given the accessible information. The number of phase 2/3 and 3 adaptive trials funded by the private sector is slightly larger than in the public sector.

Table 5.3. Type of identified adaptive designs stratified by the Funder or Sponsor.

Classification of AD	Phase 2/3		Phase 3		All phases
	Private	Public	Private	Public	Total
GSD	-	2	10	10	22
SSR	-	-	4	2	6
GSD with SSR	-	-	2	-	2
GSD with dose selection	-	1	4	2	7
Seamless	3	1	-	-	4
GSD and seamless	3	6	-	-	9
Seamless with SSR	3	2	-	-	5
Seamless with dose selection	10	1	-	-	11
Multi-arm with dose escalation	-	-	1	-	1
Two stage AD with SSR	-	-	1	-	1
	19	13	22	14	68

AD: Adaptive Design; GSD: Group Sequential Design; SSR: Sample Size Re-estimation. Dose selection relates to treatment selection.

Table 5.4 provides a brief description of the few selected exemplars of identified ADs used to highlight their scope in the confirmatory setting. Identified case studies in the confirmatory phase that used group sequential methods were included in a subsequent review of their reporting presented in Chapter 6. Appendix 5.2 is a complete list summarising identified confirmatory ADs.

Table 5.4. Some exemplars of registered confirmatory adaptive designs.

Trial registration number	Brief description of some identified confirmatory ADs and related publications
NCT01230775	A private sector funded 2 stage confirmatory AD with SSR at the 1st interim analysis applying the methodology proposed by Bauer and Kohne (1994) using p-value combination procedures briefly reflected in Appendix 2.5. It is a double-blinded RCT investigating the efficacy and safety of a drug 'Anagrelide retard' in patients with Essential Thrombocythaemia with a certain defined risk criteria.
NCT01555710	MATISSE study is a private sector sponsored, open-label, randomised trial with an active comparator, adaptive GSD with SSR at the interim analysis evaluating the efficacy of Palifosfamide-tris, in combination with carboplatin and etoposide chemotherapy in chemotherapy naive patients with extensive-stage small cell lung cancer.
NCT00268476	STAMPEDE study is a MAMS RCT, a form of platform trial (Berry et al., 2015), which started by investigating 5 treatments in combination with hormone treatment in treating patients with locally advanced or metastatic prostate cancer with options to drop futile arms or add investigative arms during the trial. The trial is predominantly funded by the UK public sector. Sydes et al (2012, 2009) describe the rationale and design aspects of the trial. James et al (2012) present 1st interim results with decisions to discontinue certain intervention arms. Further results have been reported (James et al., 2015)
NCT01545232	The PROPPR study is a GSD with SSR funded by the public sector in the USA investigating the effectiveness and safety of transfusing patients with severe trauma and major bleeding using plasma, platelets, and red blood cells in a 1:1:1 ratio compared with a 1:1:2 ratio. Baraniuk *et al* (2014) provide the detailed rationale and design of the trial. The AD had 2 efficacy interim analyses at 1/3 and 2/3 of the projected 24-hour or 30-day mortality events were observed (whichever reached its projected 1/3 and 2/3 first). The two co-primary outcomes were separately monitored using a two-sided OBF (1979) boundary with LD (1983) alpha spending function based on numbers of events for each of the two comparisons. SSR was performed prior to the first efficacy interim analysis. Holcomb et al (2015) report the trial results in the Journal of the American Medical Association.
NCT01336530	The PREVAIL study is a private sector funded, randomised, parallel group, double-blind, placebo-controlled, therapeutic confirmatory multicentre trial with 4 intervention arms, inclusive of the comparator. The trial is Bayesian adaptive group-sequential with two interim analyses, possible SSR after the 1st or 2nd interim analysis and a 'drop-the-loser' approach (option to drop futile intervention arms). Holmes et al (2014) report the results of the trial.
NCT00497146	The PRIMO study is a private sector funded trial evaluating the effects of a drug (paricalcitol) on cardiac structure and function over 48 weeks in patients with stage 3/4 chronic kidney disease with left ventricular hypertrophy. The trial is an information based GSD with SSR. Pritchett et al (2011) provide the design details and rationale , and Thadhani et al (2012) present the findings.
ISRCTN06473203	The STAR study is a multi-stage operational seamless 2/3 RCT publicly funded by the NIHR HTA. The trial investigates the effect of a novel drug free interval strategy compared to the standard treatment strategy in the first line treatment of advanced renal cell carcinoma (Collinson et al., 2012).
ISRCTN90061564	The FOCUS4 study is a MAMS, seamless 2/3 design investigating multiple treatments in multiple population enriched biomarkers in oncology. Kaplan et al (2013) provide detailed description of the design, its rationale, statistical properties, and implementation tools.
NCT01056341	This is a private sector funded RCT, placebo-controlled, double-blinded, inferential seamless phase 2/3, two-stage AD, with treatment selection of the propranolol regimen at the end of stage 1 (phase 2) interim analysis and further evaluation of the selected (promising) regimen in stage 2 (Léauté-Labrèze et al., 2015). The trial had a non-binding option for futility stopping and SSR at the interim analysis. Heritier et al (2011) discuss the statistical aspects and practical experiences during implementation of the design. The adopted statistical methodology has been described elsewhere (Posch et al., 2005).

GSD: Group Sequential Design; RCT: Randomised Controlled Trial; MAMS: Multi-Arm Multi-Stage; SSR: Sample Size Re-estimation.

Figure 5.1 displays a clustered bar chart of the number of identified ADs from the years 2001 to 2013, excluding an incomplete year 2014. The number of applied and registered ADs with at least confirmatory objectives appears to have increased slightly. In addition, the number of ADs with both exploratory and confirmatory objectives in one trial (phase 2/3) appears to have increased in more recent years. Overall, the number of standalone phase 3 trials is slightly higher than phase 2/3, however, there seems to be a trend reversal from 2010, with exception in 2012.

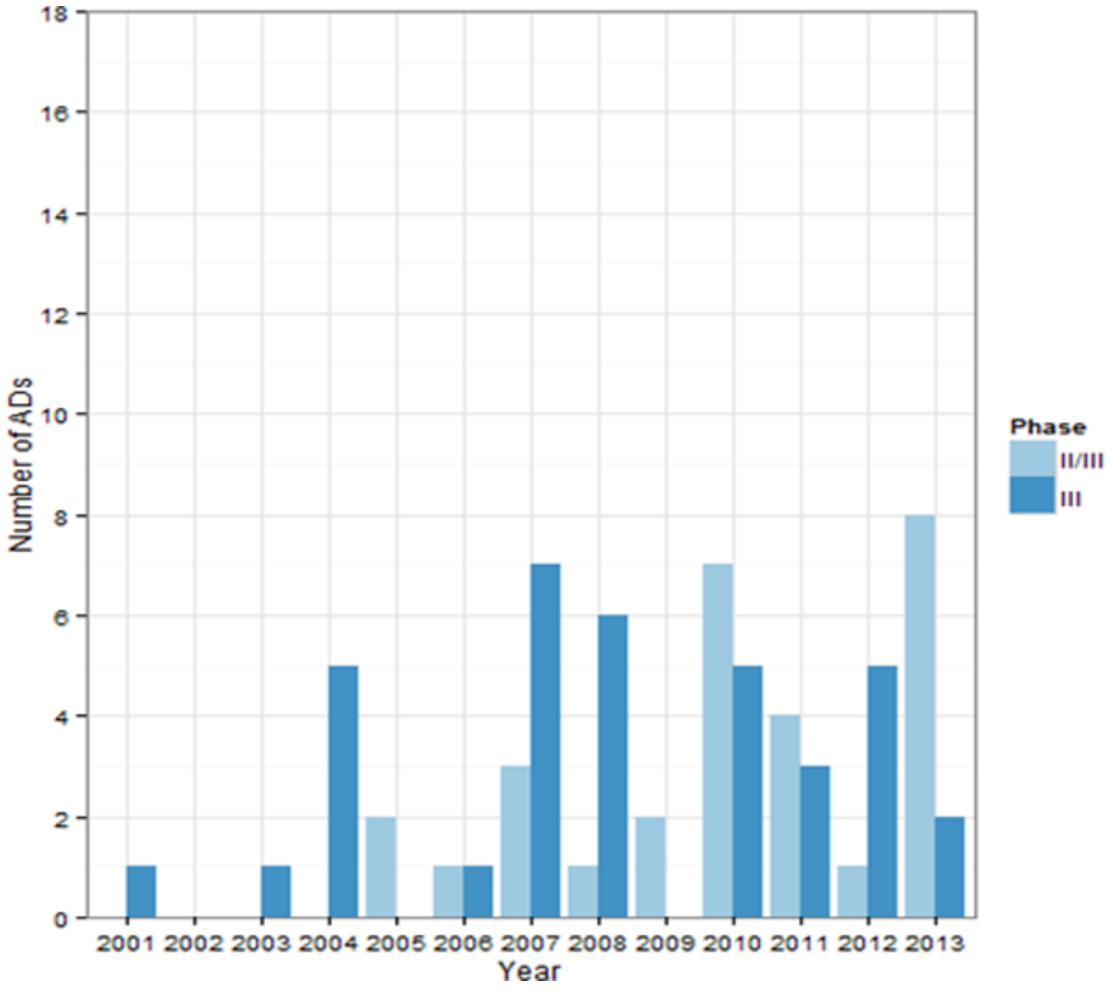

Figure 5.3. Trends in the application of confirmatory adaptive designs.

Figure 5.4 shows a bar chart of the number of ADs undertaken, which were funded by the private and public sector. It appears that the number of ADs applied is slightly higher in the private sector funded trials. However, this does not account for a possible difference in the total number of trials funded by the private and public sector.

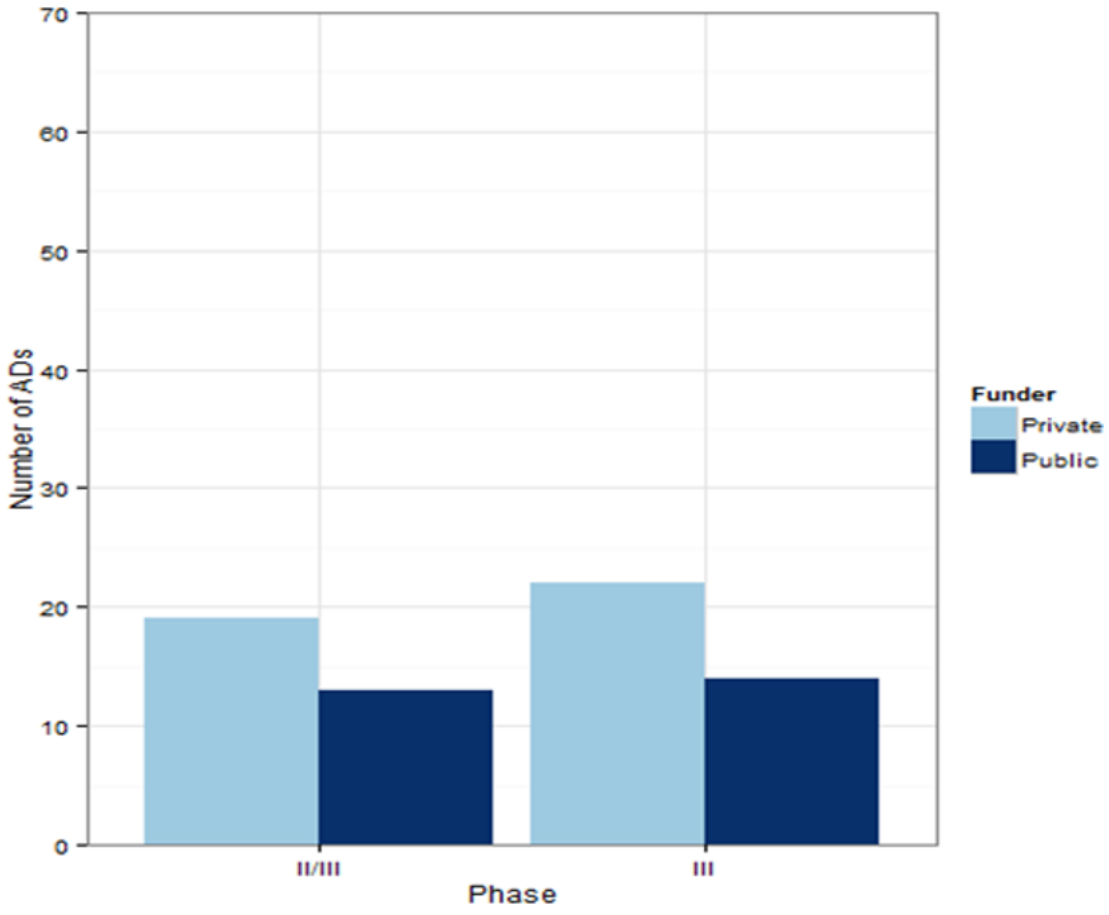

Figure 5.4. The use of confirmatory adaptive designs in the public and private sector.

5.4.5 The Publication of Confirmatory Adaptive Designs

As of the end of September 2014, 38 trials were either completed or terminated earlier than planned. Of these, 25(66%) had published their final or interim results as of the end of May 2015. Just over half, 20(53%), had either published their final or interim results within 2 years of completion or trial termination, highlighting publication delay reflected in Section 5.3.1.

5.4.6 Geographical Distribution of Identified Confirmatory Adaptive Designs

Figure 5.5 shows a bar chart of the application of identified confirmatory ADs by geographical location of the Chief or Principal Investigator. Close to half of the identified ADs were applied in the USA and Canada, whilst the number carried out in the UK appears similar to those in the rest of Europe.

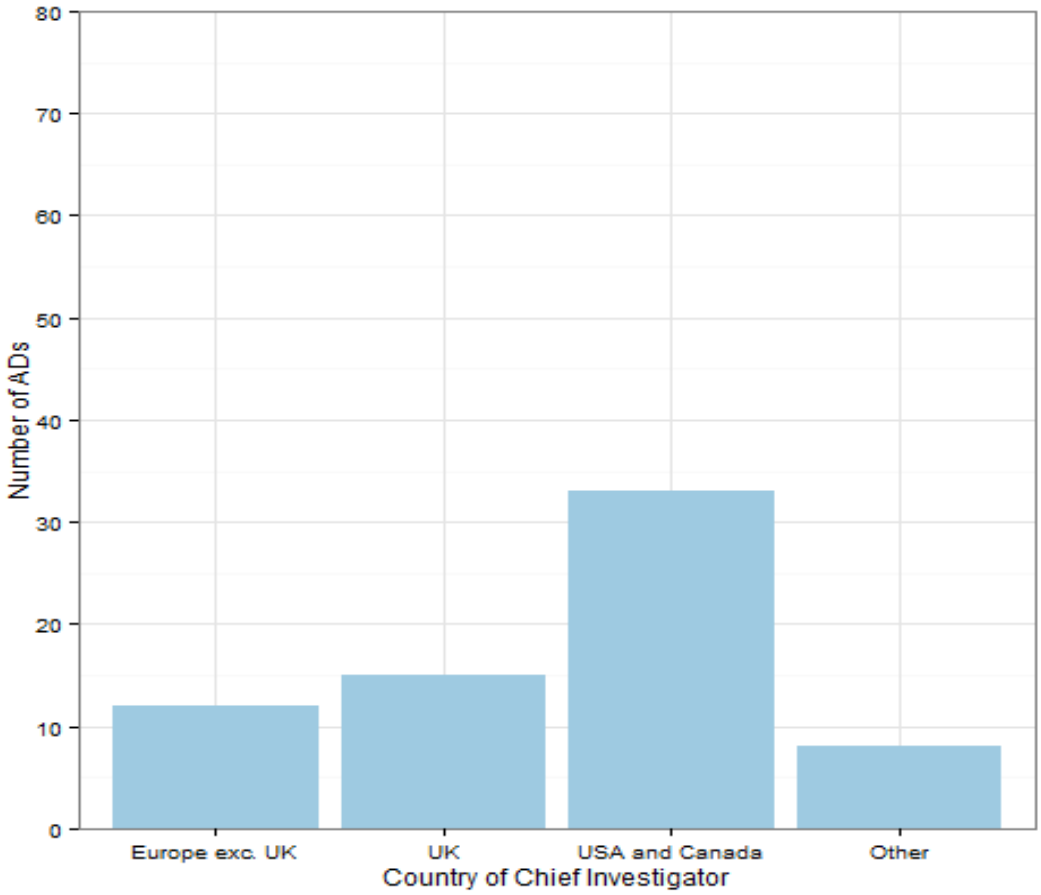

Figure 5.5. Geographical distribution of the application of confirmatory adaptive designs.

5.4.7 Efficiency of ClinicalTrials.gov in Capturing Registered Adaptive Designs

Figure 5.6 shows the flow diagram of the process to investigate the adequacy of the ClinicalTrials.gov register in capturing ADs using randomly selected Ovid MEDLINE trial publications. The Ovid MEDLINE search suggests that a number of AD trials were missed by searching ClinicalTrials.gov. Of the 300 randomly selected Ovid MEDLINE trials, 55(18%) were phase 2 or 2/3 or 3 ADs meeting the inclusion criteria. Of the 18 confirmatory ADs identified via Ovid MEDLINE, none were identified via ClinicalTrials.gov search. The remaining 42% (13/31) were either registered elsewhere or had limited information to ascertain trial registration. Only 9% (1/11) of registered phase 2 AD trials were correctly identified via the ClinicalTrials.gov database. The results indicate a very high 'false negative rate' implying that a significant number of ADs were missed by searching the ClinicalTrials.gov register. However, the scope and type of ADs identified and missed by reviewing the databases appeared similar.

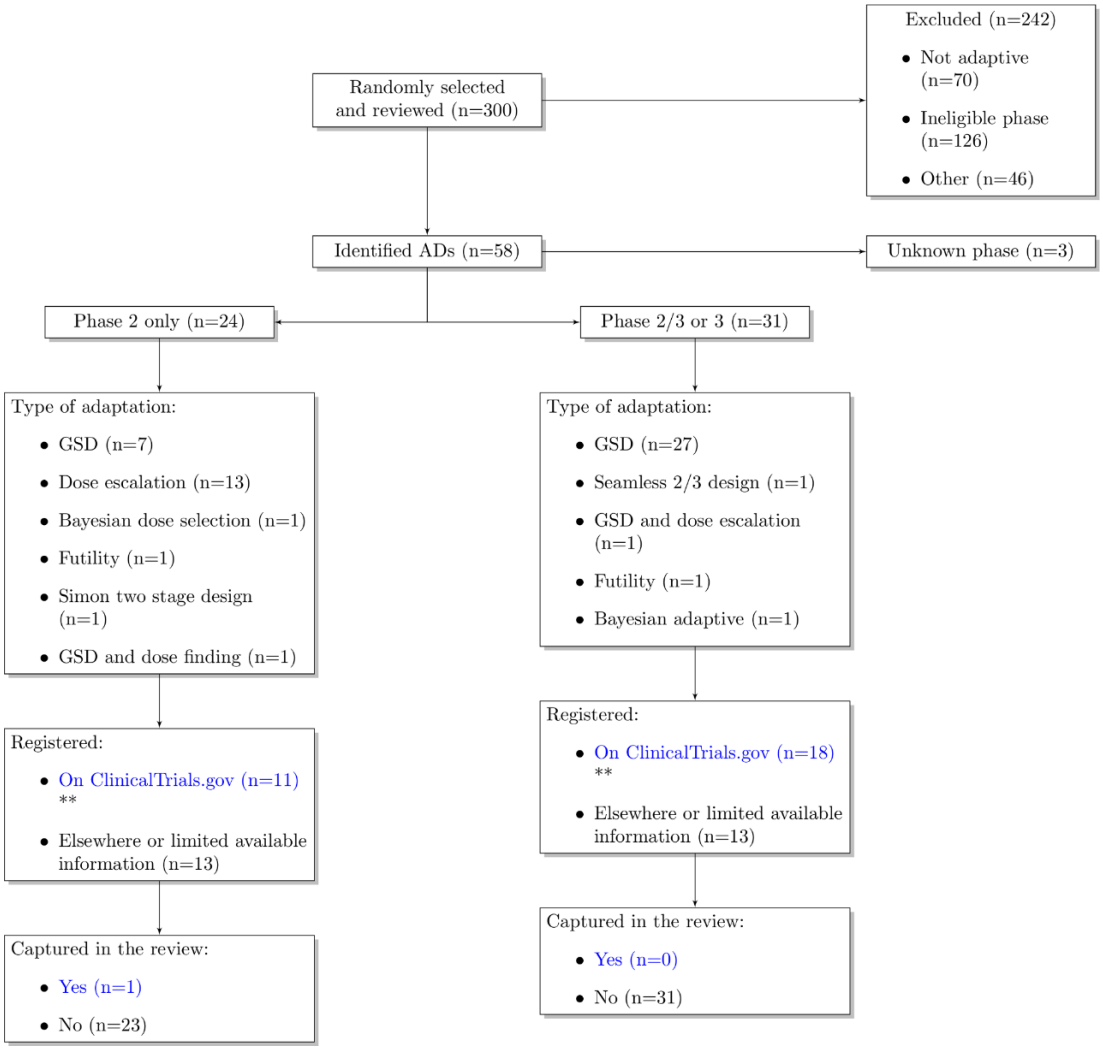

Figure 5.6. Flow diagram investigating the adequacy of ClinicalTrials.gov in capturing adaptive designs.

5.5 Discussion

This section summarises the main chapter findings, discusses their importance in relation to the results from previous chapters, and highlights the main limitations. Furthermore, it discusses how the results relate to the literature, particularly the most recently published work during the course of the thesis, as well as to other thesis chapters. To conclude, the contribution of some chapter findings to the direction of the remainder of this thesis is highlighted.

5.5.1 Main Findings and Implications

The review identified case studies of registered ADs applied in confirmatory trials funded by both the public and private sector. Although the results indicate that the use of confirmatory ADs is not wide spread in relation to the number of trials conducted, it highlighted the types of ADs applied in the confirmatory setting. The case studies identified may help the design and conduct of related ADs in the future provided that they are adequately reported. In addition to the type of ADs surveyed in Chapter 4, the chapter may help to raise awareness of opportunities for trial adaptation in confirmatory trials research.

The results, however, highlight major limitations of clinical trials registers in capturing ADs during registration. Although clinical trials registers have the potential to obviate publication bias and reduce publication time-lag, they are currently inadequate in helping Clinical Trialists identify AD case studies. This is because the details of the scope of the applied ADs are mostly unclear or missing. The poor description and indexing of the type and scope of ADs in the registers is most likely to have resulted in underestimation of particular types of ADs undertaken such as SSR and futility analyses using stochastic curtailment methods reported in cross-sector surveys presented in Section 4.4.7 of Chapter 4. Furthermore, the details provided in the clinical trials registers were inadequate to differentiate between operational and inferential seamless designs. Clinical Trialists should be encouraged to include the term 'adaptive design' in the title when registering ADs. It is also recommended that a small section be inserted in the ClinicalTrials.gov register where researchers can report the type and scope of the AD used. This could be an important resource to improve the appropriate use of ADs, obviate publication bias, and mitigate some barriers to and concerns about their use.

The use of ADs appears most common in certain therapeutic areas, such as oncology although the methodology is applicable across a variety of therapeutic areas. Reasons highlighted in Section 3.4.3.2 of Chapter 3 such as the severity of the health condition and limitations of care options in the current standard-of-care may influence receptiveness of Regulators and the research community towards ADs. The underutilisation of ADs in other therapeutic areas may also be due to limited case studies, which is a potential barrier to implementation in those settings. An increase in the number of phase 2/3 ADs in recent years may reflect the desire to speed up the evaluation of investigative interventions by removing the 'white space' between phases.

The proportion of completed and published trials, which can be used as case studies, is very small. Whilst this may not be limited to ADs, it is vital to have published case studies of ADs to help alleviate the barriers and concerns uncovered in Chapters 3 and 4. Even though this audit review was not exhaustive and limited to only a

few data sources, selected case studies of applied and registered confirmatory ADs are presented. These practical case studies are an important resource to enhance the design and conduct of future related ADs. However, such case studies can only be useful if the related publications are accessible and adequately reported.

5.5.2 Main Strengths, Limitations, and Implications on Interpretation

The major limitations of this audit review is that it did not identify all confirmatory ADs found in the literature for various reasons. For example, limited data sources were used due to time limitation in view of the wide scope of this thesis. Therefore, the results of the number of types of confirmatory ADs and trends in use gives a conservative picture of the reality.

Some more recent surveys have been published during or after the completion of this chapter's work. The Centre for Biologics Evaluation and Research (CBER), an arm of the FDA, recently reported a comprehensive survey of applied ADs both in the private and public sector based on protocol submissions to 3 product offices in the USA (Lin et al., 2015). This study identified 136 ADs from 2008 to 2013: 83(61%) and 53(39%) in phase 2 and phase 3, respectively. These ADs were predominantly GSDs and SSRs, which is consistent with the findings of this chapter and surveys presented in Chapter 4. Morgan et al (2014) reported some ADs based on a review of different sources. For example, the authors reported 179 confirmatory ADs based on a review of medical and statistical journals from 2000 to 2011, which is a larger number than found in this chapter because of the cited limitations of the clinical trials register. Elsäßer et al (2014) reported an audit of 59 AD-related scientific advice letters by the Committee for Human Medicinal Products in the EU received from 2009 to 2012, predominantly in phase 2/3 or phase 3; of which 47(80%) were accepted or 'conditionally accepted' phase 2/3 or 3 AD proposals. Quinlan et al (2010) reported some types and scope of 29 confirmatory ADs based on a limited survey of 13 private sector organisations in the USA.

There are other potential sources of publication bias. For example, the registration of trials was not mandatory and was restricted to life threatening conditions in the earlier years of registries, which was the case for the NIH (USA) before it became mandatory for all trials. In addition, the public sector led the publication registration of trials in earlier years. Hence, the trend in the application of ADs before 2004 may not reflect the true picture. However, despite the cited limitations, the findings indicate a modest improvement in the number of applied confirmatory ADs and their diversity, which is consistent with recent surveys (Elsäßer et al., 2014; Lin et al., 2015; Morgan et al., 2014; Quinlan et al., 2010).

There are strengths of the research presented here, which distinguish it from the cited related studies. First, this work identified and presented accessible case studies of confirmatory ADs, an important resource that can be used by Clinical Trialists to alleviate some cited obstacles uncovered in Chapters 3 and 4. Researchers who are interested in applying related ADs may look for related publications of these case studies, such as protocols or contact Chief Investigators for additional details. Thus, facilitating the transfer of applied knowledge on ADs. Second, the work examined the adequacy of the ClinicalTrials.gov register in capturing ADs during registration and proposed some recommendations to improve identification of ADs through the register.

5.5.3 Implications for the Research Described in the Remainder of the Thesis

It is very important for case studies of applied ADs to be made accessible and adequately reported, particularly in view of the fact that some concerns uncovered in Chapters 3 and 4 appear to be intertwined with inadequate reporting of undertaken ADs. These concerns include fear of introducing operational bias, robustness of ADs in decision-making, and acceptability to change practice when trials are stopped early. In this regard, trials identified as GSD (the most common ADs) are further reviewed to investigate completeness in their reporting, and appropriateness of the CONSORT 2010 statement to be described in Chapter 6. Lastly, the results of this chapter and surveys presented in Chapter 4 influenced the type of AD case studies to be considered for further work presented in Chapter 7.

Chapter 6. Transparency and Reporting of Confirmatory Adaptive Designs

6.1 Introduction

Research reported in Chapter 3, which has been published elsewhere (Dimairo, Boote, Julious, Nicholl, et al., 2015) highlighted a degree of conservatism as one of the perceived major barriers to the use of confirmatory ADs. This degree of conservatism seems to be influenced by, among other factors:

a) Concerns regarding the robustness of ADs in decision-making and fear of making wrong decisions when trials are stopped early;
b) Concerns about acceptability of findings from ADs to change clinical practice;
c) Worry about potential introduction of operational bias during the conduct of adaptive trials.

Follow-up research based on quantitative surveys reported in Chapter 4 and published elsewhere consolidated these findings (Dimairo, Julious, Todd, Nicholl, et al., 2015). It can be argued that the scientific rigour in the conduct of trials through transparency and adequate reporting has the potential to alleviate some of these concerns. Importantly, Chapter 4 findings overwhelmingly support transparent adequate reporting of the undertaken ADs as one of the key potential facilitators to their appropriate use.

The CONSORT statement was first published in 1996 with the aim to enhance adequate reporting of RCTs (Begg et al., 1996). Revisions have since been implemented in 2001 and 2010 (Moher, Schulz, et al., 2001; Schulz et al., 2010). There has been marked general improvement in the conduct and reporting of RCTs since the advent of the first CONSORT statement (Egger et al., 2001; Moher, Jones, et al., 2001; Turner et al., 2012). However, despite this, there are still some suboptimal areas requiring improvements (Altman et al., 2012; Turner et al., 2012). Extensions to the CONSORT statement have since been made to accommodate other trial designs and hypotheses, such as: cluster RCTs, non-inferiority and equivalence trials, and pragmatic RCTs (Campbell et al., 2012; Piaggio et al., 2012; Zwarenstein et al., 2008). As of the 12th October 2015, the EQUATOR Network (2006) were developing at least 37 reporting related guidance documents to enhance transparency in the reporting and conduct of studies.

Even though the CONSORT 2010 statement has some items relating to ‘interim analyses’, a CONSORT statement tailored for ADs does not currently exist. Furthermore, findings from Chapters 3 and 4 suggest that the

current reporting guidance framework for ADs may be inadequate for research consumers and, decision and policy makers to make informed judgements regarding the scientific quality of the research in front of them.

The need to revise the CONSORT statement to accommodate ADs has been suggested (SAACTD Workshop Committee, 2009). Detry et al (2012) proposed some aspects requiring modification for different types of ADs, including those designed using the Bayesian approach. Although some of these propositions appear robust in capturing features of a number of ADs, they were not informed by evidence on what is considered important by key research stakeholders. In addition, the authors' suggestions overlooked aforementioned concerns that influence conservatism towards ADs. Section 6.3.4 describes the aspects that were previously overlooked by the authors. This chapter therefore investigates the state of affairs in the reporting practice of the most commonly applied confirmatory AD.

The content of this chapter is based on published research during the course of this thesis (Stevely et al., 2015). The results have also been orally presented at a number of conferences; SCT (Dimairo, Stevely, Julious, Todd, et al., 2015) and 3rd ICTMC (Dimairo, Stevely, Todd, Julious, et al., 2015), and JSM (Julious et al., 2015). So, in addition to my supervisors, the chapter acknowledges the collaborative support of other researchers during the review process and write up of the published paper: Abigail Stevely, an MBChB Intern Student under my supervision and CTRU researchers (Prof Cindy Cooper and Dr Daniel Hind). This is independent research lead and supervised by myself. The advice of Helen Wood on the search strategy is also acknowledged.

6.2 Aims and Objectives

Research described in Chapters 4 and 5 illustrated that the most commonly used confirmatory AD is the GSD. As a result, this chapter aims to investigate the adequacy of the CONSORT 2010 statement in enhancing the reporting of the most commonly applied AD and to propose appropriate recommendations. The specific objectives are to:

1) Assess reporting compliance of group sequential RCTs against the CONSORT 2010 statement;
2) Investigate the shortcomings of the CONSORT 2010 statement to enhance adequate reporting of the most common AD using some researcher-led proposed modifications described in Section 6.3.4;
3) Investigate some possible facilitators to alleviate cited concerns influencing degree of conservatism to the use of confirmatory ADs;

4) Explore exemplars of well-reported aspects of group sequential RCTs, which Clinical Trialists can use to enhance adequate trial reporting and reproducibility.

6.3 Methods

This section describes the methods used to conduct the review and also explains the rationale for the decisions made.

6.3.1 Trials Eligibility Criteria

The inclusion of trials in the review was guided by a strict list of criteria. Trials meeting the following inclusion criteria were eligible for inclusion in the review:

a) Conducted to investigate interventions in humans with a comparator (or control) arm,
b) Parallel group RCTs since they are the most common and the CONSORT 2010 statement is tailored for their reporting,
c) RCTs with confirmatory objectives,
d) Prospectively planned interim analyses within the group sequential framework using the Frequentist approach,
e) Unrestricted nature of the primary endpoint(s),
f) Unrestricted number of intervention arms under investigation,
g) Unrestricted therapeutic area,
h) Trials with accessible full-text reports of primary results in peer reviewed medical journals,
i) Trial publications written in English language and published between the 1st January 2001 and 23rd September 2014.

Excluded were group sequential RCTs designed using the Bayesian approach as they are outside the scope of this thesis.

6.3.2 Searching the Literature and Data Sources

A scoping exercise was performed by searching the Ovid MEDLINE database with the support of an experienced researcher to develop an efficient search strategy. The scoping exercise found one potentially relevant MeSH term '*Early Termination in Clinical Trials'*, which could be used to index some group sequential RCTs.

One drawback of this MeSH term is that it biases the research results in favour of trials that were stopped early. Since this research focuses on the reporting of group sequential RCTs irrespective of their early stopping status, which is also an important outcome of interest, the MeSH search term was deemed undesirable. In addition, the MeSH search term was also insensitive when used via Ovid MEDLINE (Dimairo et al., 2014).

As a result of the lack of MeSH terms for consistent indexing of trial publications that utilise interim analyses using group sequential methodology, it was challenging to establish an optimal search strategy to systematically review all group sequential RCTs. As a result, a free text search of keywords often associated with group sequential methodology was preferred. The key search terms used were:

- '*Group sequential*',
- '*Interim analysis*' or '*interim analyses*',
- 'S*topping rule'* or *'stopping rules'* or *'stopping boundary'* or *'stopping boundaries*',
- '*Interim monitoring*', '*early stopping' or 'early termination*' or '*accumulating data' or 'accumulating information*".

The following more general terms were excluded because they yielded a very high number of irrelevant reports making the review impractical within the time and resources constraints:

- '*Halted*',
- 'C*losed*',
- '*Closure*',
- '*Independent data monitoring committee*',
- '*Data monitoring and safety board*',
- '*Data monitoring and ethics committee*'.

The use of specific stopping boundaries was undesirable since an exploration of the most frequently used stopping rules was also an outcome of interest. Literature searches using individual search terms were performed on the 23rd September 2014 by searching Ovid MEDLINE in combination with additional eligibility filters. The filters were:

- Publication type (*clinical trials, phase 3*),
- Check tags (*humans, full-text available, English language*),
- Publication year (*1st January 2001 to 23rd September 2014*).

The final search combined independent searches with a Boolean operator '*OR*'.

Group sequential RCTs identified as part of a review reported in Chapter 5 supplemented the trials retrieved by searching Ovid MEDLINE as shown in Figure 6.1. The screening and identification of duplicate records for exclusion was performed based on the title, first author, and year of publication.

6.3.3 Data Extraction and Quality Control

This review was conducted independently with the support of another Reviewer to meet publication standards. Trial reports were screened for eligibility, characteristics of eligible trials recorded, and reporting compliance examined. Reporting compliance was examined against the CONSORT 2010 checklist items and researcher-led proposed modifications described in Section 6.3.4. The reviewing and rectification of all discrepancies was undertaken in agreement between the two independent Reviewers. Accessible additional related reports, such as protocols and other prior publications were used to assess reporting compliance. Where possible and necessary, Chief or Principal Investigators were contacted on trial related queries through email and given 3 weeks to respond.

6.3.4 Researcher-led Proposed CONSORT Items

Chapters 3 and 4 found that the following aspects are important to alleviate some of the concerns that influence conservatism towards the use of confirmatory ADs:

a) The use of appropriate statistical methods for early stopping bias correction to obtain unbiased inference;
b) The mechanisms (processes and procedures) put in place to minimise operational bias due to the leakage or knowledge of interim results;
c) The access to prior interim results (or trend in the results prior to the interim analyses reporting);
d) The rationale put forward for choosing a group sequential RCT (and any other add-on planned adaptations such as treatment selection and SSR);

Detry et al (2012) suggest the following additional aspects viewed to be important:

e) Clarification on whether the sample size was adjusted for interim analyses;
f) Discussion of any unplanned deviations (or ad hoc protocol deviations) with reasons;

g) Discussion of the lessons learned and value of using a group sequential RCT (or any other adaptations) to inform the planning of future related trials;

h) Discussion of the generalisability of the results from a group sequential RCT- to whom the results should be generalised based upon adaptations;

Due to the poor indexing of group sequential RCTs as highlighted in Section 6.3.2, the following item was also included:

i) Identification of the design as 'group sequential' in the Title or Abstract.

In addition, the items relating to 'interim analyses', which are already covered by the CONSORT 2010 statement were also assessed for reporting compliance. However, some related aspects are not covered in the CONSORT 2010 statement as individual items. These items are:

j) Description of the decision or stopping rules or boundaries used,

k) Description of the planned stopping criteria used,

l) Description of the sample size (or number of events) at interim analyses,

m) Clarification on whether the trial or intervention arm(s) was stopped early,

n) Explanation of the reasons for early stopping of the trial or intervention arm(s) when appropriate.

All items a) to n) were assessed for completeness in their reporting.

6.3.5 Outcome Measures, Statistical Analysis and Reporting

The primary outcome of the review is to establish reporting compliance to CONSORT 2010 checklist items, as well as the additional dimensions of interest specific to group sequential RCTs. Compliance was subjectively examined and agreed upon independently with the support of another Reviewer. A predefined classification system of completeness guided the subjective examination of reporting compliance:

- *'Absent'*,
- '*Totally complete'*,
- *'Partially complete*',
- '*Cannot access*',
- *'Not applicable*'.

The number and proportion of group sequential RCTs meeting 'total' and 'at least partial' reporting compliance criteria for each checklist item was calculated. A global measure of the number and proportion of checklist items meeting 'total' and 'at least partial' reporting compliance criteria was further calculated.

A protocol version which guided the conduct of this review is accessible via White Rose repository (Dimairo et al., 2014). In addition, the PRISMA guidance framework enhanced the conduct and reporting of this chapter's research (Moher et al., 2009; Stovold et al., 2014). Descriptive summary statistics including numbers (percentages) and median (IQR) were used to assess reporting compliance. Multiple stacked bar charts and forest plots were used to aid visual interpretation. Two-sided 95% CIs around proportions were computed using the Wilson Score method (Newcombe, 1998). Fisher's exact test was used to explore differences in proportions between subgroups of interest, and estimates are presented as risk ratios (RRs), with associated 95% CIs.

A global measure of reporting compliance, based on the number and proportion of checklist items meeting a certain completeness criterion, was also used. Bootstrap methods (Efron, 1979) were used with 10 000 replicates, to compute the median difference, or median ratio (95% CI), of the total number of checklist items meeting certain reporting compliance criteria between subgroups. This approach was utilised in order to explore whether the publication journal's CONSORT endorsement policy (yes or no), and publication period (pre- or post-publication of the CONSORT 2010 statement), was associated with improved reporting compliance. The latter was used to explore the impact of the CONSORT 2010 statement in enhancing reporting. Comparability of compliance in reporting, between the standard CONSORT and researcher-led proposed items, was descriptive without any significance testing.

6.4 Results

This section presents the results of the screening process, characteristics of reviewed eligible trials, and details of reporting compliance. These findings have been reported elsewhere during the course of this thesis (Stevely et al., 2015).

6.4.1 Eligibility Screening

Figure 6.1 is a flow diagram showing the screening process to identify eligible group sequential RCTs. On the 23rd September 2014, 234 study reports were identified by searching the Ovid MEDLINE database. There were an additional 50 group sequential RCTs identified, predominantly by searching the ClinicalTrials.gov

register, and reported in Chapter 5. In total, 284 study reports were screened for eligibility, of which, 68(24%) peer reviewed publications of group sequential RCTs reporting primary results were eligible for reporting compliance examination. Figure 6.1 details the reasons for the exclusion of ineligible trials.

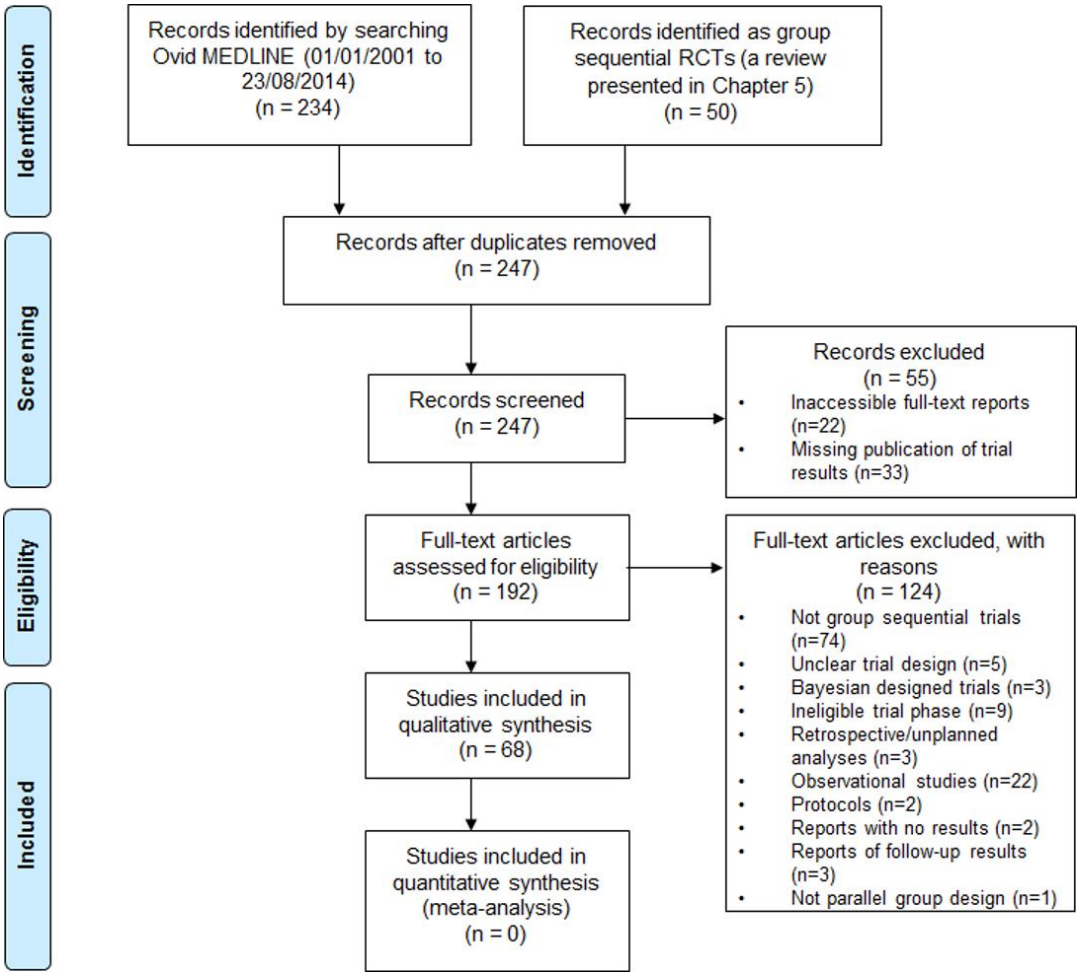

Figure 6.1. PRISMA flow diagram of the screening process.

6.4.2 Characteristics of Included Group Sequential Trials

Table 6.1 summarises detailed characteristics of examined eligible group sequential RCTs, stratified by publication period (pre- or post-publication of the CONSORT 2010 statement). The majority of RCTs were published in 'high impact' medical journals such as The New England Medical Journal, The Lancet Oncology, The American Society of Clinical Oncology, and The Journal of the American Medical Association. The median (IQR) journal impact factor for the year 2013 to 2014 was 17.5 (6.6 to 30.4), and a maximum of 54.4.

The majority of eligible group sequential RCTs (76%) were investigating intervention(s) in the therapeutic area of oncology. However, diverse therapeutic conditions under investigation included cardiovascular, respiratory, and infectious diseases. Most eligible group sequential RCTs (91%) investigated at least some form of pharmacological intervention, and 55(81%) were designed with two intervention arms, inclusive of the comparator arm. Forty-six (68%) of the publishing journals endorsed the CONSORT statement as part of their publication policy.

Table 6.1. Characteristics of eligible reviewed group sequential randomised trials.

Variable	Scoring	Publication period		Total (n=68)
		2001-2010 (n=34)	2011-2014 (n=34)	
Funder/Sponsor	Private	16(47%)	19(56%)	35(51%)
	Public	8(24%)	11(32%)	19(28%)
	Private and Public	4(12%)	4(12%)	8(12%)
	None/independent	1(3%)	-	1(1%)
	Undisclosed	5(15%)	-	5(7%)
Nature of primary outcome(s)	Time-to-event	23(68%)	28(82%)	51(75%)
	Binary	6(18%)	3(9%)	9(13%)
	Continuous	3(9%)	3(9%)	6(9%)
	Binary and continuous	1(3%)	-	1(1%)
	Binary and time-to-event	1(3%)	-	1(1%)
Number of intervention arms	2	26(76%)	29(85%)	55(81%)
	3	6(18%)	3(9%)	9(13%)
	4	1(3%)	1(3%)	2(3%)
	5 or 6	1(3%)	1(3%)	2(3%)
Therapeutic area	Oncology	28(82%)	24(71%)	52(76%)
	HIV/AIDS	3(9%)	-	3(4%)
	Cardiac	-	2(6%)	2(3%)
	Musculoskeletal	1(3%)	1(3%)	2(3%)
	Optical	-	2(6%)	2(3%)
	Stroke	-	1(3%)	1(1%)
	Respiratory	1(3%)	-	1(1%)
	Diabetes	-	1(3%)	1(1%)
	Multiple Sclerosis	1(3%)	-	1(1%)
	Degenerative	-	1(3%)	1(1%)
	Epilepsy	-	1(3%)	1(1%)
	Kidney	-	1(3%)	1(1%)
Journal CONSORT endorsement status	No	13(38%)	9(26%)	22(32%)
	Yes	21(62%)	25(74%)	46(68%)
Publishing journal	The Lancet Oncology	3(9%)	9(26%)	12(18%)

(Table continued)

	The New England Journal of Medicine	5(15%)	7(21%)	12(18%)
	American Society of Clinical Oncology	8(24%)	4(12%)	12(18%)
	Annals of Oncology	3(9%)	2(6%)	5(7%)
	The Journal of the American Medical Association	1(3%)	4(12%)	5(7%)
	Breast Cancer Research Treatment	2(6%)	1(3%)	3(4%)
	Journal of Clinical Oncology	2(6%)	1(3%)	3(4%)
	The Lancet	1(3%)	1(3%)	2(3%)
	The American Academy of Ophthalmology	-	2(6%)	2(3%)
	Journal of the National Cancer Institute	2(6%)	-	1(3%)
	Arthritis and Rheumatology	-	1(3%)	1(1%)
	British Journal of Surgery	1(3%)	-	1(1%)
	Clinical Breast Cancer	-	1(3%)	1(1%)
	Clinical Cancer Research	1(3%)	-	1(1%)
	European Journal of Cancer	-	1(3%)	1(1%)
	HIV Clinical Trials	1(3%)	-	1(1%)
	Journal of Urology	1(3%)	-	1(1%)
	Nutrition	1(3%)	-	1(1%)
	Radiotherapy and Oncology	1(3%)	-	1(1%)
	The Journal of Infectious Diseases	1(3%)	-	1(1%)
Type of intervention	Drug	29(85%)	30(88%)	59(87%)
	Dietary	1(3%)	1(3%)	2(3%)
	Device	-	1(3%)	1(1%)
	Physiological	1(3%)	-	1(1%)
	Radiotherapy	1(3%)	-	1(1%)
	Drug and radiotherapy	-	1(3%)	1(1%)
	Drug and dietary	1(3%)	-	1(1%)
	Surgical	1(3%)	-	1(1%)
	Vaccine	-	1(3%)	1(1%)
Class of intervention	Pharmacological	30(88%)	32(94%)	62(91%)
	Non-pharmacological	4(12%)	2(6%)	6(9%)
Stage of reporting	Interim analysis	25(74%)	22(65%)	47(69%)
	Final analysis	7(21%)	6(18%)	13(19%)
	Unplanned interim analysis	2(6%)	6(18%)	8(12%)
Number of planned interims	1	16(47%)	12(35%)	28(41%)
	2	9(26%)	14(41%)	23(34%)
	3	3(9%)	2(6%)	5(7%)
	4	-	4(12%)	4(6%)
	5 or 7	3(9%)	-	3(4%)
	Undisclosed	3(9%)	2(6%)	5(7%)
Trial stopped early	No	11(32%)	9(26%)	20(29%)

(Table continued)

	Yes	22(65%)	24(71%)	46(68%)
	No, but interim arm discontinued at interim	1(3%)	1(3%)	2(3%)
Reasons for early stopping (N=46)	Futility	12(55%)	10(42%)	22(48%)
	Efficacy	5(23%)	5(21%)	10(22%)
	Safety	1(5%)	1(4%)	2(4%)
	Futility and safety	-	5(21%)	5(11%)
	Poor recruitment and/or financial	3(14%)	3(13%)	6(13%)
	Futility and external information	1(5%)	-	1(2%)
Planned stopping criteria	Undisclosed	16(47%)	6(18%)	22(32%)
	Futility or efficacy	8(24%)	12(35%)	20(29%)
	Futility	3(9%)	6(18%)	9(13%)
	Efficacy	-	6(18%)	6(9%)
	Efficacy or safety	3(9%)	1(3%)	4(6%)
	Futility or efficacy or safety	1(3%)	3(9%)	4(6%)
	Non-inferiority	2(6%)	-	2(3%)
	Safety	1(3%)	-	1(1%)
Planned total sample size	Min to Max	160-8028	100-15000	100-15000
	Median(IQR)	604(350-1071)	784(428-1200)	724(357-1155)

IQR: Interquartile range; Min: Minimum; Max: Maximum.

6.4.3 Reporting of Universal CONSORT 2010 Checklist Items

Figure 6.2 is a multiple bar chart showing reporting compliance against the CONSORT 2010 checklist items. Appendix 6.1 provides summary data on reporting compliance supporting Figure 6.2. In general, most checklist items were well reported. The median proportions (IQR) of group sequential RCTs meeting 'total' and 'at least partial' reporting compliance criteria of checklist items was 81%(53% to 91%) and 93%(78% to 97%), and a minimum of 12% and 22%, respectively. As evident in Figure 6.2, suboptimal reporting compliance was observed in items relating to:

a) Details of the randomisation concealment 50(74%),

b) Implementation of randomisation 40(59%),

c) The disclosure of and access to full trial protocols 53(53%),

d) Methods used to generate the randomisation list(s) 32(47%),

e) Details of additional analyses 29(43%),

f) Disclosure of trial registration information 26(38%).

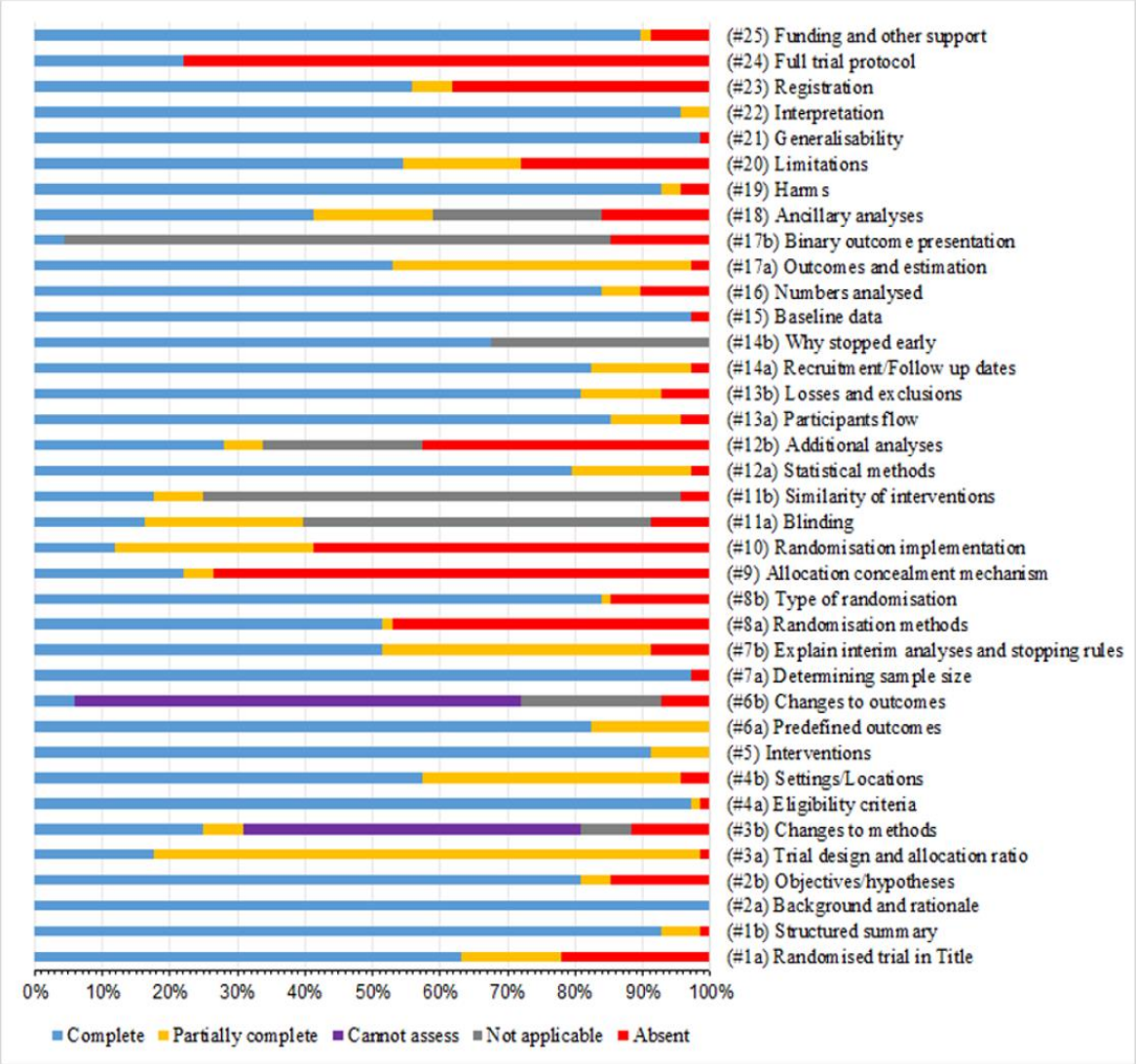

Figure 6.2. Reporting compliance of universal CONSORT 2010 checklist items.

Figure 6.3 is a forest plot showing the proportions of group sequential RCTs meeting the 'total' reporting compliance criterion. Reporting compliance point estimates are presented with their associated 95% CIs.

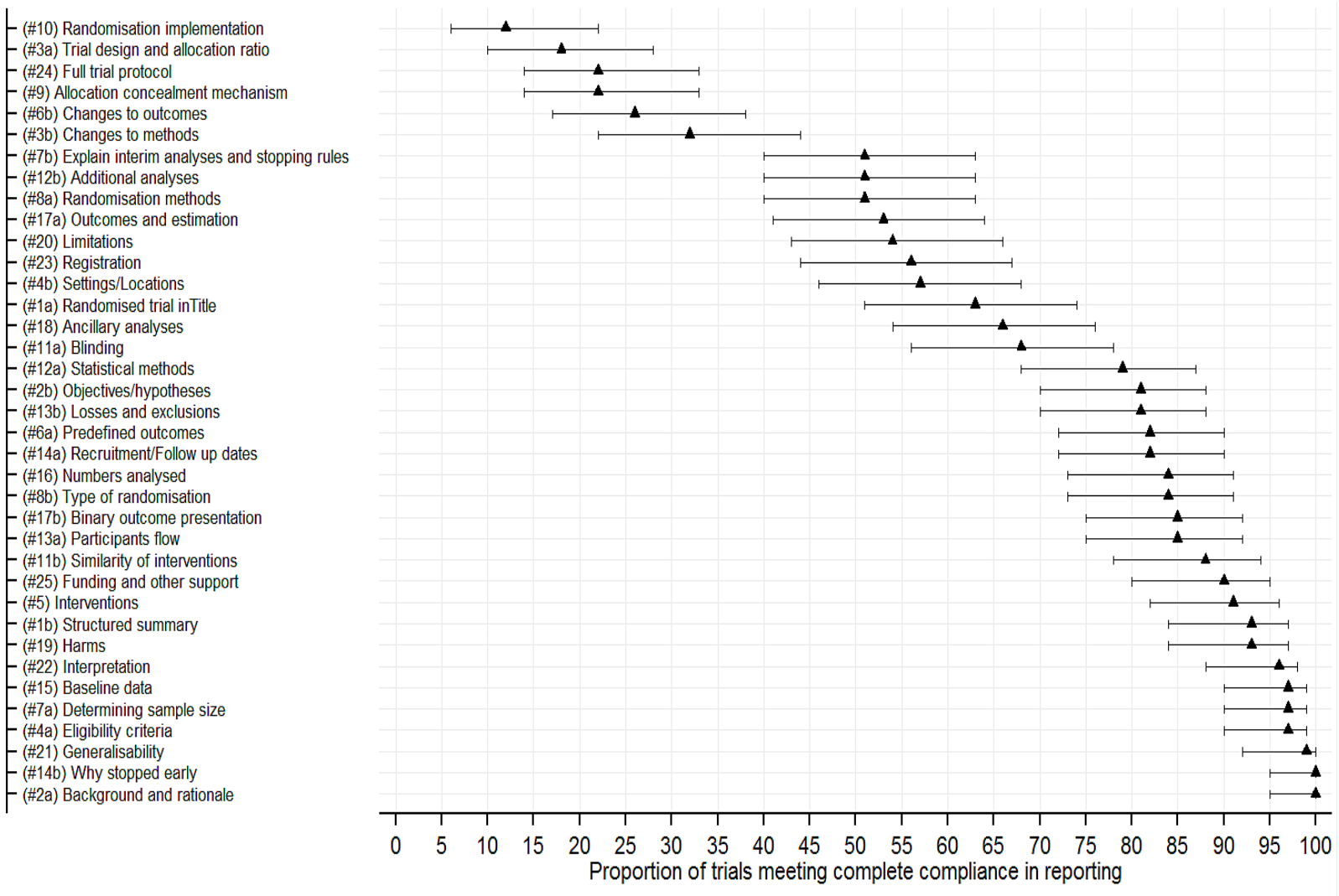

Figure 6.3. Trials meeting 'total' reporting compliance of universal CONSORT checklist items.

Changes to methods and outcomes could not be examined in 34(50%) and 45(66%) group sequential RCTs, respectively. This was because of inaccessible protocols and related amendments for most group sequential RCTs. Of the 37 CONSORT checklist items, the median number (proportion) [IQR] that were completely reported was 26(70%) [24(65%) to 28(76%)], and a minimum of 15(41%). The median number (proportion) [95% CI] of items that met complete compliance increased by 2(5%) [1(1%) to 4(10%); p-value=0.009] post-publication of the CONSORT 2010 statement. Aspects relating to the trial design were partially reported in 55(81%) RCTs.

Figure 6.4 shows a forest plot of the proportions of group sequential RCTs meeting 'at least partial' reporting compliance criteria.

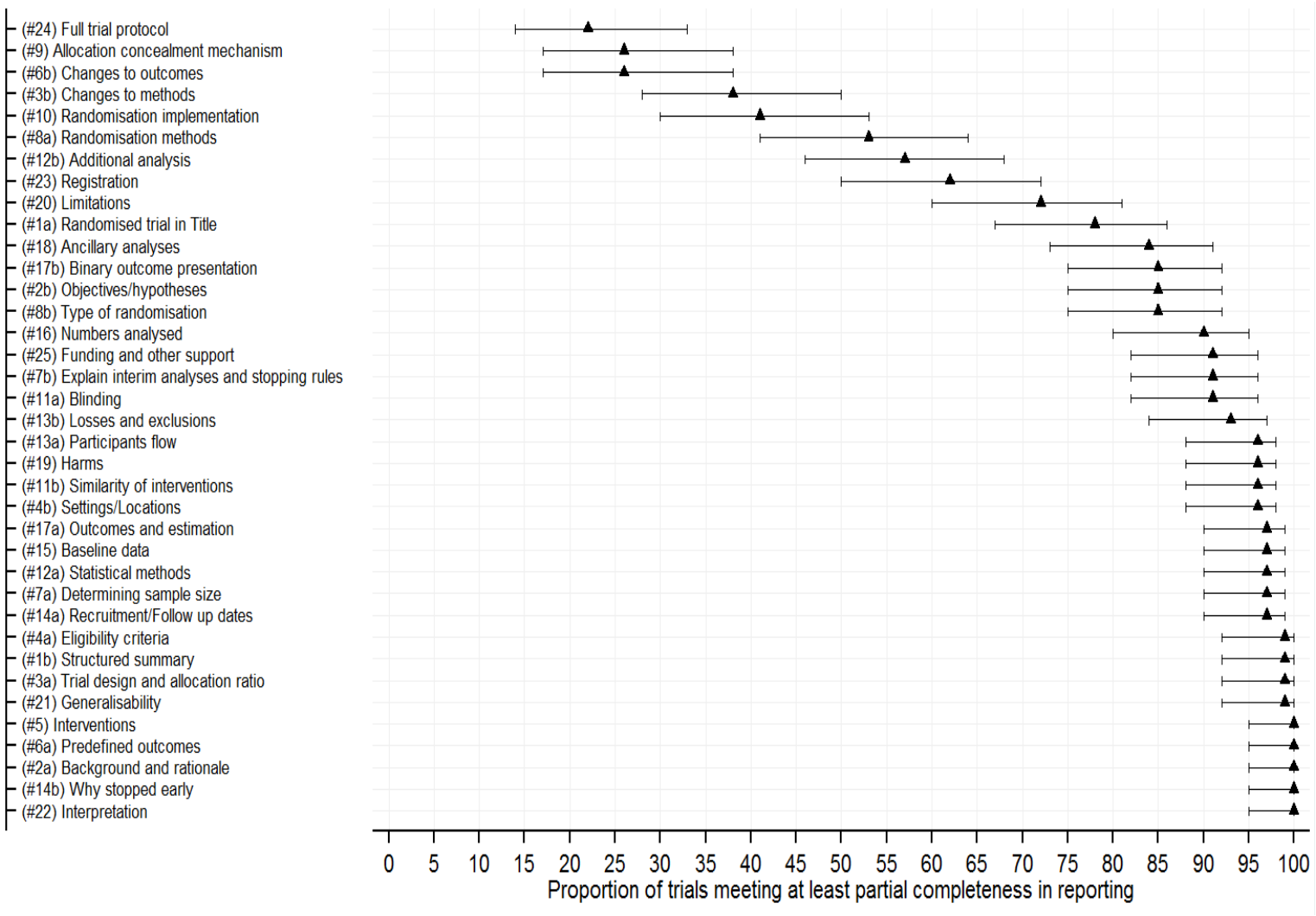

Figure 6.4. Trials meeting ‘at least partial’ reporting compliance of universal CONSORT checklist items.

The median [IQR] distribution of checklist items meeting ‘at least partial’ reporting compliance criteria was 30(81%) [29(78%) to 32(87%)], and a minimum of 24(65%). The median difference [95% CI] in items that met complete compliance in favour of journals that endorse the CONSORT statement as part of their publication policy was 1.5(4.1%) [-0.3(-0.9%) to 3.3(9.0%)]; p-value=0.112.

6.4.4 Reporting of Group Sequential Specific Items and Proposed Modifications

Figure 6.5 displays a multiple stacked bar chart of reporting compliance of GSD specific aspects as described in Section 6.3.4.

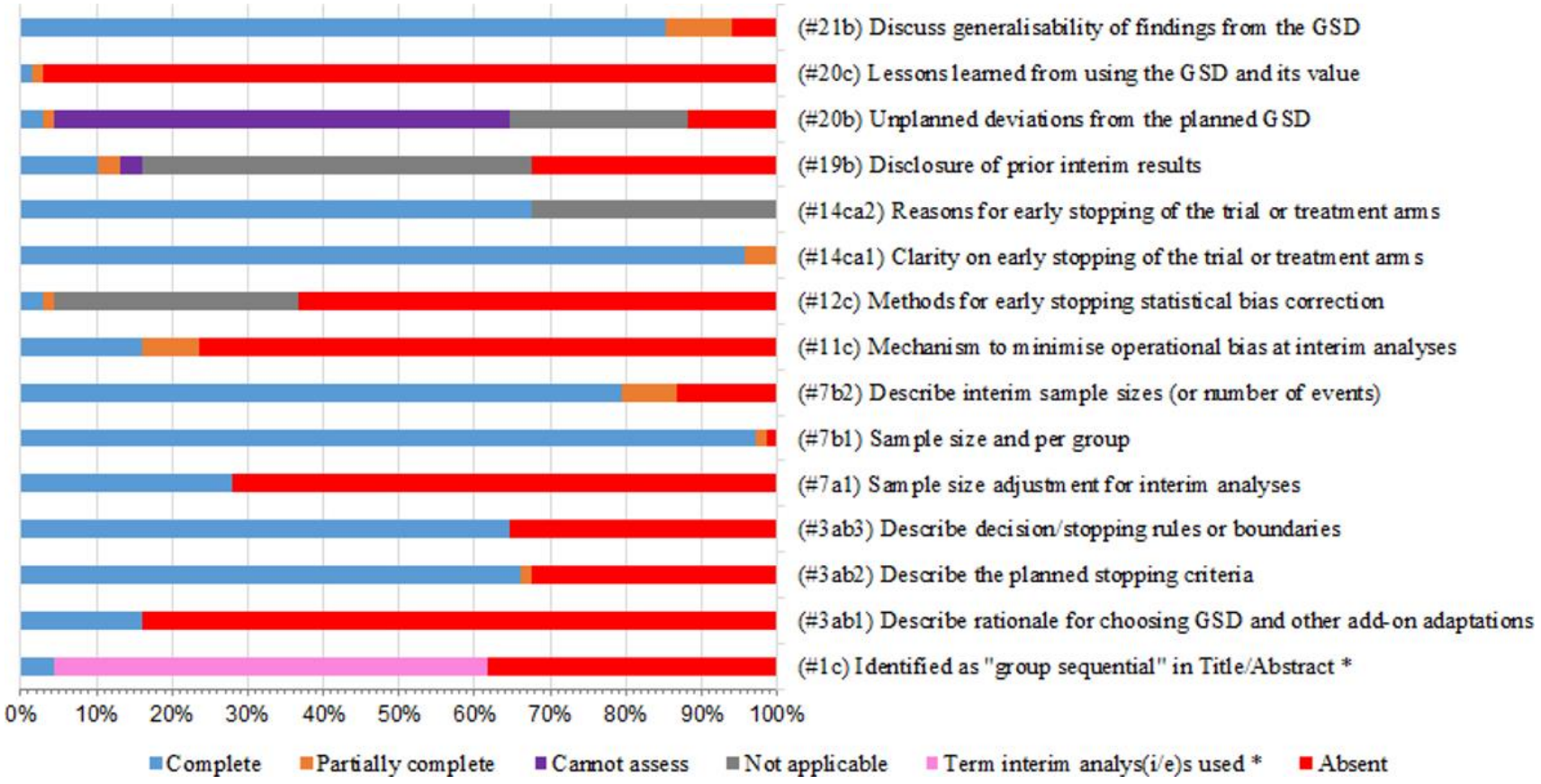

Figure 6.5. Reporting compliance of group sequential specific checklist items.

As evident in Figure 6.5, the reporting of most items relating to GSD specific aspects was suboptimal. Only 3(4%) group sequential RCTs were identifiable by the term “group sequential” in the Title or Abstract. An additional 39(57%) were identifiable by the terms “interim analyses” or “interim analysis”. The rationale for choosing a GSD (with any other add-on forms of trial adaptation) was only explained in 11(16%) group sequential RCTs.

Just 11(16%) group sequential RCTs adequately reported the mechanism used to minimise operational bias due to the knowledge or leakage of the interim results; 7 of these cited relevant prior publications. Of the 33 group sequential RCTs that were reporting interim results after the first interim, 9(27%) reported or disclosed prior interim results. Only 3 group sequential RCTs reported unplanned deviations from the planned GSD and their potential implications for the findings. However, unplanned deviations could not be assessed for in 41(60%) group sequential RCTs, due to inaccessibility of protocols and associated amendments. Only 2 group sequential RCTs described the lessons learned from using the GSD, and their value to enhance the planning of future group sequential RCTs. Table 6.2 summarises supporting data of reporting compliance of group sequential specific aspects.

Table 6.2. Summary data of reporting compliance of group sequential specific aspects.

Checklist item	Completeness in reporting of group sequential specific aspects						
	Partially/Totally Complete	Absent	Totally Complete	Partially complete	Cannot assess	Not applicable	Interim analys(i/e)s
(#1c) Identified as "group sequential" in Title/Abstract * §	3(4%)	26(38%)	3(4%)	-	-	-	39(57%)
(#3ab1) Describe the rationale for choosing the group sequential design (and other add-on adaptations) §	11(16%)	57(84%)	11(16%)	-	-	-	-
(#3ab2) Describe the stopping criteria employed	48(70%)	20(29%)	47(69%) †	1(1%)	-	-	-
(#3ab3) Describe the stopping rules employed	44(65%)	24(35%)	44(65%) †	-	-	-	-
(#7a1) Sample size adjusted for interim analyses §	19(28%)	49(72%)	19(28%)	-	-	-	-
(#7b1) Sample size and per group	67(99%)	1(1%)	66(97%)	1(1%)	-	-	-
(#7b2) Describe interim sample sizes (or number of events)	59(86%)	9(13%)	54(79%)	5(7%)	-	-	-
(#11c) Measures to minimise operational bias due to leakage or knowledge of interim results §	16(23%)	52(76%)	11(16%) ϙ	5(7%)	-	-	-
(#12c) Describe use of statistical methods for early stopping bias correction §	3(4%)	44(65%)	2(3%)	1(1%)	-	21(31%)	-
(#14ca1) Clarification on whether the trial (or treatment arms §) were stopped early	68(100%)	-	65(96%)	3(4%)	-	-	-
(#14ca2) Reasons for early stopping of the trial (or treatment arms §)	46(68%)	-	46(68%)	-	-	22(32%)	-
(#19b) Describe prior interim results when applicable §	9(13%)	22(32%)	7(10%)	2(3%)	2(3%)	35(51%)	-
(#20b) Described unplanned deviations from the planned group sequential design §	3(4%)	8(12%)	2(3%)	1(1%)	41(60%)	16(24%)	-
(#20c) Discuss lessons learned and the value of the group sequential design §	2(2%)	66(97%)	1(1%)	1(1%)	-	-	-
(#21b) Discuss generalisability of the results from the group sequential design §	64(94%)	4(6%)	58(85%)	6(9%)	-	-	-

† 11 trials described stopping boundaries/rules and/or stopping rules elsewhere in cited material;

ϙ 7 trials described measures to minimise operational bias due to leakage or knowledge of interim results elsewhere in cited material;

* 39(57%) trials identified by the terms "interim analysis" or "interim analyses";

§ Marked items (or parts) are researcher-led proposed modifications.

Adequately reported group sequential specific aspects include description of the total (and per group) sample size, the planned number of interims and associated interim sample sizes (or number of events), and clarification on whether the trial was stopped early and reasons where applicable. Description of the planned stopping criteria and rules or boundaries were moderately well reported.

6.4.4.1 Early Stopping of Trials or Treatment Arms

Of the 68 group sequential RCTs, 46(68%) were stopped early, predominantly for futility, 61% (28/46), and efficacy, 22% (10/46). The proportion of group sequential RCTs that stopped early for any reasons before and after 2010 appeared to be similar; 22(65%) versus 24(71%), respectively: RR (95% CI, p-value); 0.92(0.66 to 1.27, p-value=0.796). Of the 22 group sequential RCTs which were not stopped early, 6(27%) had multiple intervention arms. Of these 6 group sequential RCTs, 2 had discontinued one intervention arm at previous interim analyses. In 46 group sequential RCTs that were stopped early, the median (IQR) of the distribution of the proportion of interim sample size (or observed interim events) at the time of trial stopping relative to that planned was 65% (50% to 85%), and a minimum of 19%.

6.4.4.2 Type of Planned Stopping Criteria

Stopping criteria and rules or boundaries planned at the design stage, were unreported in 22(32%) and 24(35%) group sequential RCTs, respectively. Of the group sequential RCTs that reported planned stopping criteria and/or stopping boundaries, 11(16%) cited additional relevant information accessible in the form of prior publications or protocols. Thirty-three (49%) group sequential RCTs were planned with at least some form of futility early stopping criteria; 9(13%) for futility only, 20(29%) for either futility or efficacy, and 4(6%) for futility, efficacy or safety.

6.4.4.3 Type of Planned Stopping Rules or Boundaries

Table 6.3 summarises the criteria used to make early stopping decisions at interim analyses and associated methodological references. Twenty-four (35%) group sequential RCTs did not disclose the stopping rules or boundaries used. Of the 44(65%) group sequential RCTs that reported stopping rules or boundaries, the most frequently used stopping boundaries were: 15(34%) LD (1983) error spending function mimicking OBF (1979) type properties, and 12(27%) OBF. These have been described in Section 2.7.5 of Chapter 2.

Table 6.3. Type of stopping boundaries utilised in those with complete information.

Type of stopping boundaries	Total (N=44)
LD (1983) error spending function of the OBF (1979) type	15(34.1%)
OBF (1979)	12(27.3%)
HP (Haybittle, 1971; Peto, Pike, Armitage, et al., 1977)	3(6.8%)
Based on CP or number of cases or hazard ratios	3(6.8%)
Pocock (1977)	2(4.5%)
PT (1994), in combination with LD error spending function of the OBF type	1(2.3%)
PT (1994)	1(2.3%)
LD (1983) error spending function	1(2.3%)
Gamma family (γ=-8) (Hwang et al., 1990)	1(2.3%)
Rho family (ρ=3) (Jennison and Turnbull, 2000g)	1(2.3%)
Fleming (1982) in combination with LD error spending function of the OBF type	1(2.3%)
WT (1987) (δ = 0) in combination with OBF	1(2.3%)
Whitehead double triangular (Whitehead, 1999)	1(2.3%)
Lee type †	1(2.3%)

CP: Conditional power; LD: Lan and DeMets; HP: Haybittle–Peto; OBF: O'Brien and Fleming; PT: Pampallona and Tsiatis; WT: Wang and Tsiatis; † unclear details

6.4.4.4 Number of Planned Interim Analyses and Stage of Reporting

The majority (75%) of group sequential RCTs were planned with either one or two interim analyses. There were very few group sequential RCTs planned with large numbers of interim analyses (Table 6.1). Only 5(7%) RCTs did not report the number of planned interim analyses. Fifty-five (81%) group sequential RCTs were reporting interim results; of which, 47(69%) were as intended. Poor recruitment and/or financially related issues were the main reasons for reporting unplanned interim results in the remaining 8(12%) group sequential RCTs.

6.4.4.5 Early Stopping Statistical Bias Correction

Forty-six (68%) group sequential RCTs were stopped early at an interim analysis. Of these, only 3(7%) reported the use of appropriate statistical methods for bias correction of point estimates of the intervention effect, and associated CIs and P-values; 2 of these were stopped early for futility and/or safety. Only 1 of the 10 group sequential RCTs that were stopped early for efficacy reported the use of bias corrected statistical methods to conduct inference. These methods have been described in Section 2.7.11 of Chapter 2.

6.4.5 Exemplars to Enhance the Reporting of Group Sequential Trials

Although there were no publications that met complete compliance on all checklist items, there are a few exemplars of group sequential RCTs that reported most items adequately (Butts et al., 2014; Middleton et al., 2014; Tröger et al., 2013). The PRIMO trial is an exemplar that provided a comprehensive rationale of choosing,

and detailed description of a GSD incorporating a SSR (an information based GSD) and also cited a prior relevant publication (Pritchett et al., 2011; Thadhani, Wenger, et al., 2012). Some publications described aspects of the randomisation process, allocation concealment and its implementation, which were found to be problematically reported in most group sequential RCTs, better than others (Middleton et al., 2014; Wolff et al., 2013).

Mehta et al (2012) is an exemplar that reported detailed description of protocol changes. Chew et al (2014) gave a useful exemplar of the description, and graphical representation, of prior interim trends of the intervention effect, and explored the trends using piecewise linear regression. In the next chapter, Figure 7.10 and Figure 7.11 demonstrate graphically interim trends of results superimposed onto the stopping decision-making boundaries when results are only presented at or after the point of early stopping.

Roger et al (2013) gave a clear description of an exact statistical method, used to obtain unbiased inference following early stopping of group sequential RCTs and referenced the underlying methodology:

> "*The study design is based on a group-sequential test procedure with pre-planned analyses after 220, 320 and 428 patients meeting one of the off-study criteria. An alpha-spending approach as suggested by Lan and DeMets* (1983) *with an OBF-like alpha spending function was used to define the test boundaries of the group-sequential procedure. The primary analysis regarding OS uses a Cox Proportional Hazard Model with treatment and prognosis groups as predictor variables to calculate the Z score needed for the group- sequential procedure. Stagewise ordering was used to compute the unbiased median estimate and confidence limits for the prognosis-group-adjusted hazard rates* (Emerson and Fleming, 1990)."

Mascia et al (2010) is an exemplar that provided an explanation of deviations from the planned interim analyses and the implications on their findings. A couple of group sequential RCTs somewhat discussed lessons learned from, the value of, and implications of using a group sequential approach (Markman et al., 2003; Moore et al., 2003).

6.5 Discussion

This section presents key findings and interpretation in the context of existing literature. The implications of the findings are discussed in relation to the use of confirmatory ADs. Some research strengths and limitations are highlighted and a road map of the direction of the research summarised.

6.5.1 Main findings

A number of confirmatory group sequential RCTs are utilised in oncology, although the therapeutic areas of application are diverse. Most of these group sequential RCTs are published in 'high impact' peer reviewed medical journals. Moreover, a sizable number are stopped early, predominantly for futility followed by efficacy.

The findings highlight inadequate reporting of CONSORT checklist items relating to the disclosure and access to trial protocols, methods used to generate the randomisation list(s), details of randomisation concealment and its implementation, and other trial design aspects. Most concerning, is the lack of access to full trial protocols, with related amendments, in the public domain, for most of the examined group sequential RCTs. As a result, reporting compliance of some key checklist items such as changes to methods and outcomes could not be examined. Despite this, reporting compliance of most universal CONSORT 2010 checklist items appeared optimal.

Equally important, however, the findings highlight very poor reporting of important features of group sequential RCTs considered. These include:

a) Rationale for choosing a GSD with any other additional planned adaptations,
b) Mechanisms put in place to minimise operational bias due to the knowledge of interim results,
c) The use of appropriate statistical methods to obtain unbiased or bias-corrected results (point estimates, CIs and P-values),
d) Lessons learned from using a GSD and their value to help the planning of future related trials.

In addition, reporting compliance of planned stopping criteria, and stopping rules or boundaries used, is still unsatisfactory, even though these aspects are partly covered in the CONSORT 2010 statement, but not as standalone items. Nonetheless, there is adequate reporting compliance of some aspects of interim analyses covered by the CONSORT 2010 statement. These include clarification of early stopping with reasons where applicable and the description of interim sample sizes (or number of events), number of planned interim analyses, and timing.

Table 6.4 is a preamble summarising additional general checklist items which should be considered when developing a CONSORT guidance tailored for adaptive randomised trials. These general additional items are marked in purple as items xx or xxx. The preamble template has been adapted from Moher et al (2010). Extensions specific to certain ADs such as SSR, GSD, MAMS, population enrichment, and operational and inferential seamless designs should also be considered.

Table 6.4. Modified CONSORT checklist of information to include when reporting an adaptive randomised trial.

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract	1a	Identification as a randomised trial in the title	
	xx	Identification as an adaptive design in the title and/or abstract	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	
	xx	Description of the type of AD used and its scope	
	xx	Rationale on why an AD was considered and its appropriateness to address research questions	
	xx	Clear description of the interim decision-making criteria guiding the adaptation(s) and decision-making process	
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	
	xx	Describe any important changes to the design or methods outside the scope of the planned AD	
Participants	4a	Eligibility criteria for participants	
	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	
Sample size	7a	How sample size was determined	

	xx	Whenever simulation was used, provide rationale for the simulation scenarios considered. Reference an accessible simulation protocol and report, and statistical programs or code used	
	7b	Explanation of any interim analyses, planned sample sizes, and stopping guidelines	
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)	
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	
	11b	If relevant, description of the similarity of interventions	
	xxx	Description of the systems, procedures, and processes put in place to minimise operational bias during the course of the trial due to knowledge of the interim results	
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	
	xxx	Description of the appropriate statistical methods used to account for the implemented trial adaptations in order to obtain unbiased or bias adjusted results (point estimates, CIs and p-values) when appropriate	
	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 at the point of interim analysis	
	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	
	xxx	Why recruitment in certain treatment arms was stopped where appropriate	

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% CI)	
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	
	xxx	Provide an accessible reference of previous interim primary results where appropriate	
	xxx	When a trial is stopped early as part of the adaptation process, provide a figure showing a trend of the primary results, point estimates with Cis, up to the time of interim early stopping	
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)	
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	
	xx	Discuss the generalisability of the results from the adaptive trial and to whom the results pertain to	
	xx	Discuss the lessons learned from using the implemented AD to help the design and planning of future related trials	
Interpretation	22	Interpretation consistent with results (in the context of the planned decision-making criteria and stopping rules), balancing benefits and harms, and considering other relevant evidence	
Other information			
Registration	23	Registration number and name of trial registry	
Protocol	24	Where the full trial protocol can be accessed, if available	
Simulation protocol and report	xx	Where the simulation protocol and report can be accessed when trial simulation was used to estimate the sample size and explore the statistical characteristics (such as type I error and power) and operational characteristics of the implemented AD	
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	

6.5.2 Interpretation of the findings

The findings are predominantly based upon group sequential RCTs published in 'high impact' medical journals and in oncology. Therefore, the general quality of reporting compliance may exaggerate what might be observed based on reports in other therapeutic areas, or lower impact journals for some checklist items (Hurst, 2011; Sjögren and Halling, 2002; Yao et al., 2014). For example, suboptimal reporting compliance to most checklist items has been highlighted in previous reviews in other therapeutic areas (Hurst, 2011; Sjögren and Halling, 2002). Regardless of the impact factor of the publishing journal, trial design and therapeutic area; suboptimal reporting of randomisation methods, and details of randomisation concealment and its implementation has been widely reported and is consistent with the findings of this chapter (Camm et al., 2013; Hurst, 2011; Mhaskar et al., 2012; Sjögren and Halling, 2002; Yao et al., 2014). Similar findings on these checklist items were also found in oncology, and moreover, inadequate reporting was associated with exaggerated, biased intervention effects (Mhaskar et al., 2012).

Most importantly, the findings of very poor reporting and use, of statistical methods to obtain unbiased or bias-corrected results, are consistent with a previous systematic review, focusing on trials which stopped early for benefit (Montori et al., 2005). This aspect is one of those which have been overlooked by Detry et al (2012) who proposed some modifications to the CONSORT statement. The thesis findings on areas requiring improvement provide a conservative picture of the scale of the problem regarding reporting compliance of group sequential specific aspects, which are vital for research consumers to make informed judgements about the quality of research findings.

6.5.3 Implications to Practice

This review uncovered group sequential RCTs published predominantly in 'high impact' medical journals. To some extent, this may provide assurance to sceptical researchers, who may have concerns pertaining to poor receptiveness by Journal Editors and Reviewers towards ADs, when trials are stopped early. In contrast, suboptimal reporting of appropriate statistical methods for early stopping bias correction, may influence some research consumers, who are aware of the phenomenon of exaggerated intervention effects when a naïve statistical approach is used, to consider findings from group sequential RCTs with a degree of scepticism. Research consumers, such as Clinicians and Regulators, may be reluctant to accept findings in order to change medical

practice, when trials are stopped early coupled with failure to implement bias correction, and poor communication and reporting of the corrective actions taken (Dimairo, Boote, Julious, Nicholl, et al., 2015; Montori et al., 2005).

The phenomenon of exaggeration of the intervention effects in group sequential RCTs, following early stopping when a naïve statistical approach is used, has been widely debated and highlighted (Bassler et al., 2008, 2010; Freidlin and Korn, 2009; Montori et al., 2005; Wears, 2015; Wittes, 2012; Zannad et al., 2012; Zhang et al., 2012). Although much attention has been paid to group sequential RCTs that are stopping early for benefit (Bassler et al., 2008; Montori et al., 2005; Zannad et al., 2012; Zhang et al., 2012), the consequences could be similar when trials are stopped early for futility, since the evidence can be used to withdraw intervention(s) already in the care pathway. More so, it could be argued that the consequences on future evidence synthesis, through meta-analysis, should be treated similarly regardless of the reasons for early stopping.

Chapter 2 (Section 2.7.11 onwards) described a number of proposed statistical methods to conduct inference following a GSD. Despite the existence of these methods, this chapter highlights that the methods are rarely used or reported in practice. The extent of the impact of statistical bias correction, on the results and decision-making of these group sequential RCTs, particularly those that are stopped early is unclear. Montori et al (2005) illustrate some RCT examples, where interpretation of findings changed after bias correction. Wittes (2012) reports a trial that produced consistent interpretation of findings when both naïve and statistical bias correction methods were used. In addition, the results of case studies in Chapter 7 are consistent despite the method used. The lack of knowledge of the impact that inaccurate analysis has on decision-making among Statisticians, lack of awareness of bias adjustment methods, and perhaps the unfamiliarity with mainstream statistical software(s) offering options to implement these procedures, could be contributing factors to their poor uptake.

Findings from this chapter support initiatives for mandatory publication of, not only full trial protocols, but also all related protocol amendments. Assessing the quality of reporting of some key aspects of group sequential RCTs, proved challenging without access to these important trial documents, which is most imperative for complex ADs. This is consistent with previous findings highlighting that the assessment of methodological quality should be based on evaluation of both protocols and publications (Mhaskar et al., 2012).

As highlighted in Chapter 3, interim analyses heighten anxiety among some research consumers, due to the potential for introducing operational bias to trial conduct, which undermines the scientific integrity and validity of the findings. The potential introduction of operational bias, due to the leaking or knowledge of interim results in adaptive trials, has been well described (Chow and Chang, 2008; Chow and Corey, 2011). However, Detry et

al (2012) highlight that the extent and impact of operational bias on the findings and decision-making is less well-understood in practice. Therefore, it is imperative to report mechanisms put in place, to minimise and/or control for operational bias such as who conducted the interim analyses, how data were transferred and results communicated, who the stakeholders were in the interim decision-making process, and how decisions were made. Although it is difficult to prevent indirect inference of interim results, due to decisions made following an interim analysis, the authors state that careful planning, implementation, and optimal reporting of the mechanisms put in place, may go a long way in alleviating research consumers' worries about operational bias.

The poor reporting or inaccessibility of prior interim results at the interim reporting or time of early stopping may hinder the ability of consumers of research findings, to assess trends of the direction of the treatment effects and the potential effect of population drift. Even though it is challenging to distinguish between natural and population drift induced by operation bias, access to prior interim results may help research consumers to make their own informed judgements, and alleviate some of the cited concerns.

6.5.4 Strengths and Limitations

This chapter is based upon concerns raised by key stakeholders which have been reported in Chapters 3 and 4, which have been already published elsewhere (Dimairo, Boote, Julious, Nicholl, et al., 2015; Dimairo, Julious, Todd, Nicholl, et al., 2015). The examination of the completeness in reporting used all accessible publication related reports, using an improved classification system, and with the support of an independent Reviewer for quality control. In addition, exemplars that can be used by Clinical Trialists as a resource to enhance adequate reporting of group sequential trials are highlighted.

One of the main limitations is that the literature search was restricted to Ovid MEDLINE due to resource and time limitations. Moreover, the poor indexing of group sequential RCTs means that the Ovid MEDLINE search could have missed a number of eligible trials. However, group sequential trials identified in Chapter 5 supplemented systematically reviewed trials. Inaccessibility of trials protocols and associated amendments hampered the examination of some key CONSORT checklist items, such as changes to methods and outcomes. Lastly, the factors associated with suboptimal reporting were not explored due to time constraints.

6.5.5 Summary and Direction of the Remainder of the Thesis

The findings highlighted in this chapter may partly explain cited concerns regarding the robustness in decision-making and acceptability of confirmatory ADs to change practice, when trials are stopped early. The assurance of scientific rigour through transparent and adequate reporting of adaptive trials is paramount to the acceptability of their findings. There is an urgent need for a CONSORT statement tailored for ADs. General recommendations based upon thesis findings and the desire to improve the planning of future trials, reproducibility of trials, the acceptability of findings from group sequential RCTs, and to reduce waste in trials research are presented in Chapter 9.

In the next chapter, retrospective planned case studies are used to illustrate the design and implementation of ADs with great potential in confirmatory trials. Potential benefits of these ADs, mainly in terms of patients and time savings, are explored. More so, robustness of the ADs considered is established by exploring consistency in decision-making with the trial results already known. Some findings such as the timing of interim analyses shall guide the work of Chapters 7 and 8 and the design of related future trials. Another motivation for the work of Chapters 7 and 8 is that some of the case studies identified so far did not give enough detail or depth for other researchers to adequately learn from them. Therefore, it is important to create further case studies based on reanalysis of retrospective and prospective planned adaptive trials.

Chapter 7. Design and Implementation of Retrospective Case Studies

7.1 Introduction

Chapter 4 highlighted a number of obstacles to the appropriate use of confirmatory ADs. Some of the barriers included: the lack of practical implementation knowledge; insufficient access to case studies; a lack of applied training, and challenges in marketing ADs to key stakeholders. Additional concerns included the robustness of ADs in decision-making and the credibility of trial results to change practice when trials are stopped early. Work described in Chapter 3 based on interviews of key stakeholders also underscored the need to raise awareness of opportunities and potential benefits of ADs in clinical trials research.

Section 4.5 of Chapter 4 highlighted that addressing the cited obstacles requires a number of approaches. Retrospectively designed and analysed case studies aided with simulation work may help demonstrate lost opportunities and provide some methodological reassurance of the robustness of ADs in decision-making. The case studies may also facilitate practical learning and enhance communication of statistical and operational aspects of ADs, and highlight some pitfalls during implementation and decision-making. Equally important, lessons learned will help to draw recommendations to enhance the planning of future adaptive trials of similar characteristics. This chapter endeavours to address some of the above cited obstacles by exploring retrospectively planned confirmatory ADs. The work of Chapters 2 and 6 guides the statistical implementation of case studies.

The chapter acknowledges the consent of Prof Steve Goodacre, Prof Elizabeth Goyder, and Prof Alasdair Gray to use anonymised trial data of studies undertaken in ScHARR.

7.2 Aims and Objectives

Through the retrospective design and analysis of a number of trials, this chapter aims to expand on the applied knowledge of ADs. The focus is on the most commonly used ADs with significant potential in the confirmatory setting based on survey results (Chapter 4) and review of case studies (Chapter 5). These ADs include SSR, futility analysis based stochastic curtailment using CP, and GSD. More so, Chapters 3, 4, and 5 reiterated the importance given to the minimisation of operational bias during trial conduct introduced by the knowledge of interim results. Hence, SSR will be conducted in a blinded manner. In addition, it is the approach most recommended in the literature and favoured by Regulators as reflected in Chapter 2. Based on the literature

review findings in Section 2.6, one interim futility analysis based on stochastic curtailment using CP will be used because the method has negligible impact on the statistical properties of the design. GSDs are well developed with known statistical properties and are accepted by the research community and Regulators.

The specific objectives of the chapter are to:

1) Demonstrate the design and analysis aspects of ADs;
2) Illustrate lost opportunities and potential benefits of certain types of ADs, particularly in terms of reduction in trial duration and savings in trial participants;
3) Explore the pitfalls, limitations, and robustness of particular types of ADs in decision-making using retrospectively planned case studies given that the final trial results are known. Here, interim results and decision-making are checked for consistency against the final results;
4) Draw recommendations from lessons learned to help the planning of future adaptive ADs with similar characteristics.

7.3 Brief Description of the Retrospective Case Studies

The trials described here were originally designed with fixed sample sizes and have been completed with the main results published elsewhere. Detailed background information about the trials and their findings are given in cited references. In this chapter, the trials are redesigned retrospectively as if they were pre-planned ADs, depending on the type of AD considered as guided by the literature review findings in Chapter 2.

7.3.1 RATPAC Trial

Goodacre et al (2011) conducted the RATPAC trial and report the main findings. The study was a two arm, multicentre, parallel group, superiority, pragmatic, open-label RCT investigating the effectiveness of a Point-of-care (PoC) intervention against Standard Care (SC) in increasing successful hospital discharge of participants with suspected Myocardial Infarction. The primary endpoint was successful hospital discharge within 4 hours of attendance with no major adverse event during the following 3 months. A major adverse event was defined based on some criteria described in the study protocol. The investigators assumed a 50% successful hospital discharge in the SC arm. A 5% increase in successful hospital discharge was viewed as clinically relevant to detect in order to declare superiority in favour of the PoC or SC arm. This corresponds to an OR of 1.22 or RR of 1.10. With these assumptions, the research team justified the planned recruitment of 3130 participants (1565 per arm) to

preserve 80% power and a 5% two-sided type I error. The research team managed to enrol 2243 participants before research funding ran out. The trial was terminated early after the Public Funders declined the request for an extension. Despite this, the final analysis based on 2243 participants showed superiority of the PoC intervention in increasing successful hospital discharge.

7.3.2 3CPO Trial

Gray et al (2008, 2009) report the main results of the 3CPO trial. The study was a three arm, multicentre (involving 27 emergency departments), open-label, parallel group, superiority RCT investigating whether the two non-invasive ventilation procedures: non-invasive intermittent positive-pressure ventilation (NIPPV) and continuous positive airway pressure (CPAP) improve survival in participants with Acute Cardiogenic Pulmonary Edema. The primary comparison was to test standard oxygen therapy (SOT) versus CPAP or NIPPV in reducing mortality observed within 7 days. A secondary outcome was to determine whether NIPPV is superior to CPAP based on a composite outcome of mortality or intubation within 7 days. The research team assumed a 15% mortality rate in the SOT arm and considered a 6% absolute reduction in mortality within 7 days to be clinically relevant, to declare superiority in favour of CPAP/NIPPV or SOT. A 6% absolute mortality difference relative to 15% translates to an OR of 0.56 or RR of 0.60. The research team justified the recruitment of a total sample size of 1200 patients (400 per arm) to preserve 80% power and a 5% two-sided type I error [1]. The trial recruited 1069 participants within the planned trial duration before research funding ran out. Final analysis based on 1069 participants showed no statistically significant difference in mortality between the CPAP or NIPPV and SOT arms. Furthermore, no statistically significant difference in mortality or intubation within 7 days was observed between the CPAP and NIPPV arms.

7.3.3 3Mg Trial

Goodacre et al (2013, 2014) report the primary findings of the 3Mg trial, which was aimed to determine whether intravenous (IV) or nebulised (NEB) magnesium sulphate ($MgSO_4$) should be a standard first-line intervention for patients with acute severe asthma compared to a placebo. 3Mg study was a three arm, multicentre RCT involving 34 emergency departments in acute hospitals, double-blind and, and placebo-controlled. The trial

[1] Note: estimated planned sample size of 1200 could not be accurately replicated.

tested whether IV or NEB $MgSO_4$ could reduce the proportion of participants who required admission at initial presentation or during the following week, and/or improve the participants´ assessment of their breathlessness over two hours after initiation of treatment. The trial was designed with co-primary outcomes: a) admission to hospital at presentation or any time within a week, and b) visual analogue scale (VAS) breathlessness over two hours after the intervention.

Although these co-primary endpoints were considered when the trial was designed, discussions with the investigators highlighted that admission to hospital at presentation or any time within a week was the most important. The research team assumed 80% of patients with severe acute asthma were admitted after emergency department management and hospital admission is recorded for all patients. In addition, the investigators sought to detect a clinically relevant 10% absolute reduction in the proportion of admitted participants between any pair of intervention arms compared to the placebo; from 80% to 70%, corresponding to an approximate OR of 0.58 or RR of 0.88. Under these assumptions, the research team aimed to recruit 1200 participants (400 per arm) to preserve a power of 90% to detect a 10% absolute reduction in hospital admission for any pair of comparisons compared to placebo at 5% two-sided significance level. The research team however achieved a recruitment of 1109 participants of the targeted 1200 when funding ran out. The study did not show a clinically meaningful difference in hospital admission from either IV or NED $MgSO_4$ compared with a placebo.

7.3.4 Booster Trial

The Booster study was a three arm, parallel group, pragmatic, superiority RCT. The trial investigated whether objectively measured physical activity, 6 months after a brief intervention, is increased in those receiving physical activity ‘booster’ consultations delivered in a motivational interviewing style, either face-to-face (‘full booster’) or by telephone (‘mini booster’) compared to the control (‘no further intervention’) (Goyder et al., 2014). Booster sessions were delivered at 1 and 2 months post-randomisation. In addition, the trial had an internal pilot assessing operational feasibility and estimation of the primary outcome variability to anchor the effect size. The primary outcome was total energy expenditure (TEE) per day in kcal from 7-day accelerometry at 6 months post-randomisation.

The research team viewed that a 1/3 SD increase in TEE was clinically relevant to declare superiority in favour of the combined booster interventions. Assuming an approximate 25% loss to follow-up by three months, the research team aimed to recruit 600 participants (200 per arm) to yield 90% power to detect a 1/3 SD mean

difference in TEE between the combined booster interventions and control arm as statistically significant at 5% two-sided type I error. However, the trial struggled with recruitment and only managed to recruit 240 participants (40% of the target) before research funding ran out.

7.4 Methods

This section describes the approaches used to redesign and reanalyse case studies as guided by the literature review findings and recommendations in Chapters 2 and 6.

7.4.1 Sample Size Re-estimation for Binary Outcomes

Here, the degree of uncertainty around the assumed nuisance parameters for sample size estimation at the design stage was investigated. For illustrative purposes, nuisance parameters and sample sizes were re-estimated after the enrolment of every one participant (sequentially). This enabled exploration of the performance of the method as the number of participants increased in order to learn lessons for the future.

The re-estimation of nuisance parameters was carried out for the pooled event rate (in a 'blinded' manner) rather than for the control event rate. As highlighted under Section 2.5 of Chapter 2, this approach minimises operational bias and can be simply implemented by in-house Trial Statisticians. The sample size was re-estimated using versions of equation (2:4), depending on the allocation ratio under consideration. This process was performed just once at a single interim, however for illustration, sample size was re-estimated after enrolment of every one participant. The summary statistics of re-estimated nuisance parameters and sample sizes are presented. assuming a minimum threshold of the number of participants required to yield stable and reliable estimates. The summarised average estimates obtained after imposing the minimum threshold are benchmarked against the assumptions made at the design stage. The minimum threshold imposed forms part of lessons learned for a reliable blinded SSR procedure for binary outcomes to complement literature review findings.

The influence of the blinded SSR conducted in an 'unrestricted' manner was explored and the number of trials that made reasonably accurate assumptions on nuisance parameters are reported. 'Unrestricted' SSR was considered when sample size review was the only trial adaptation. All analyses were performed in Stata 14.1.

7.4.2 Conditional Power Based Stochastic Curtailment Futility Analysis

This section was guided by the literature review findings in Section 2.6 of Chapter 2. For illustration, the CP was calculated once after enrolment of every participant during the entire trial duration conditional on the planned sample size. Computation of the CP was performed under a number of assumptions regarding the future trend of unobserved outcome data of participants yet to be recruited or whose outcomes are yet to be observed. That is, the future unobserved trend is assumed to have an effect under H_0, $H_{0.5}$, H_1 or observed interim effect. $H_{0.5}$ represents half the effect assumed under H_1. The CP trends under these assumptions are displayed over the entire trial duration. The trend in type I error is presented for illustration only for the first case study. The robustness of futility early stopping decision-making is examined for thresholds ranging from less than 10% to 30%, as informed by the literature review findings. The influence of the timing of the futility analysis is also explored. The potential benefits in terms of patient and trial duration savings were investigated under a number of scenarios. All analyses were performed in Stata 14.1 using a code written by the author, which shall be made publicly available as a Stata and R package.

7.4.3 Group Sequential Design for Binary Outcomes

The literature review findings in Section 2.7 of Chapter 2 guide the statistical design and implementation of a GSD. For illustrative purposes, trials were redesigned with two-sided group sequential superiority tests as planned for the primary studies to allow for options to stop either for futility or efficacy although in practice the choice is trial dependent. Unless stated otherwise, the most common stringent efficacy boundaries constructed using LD spending functions similar to OBF type highlighted in Section 6.4.4.3 of Chapter 6 were used. 'Non-binding' futility stopping boundaries were used in the sense that they can be overruled during interim decision-making without undermining the type I error and power. The inner wedge 'non-binding' futility boundaries constructed using LD spending functions extended by Pampallona and Tsiatis (Section 2.7.5 of Chapter 2) were used to allow for futility early stopping. Simulations were performed for sensitivity analysis, for instance, to understand the statistical properties and performance of the design.

Since the trials considered here had binary primary endpoints, monitoring was based on the number of participants with outcome data relative to the planned sample size. This approach works well when the design assumptions on nuisance parameters are accurate. Otherwise, the power of the study cannot be precisely controlled as desired. The influence of inaccurate assumptions on nuisance design parameters was investigated through

simulations. Importantly, an alternative information based GSD discussed in Section 2.8 of Chapter 2 is illustrated using RATPAC trial in Section 7.5.4. This trial was selected because the design assumptions were markedly inaccurate.

Median unbiased results at the time of early stopping using the stagewise ordering described in Section 2.7.11 onwards of Chapter 2 are presented together with naïve estimates unless stated otherwise. The design and implementation was conducted in East 6.3 and cross-validated in ADDPLAN 6.1. Enhanced plots were constructed in Stata 14.1. Potential lost opportunities in terms of savings; participants and trial duration were explored. Finally, lessons learned are highlighted.

7.5 Results

This section presents results of retrospectively designed and analysed case studies in the order; SSR, CP futility analysis based on stochastic curtailment, standard GSD, and information based GSD.

7.5.1 Sample Size Re-estimation for Binary Outcomes

In this section, only two case studies with contrasting findings are used to illustrate retrospective application of SSR. That is, RATPAC and 3Mg trials. The 3CPO trial, which had results consistent with the findings from RATPAC is provided as supplementary material in Appendix 7.1. SSR for the Booster trial is not applicable since the research team used a standardised effect size as a function of the SD.

7.5.1.1 RATPAC Trial

In Figure 7.1, the estimated hospital discharge in the SC and pooled arms are plotted against observed recruitment and compared to the assumed parameters at the design stage. In Table 7.1, the summaries of the pooled successful hospital discharge, total sample size, and likely overestimation of hospital discharge and sample size are presented. This assumes that SSR is performed at any single point after the recruitment of at least 300 participants. As noticeable from Figure 7.1 and Table 7.1, there are marked discrepancies between the assumed pooled (52.5%) and observed successful hospital discharge proportions. After a total recruitment of 300 participants in total (150 per group), the pooled successful hospital discharge has a mean (SD) of 23.5% (1.0%), corresponding to an overestimation of 29.0% (1.0%). The estimates of pooled hospital discharge rate seem unreliable and unstable during the recruitment of approximately the first 250 to 300 participants.

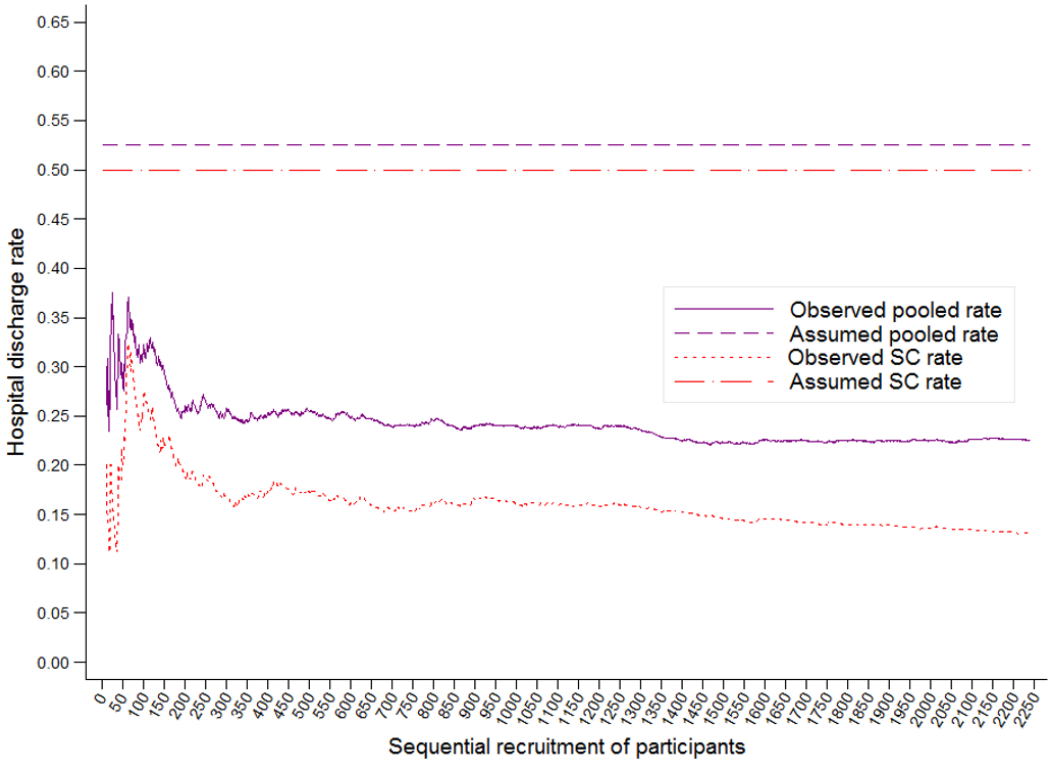

Figure 7.1. Uncertainty around assumed successful hospital discharge for RATPAC trial.

Figure 7.2 displays the pattern of the total re-estimated sample size. Summaries of the total re-estimated sample size assuming SSR is performed at any one point after the recruitment of at least 300 participants are shown in Table 7.1.

Table 7.1. Summary statistics of sample size re-estimation for RATPAC trial.

Variable	Summary Statistics		
	Median (IQR)	Mean (SD)	Min to Max
Overall hospital discharge (assumed =52.5%)	23.7% (22.5% to 24.1%)	23.5% (1.0%)	22.1% to 25.8%
Overestimation of hospital discharge	28.8% (28.4% to 30.0%)	29.0% (1.0%)	26.7% to 30.4%
Re-estimated total sample size (assumed=3131)	2271 (2190 to 2298)	2255 (68)	2163 to 2405
Overestimation in total sample size	861 (835 to 942)	878 (67.7)	728 to 970

SD: Standard Deviation; IQR: Interquartile range; Min: Minimum; Max: Maximum.

As evident, the potential overestimation in the total sample size has a median (IQR) of 861 (835 to 942) participants, ranging from 728 to 970. This assumes that the 5% difference in successful hospital discharge assumed by the research team at the design stage is still clinically relevant despite the overestimated pooled hospital discharge.

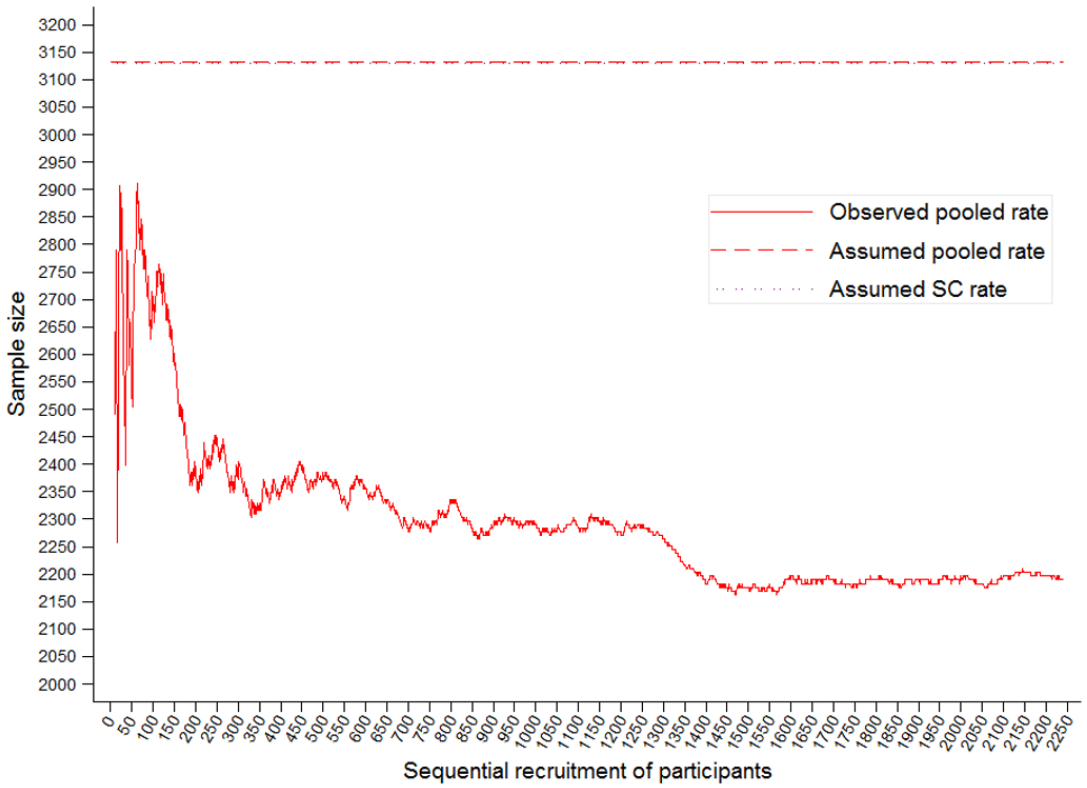

Figure 7.2. Pattern of the re-estimated total sample size for RATPAC trial.

Note: The planned sample sizes calculated using the assumed SC or pooled hospital discharge are similar, so the patterns are superimposed.

So far, a 5% risk difference has been assumed to be a fixed effect size to detect irrespective of the observed pooled or SC successful hospital discharge. This corresponds to a fixed OR of 1.22 in favour of the PoC in increasing successful hospital discharge from 50% to 55%. The question can be asked as to what would happen if investigators wanted to preserve an effect size on a fixed OR scale of 1.22 regardless of the observed event rate. Here, SSR was performed to preserve an OR of 1.22. That is, the absolute risk difference sought changes depending on the underlying observed event rate while the OR remains the same. Given the observed average pooled successful discharge of 23.5% rather than 52.5%, SSR would have resulted in a median increase (IQR) in sample size of 972 (918 to 1114). The RR would have changed from 1.1 under the design to 1.16 based on observed event rates. This corresponds to a risk difference of 3.8% (from 23.5% to 27.3%) for a fixed OR of 1.22 rather than the planned 5% (50% to 55%). This illustrates how sensitive the SSR procedure is to the operating scale of the effect size chosen when event rates are estimated with marked uncertainty.

In summary, the pooled successful hospital discharge for RATPAC trial was potentially overestimated at the design stage by more than 50% of the planned proportion. As a result, if the trial had been completed as

planned, it would have over recruited a mean (SD) of 878 (67.7) more participants than required to preserve a power of 80% to address the primary research objective. If research team proceeded to recruit as planned, the trial would have had 99.5% power. However, this is based on the premise that the research team stick to a 5% risk difference as planned regardless of the observed pooled or SC successful hospital discharge rates.

7.5.1.2 3Mg Trial

This example offers a contrasting perspective to the RATPAC SSR presented in Section 7.5.1.1. Figure 7.3 is a plot of the hospital admissions (assumed and observed in the control or pooled arms) against recruited sample size. As noticeable, the assumption of the 75% pooled hospital admission at the planning stage was relatively accurate. The observed mean (SD) pooled hospital admission was 77.1% (0.6%), translating to a relatively small underestimation of 2.1% (0.6%). This assumes that SSR is conducted at any single point after the recruitment of at least 300 participants in total.

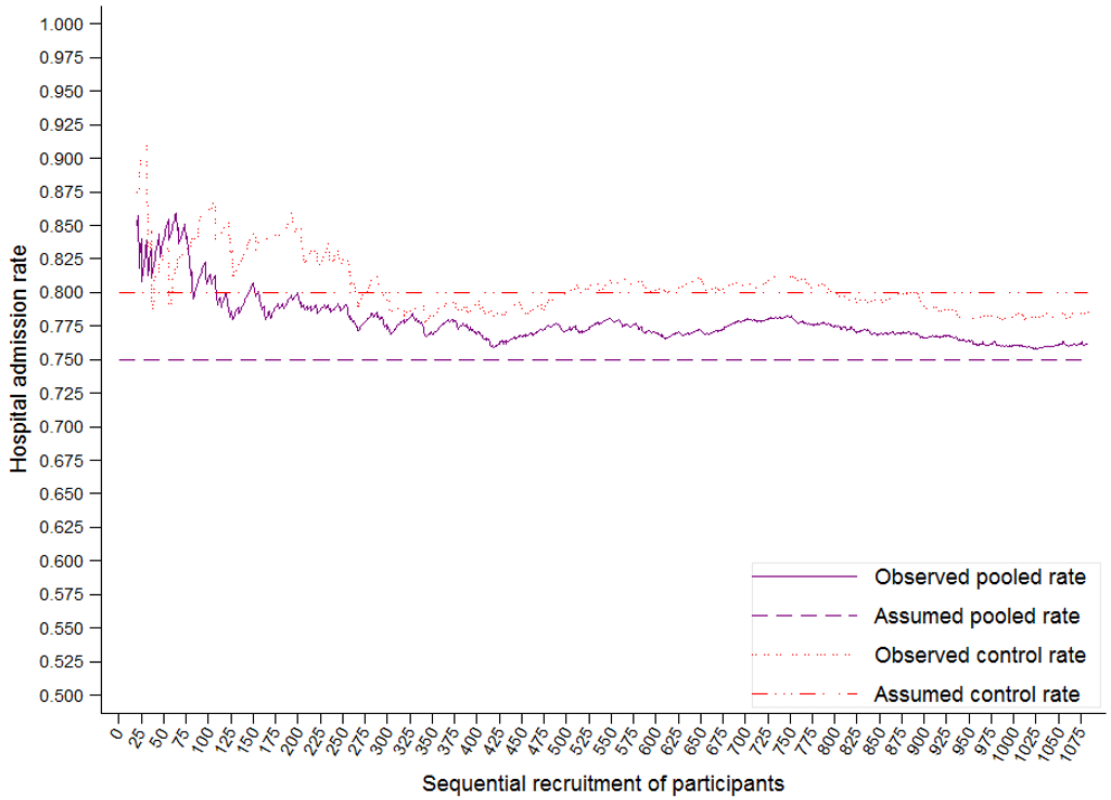

Figure 7.3. Uncertainty around assumed hospital admissions for 3Mg trial.

Figure 7.4 shows the pattern of the sample size re-estimated using the observed pooled hospital admissions shown in Figure 7.3. Here, equation (2:4) was invoked and sample size multiplied by 3 to preserve the

power of the pairwise primary comparisons against the placebo arm, assuming that a 10% absolute difference in hospital admissions is still clinically relevant to detect. Assuming SSR is performed at any single point after the recruitment of at least 300 participants, the median (IQR) re-estimated sample size is 1114 (1096 to 1131), ranging from 1068 to 1157. This corresponds to a relatively small overestimation in sample size with a median (IQR) of 69 (53 to 87). Recruitment as planned would increase the power from the planned 90% to 97.5%.

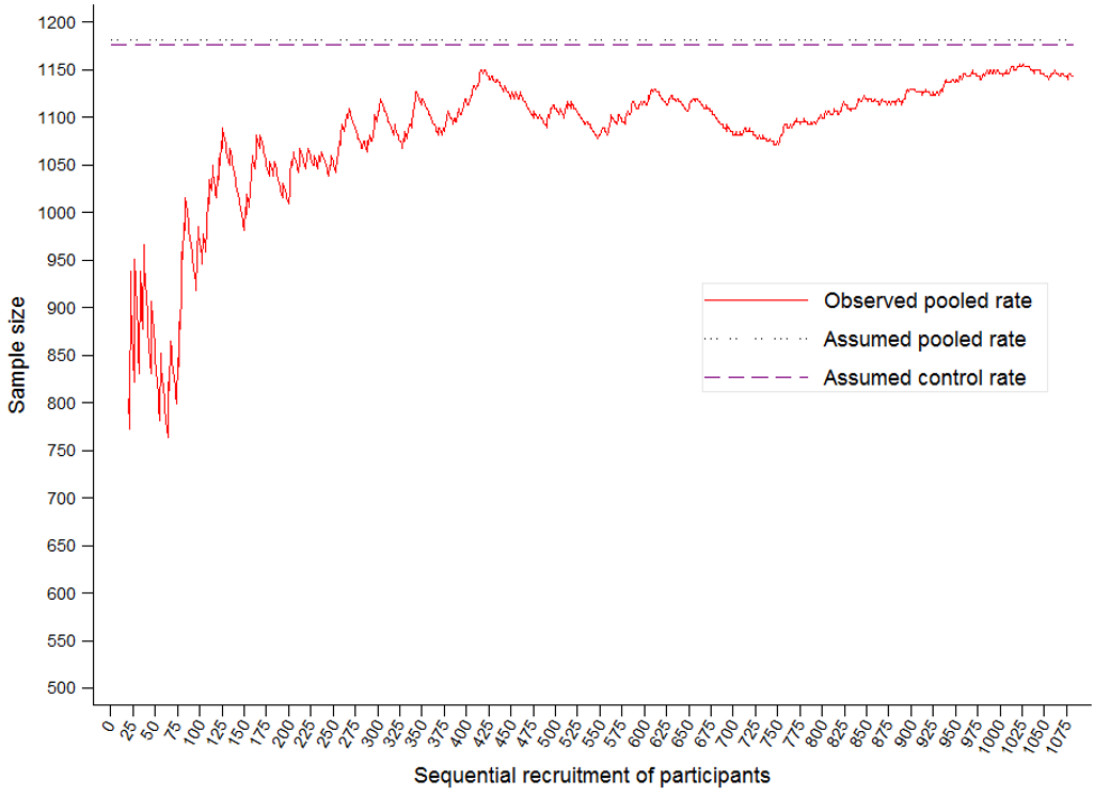

Figure 7.4. Pattern of the re-estimated total sample size for 3Mg trial.

Note: The planned sample sizes calculated using the assumed control or pooled hospital admissions are similar.

Under the planned design, an assumed 10% reduction in hospital admissions (80% to 70%) corresponds to an OR of 0.58 and RR of 0.88 in favour of the IV or NEB $MgSO_4$ interventions. The influence of preserving a fixed effect on an OR scale (0.58) rather than absolute risk reduction was investigated. For a fixed OR of 0.58 and observed average pooled event rate of 77.1%, the corresponding RR would be about 0.86 and a risk difference of 6.6% rather than the assumed 10%. If investigators wished to detect a fixed OR of 0.58, SSR would yield a potential relatively small overestimation in sample size distributed with a median (IQR) of 90 (78 to 105). This assumes that SSR is performed only once after at least 300 participants in total with primary outcome data.

In summary, the underestimation of the pooled hospital admission was relatively small. As a result, the planned sample size potentially slightly overestimated what was required to preserve a 90% power under the design assumptions. However, the sample sizes estimated assuming the planned and observed hospital admissions are consistent and almost similar regardless of the assumed scale of the effect size sought. In such circumstances, the research team may stick to the planned sample size.

7.5.2 Conditional Power Based Stochastic Curtailment Futility Analysis

This section demonstrates the application of one futility analysis at a single interim using the stochastic curtailment approach based on CP as described in Section 2.6 of Chapter 2. Here, 3CPO, RATPAC and Booster trials are considered. 3Mg has been excluded here since the primary outcome had two pairwise comparisons of IV or NEB versus the control. One stochastic curtailment futility analysis in the context of multi-arm trials has not been covered in this thesis.

7.5.2.1 3CPO Trial

Figure 7.5 (a) shows trends in the CP estimated based on the interim intervention effect (CPAP or NIPPV against SOT) and four assumptions made about the intervention effect of participants to be recruited or whose outcomes are yet to be observed: no effect under H_0, effect consistent with interim results, 50% of the effect under H_1 ($H_{0.5}$=3%), and effect assumed under H_1 (6%). Figure 7.5 (b) displays the trend of the Brownian motion defined by equation (2:9), which is a function of the standardised Z statistic weighted according to the interim information fraction. The function indicates the direction and strength of the results as the trial progresses. When H_1 is true (6% mortality difference), the function is expected to follow a linear trend $2.802t$. At the planned end ($t = 1$), the expected standardised Z statistic for a trial designed with 80% power and 5% two-sided type I error is 2.802 ($Z_\beta + Z_{\alpha/2} = 0.842 + 1.95$). Clinically relevant and statistically significant results are observed when the interim Brownian motion trend is above the expected linear trend $2.802t$ (solid red line).

In Figure 7.5 (c), the pattern of the observed interim standardised intervention effect expressed in terms of survival favouring the CPAP or NIPPV interventions is displayed as the trial progresses. As evident in Figure 7.5 (b), the observed Brownian motion trend drifts away drastically from the expected trend after the enrolment of the first 500 participants. This indicates that the observed results are getting worse, even though the survival of participants is slightly in favour of the investigative interventions (CPAP or NIPPV). Hence , the corresponding CP of the study drops sharply under various assumptions as shown in Figure 7.5 (a).

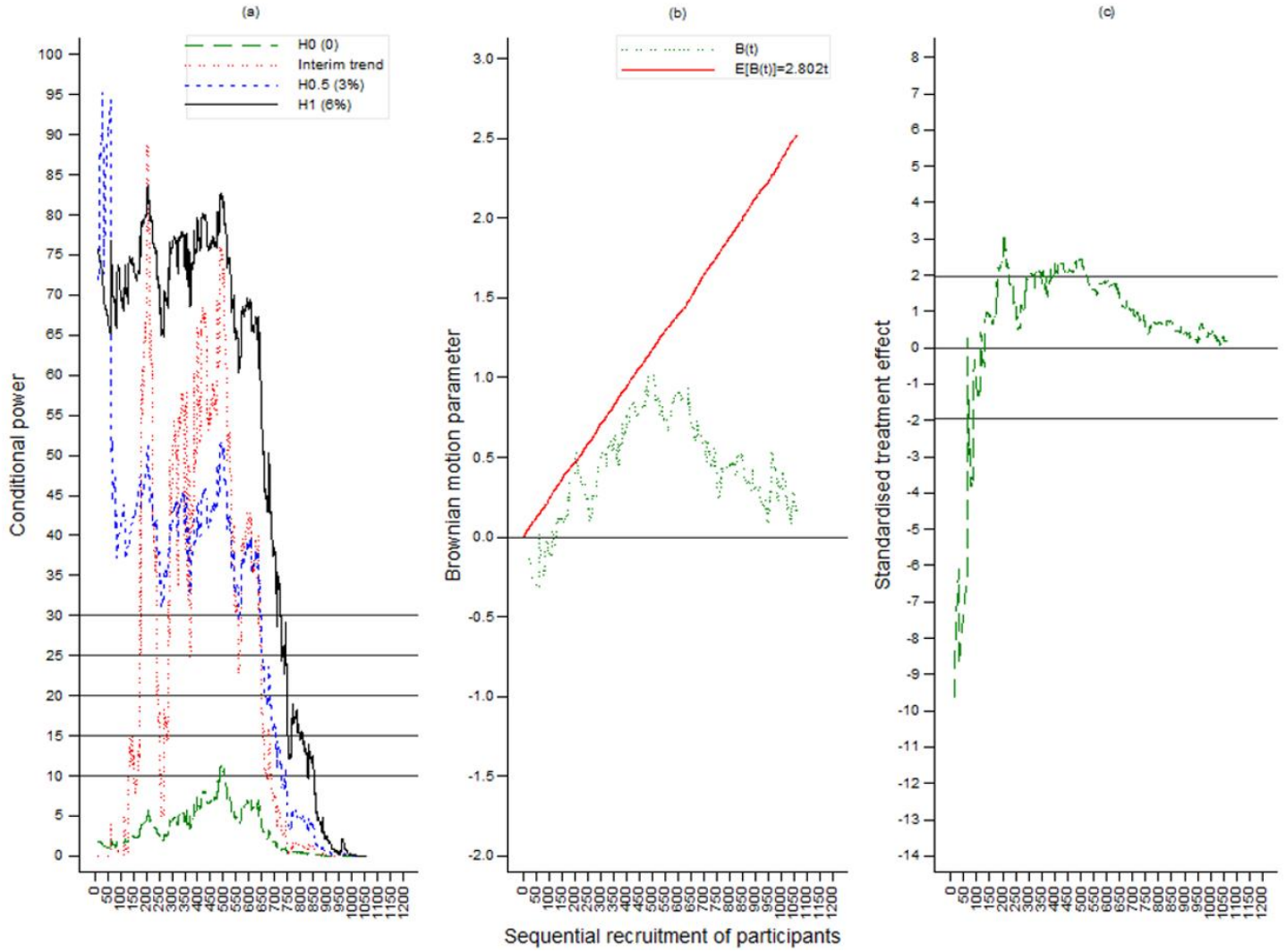

Figure 7.5. Trends of conditional power and intervention effect for 3CPO trial.

Table 7.2 summarises the potential savings in participants by performing one CP based futility analysis shown in Figure 7.5 (a). Here, the savings are presented for 5 CP futility stopping thresholds ranging from 10% to 30%, consistent with the literature review findings (Section 2.6.6 of Chapter 2). In addition, only 3 conservative assumptions about future unobserved results are considered excluding H_0. This reflects a realistic decision-making approach to be used in practice.

For interpretation (row 1), if the trial was designed to be terminated early for futility when CP$\leq$ 10%, it could have been stopped at 69% (826) of the planned total sample size assuming interim futility analysis was conducted at this point. This is because the estimated CP given the interim results and assuming an unobserved mortality difference for the remaining data (6%) is only 9.8%. The trial could have potentially saved recruitment of 374 participants relative to planned total recruitment of 1200. This corresponds to 243 participants relative to the achieved recruitment of 1069 participants. The expected participant savings appeared sizeable under various assumptions considered.

In summary, the trial could have been stopped early for futility assuming the conduct of interim analysis between 53% and 69% of the planned recruitment with a CP futility threshold within the range 10% to 30%.

Table 7.2. Participants savings under various scenarios of conditional power futility threshold stopping criteria.

Assumptions about future unobserved data	CP futility threshold	Information fraction at interim stopping	Estimated interim CP	Interim stopping sample size	Expected total sample size saving relative to:	
					Planned 1200 participants	Achieved recruitment (N=1069)
Under H_1(6%)	10%	69%	9.8%	826	374	243
	15%	63%	13.8%	752	448	317
	20%	63%	19.9%	748	452	321
	25%	61%	24.3%	727	473	342
	30%	59%	29.2%	708	492	361
50% of H_1 ($H_{0.5} = 3\%$)	10%	61%	8.5%	727	473	342
	15%	59%	13.0%	706	494	363
	20%	55%	19.9%	662	538	407
	25%	54%	25.0%	651	549	418
	30%	53%	30.0%	646	554	423
Interim intervention effect	10%	55%	9.5%	668	532	401
	15%	54%	14.9%	655	545	414
	20%	53%	19.4%	649	551	420
	25%	53%	24.8%	648	552	421
	30%	53%	26.8%	645	555	424

CP: Conditional Power.

For illustrative purposes, Figure 7.6 shows the overall type I error trend for performing one CP based futility analysis as described in Section 2.6.9 of Chapter 2, assuming the trial proceeded to the planned end to find a statistically significant result. As observed, the type I error is always less than the pre-planned nominal 5% level when futility is conducted once and positively correlated with regions where the probability of stopping is low. There is no type I error committed if the trial is stopped early for futility.

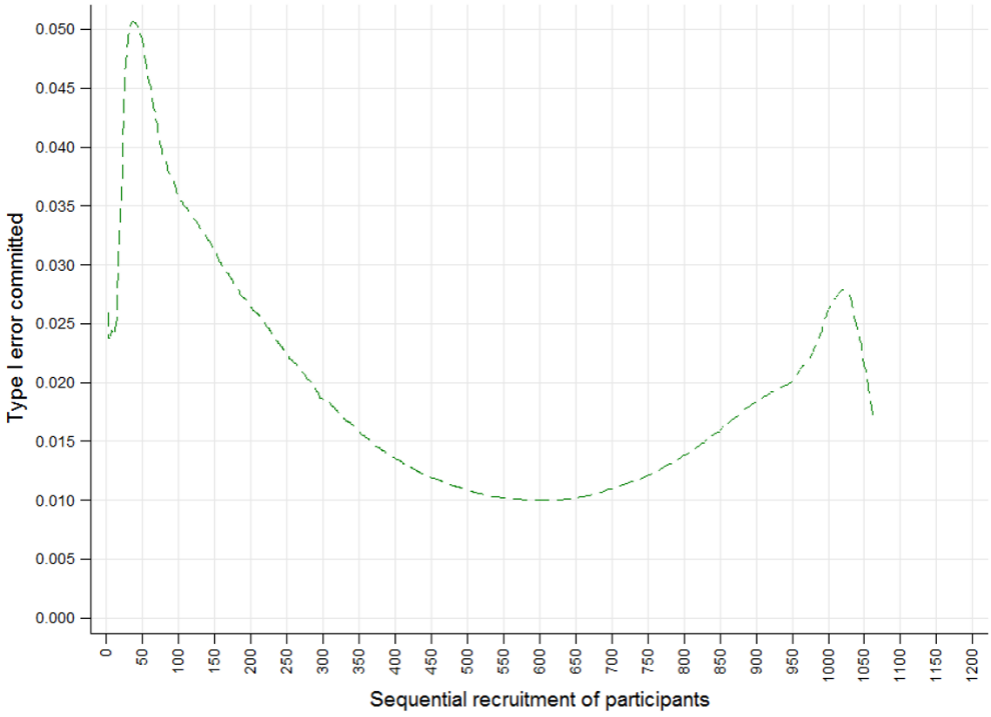

Figure 7.6. Approximate type I error committed for conducting one futility analysis.

In summary and hindsight, there were potentially lost opportunities as a result of not using one stochastic curtailment futility analysis using CP assuming analysis was performed after 50% of the target recruitment. The trial would have potentially saved recruitment of a significant number of patients. The results were highly unstable during the first 300 participants, which is consistent with SSR findings. It is important to note that the research team assumed the effect of CPAP and NIPPV interventions to be similar, hence the rationale to combine the two interventions. Luckily, this assumption was accurate for the 3CPO study. Otherwise, suboptimal decisions could be made when grouped investigative interventions have conflicting effectiveness.

7.5.2.2 RATPAC Trial

This example illustrates a contrasting perspective to the 3CPO above. The interpretation of Figure 7.7 is similar to Figure 7.5 described in Section 7.5.2.1. Here, the observed trend in successful hospital discharge difference is overwhelmingly above the expected trend assuming 5% difference under H_1. The overwhelming trend of benefit is consistent regardless of the assumptions made about the intervention effect for the future unobserved data. The estimated CP under various scenarios reaches 100% before the recruitment of 1200; at approximately 37% of the targeted recruitment as shown in Figure 7.7 (a). As a result, there was no opportunity to stop the trial for futility. The overwhelming trend of the intervention effect strongly suggests opportunities for efficacy early stopping. However, since early stopping for efficacy based on stochastic curtailment is discouraged as highlighted in Section 2.6.2 of Chapter 2, the use of standard and information based GSDs allowing for early stopping for efficacy is demonstrated in Sections 7.5.3.1 and 7.5.4, respectively.

In summary, RATPAC trial demonstrated that the use of CP can also indirectly reveal the effectiveness of an investigative intervention which may introduce operational bias. For instance, a CP of 99% for a trial designed with 80% to 90% power indicates that the trial has already addressed its objectives and the investigative intervention is overwhelmingly beneficial. It highlights the importance of adequate processes and procedures to guide the conduct of futility analysis and communication of results to key stakeholders such as Funders and the research teams.

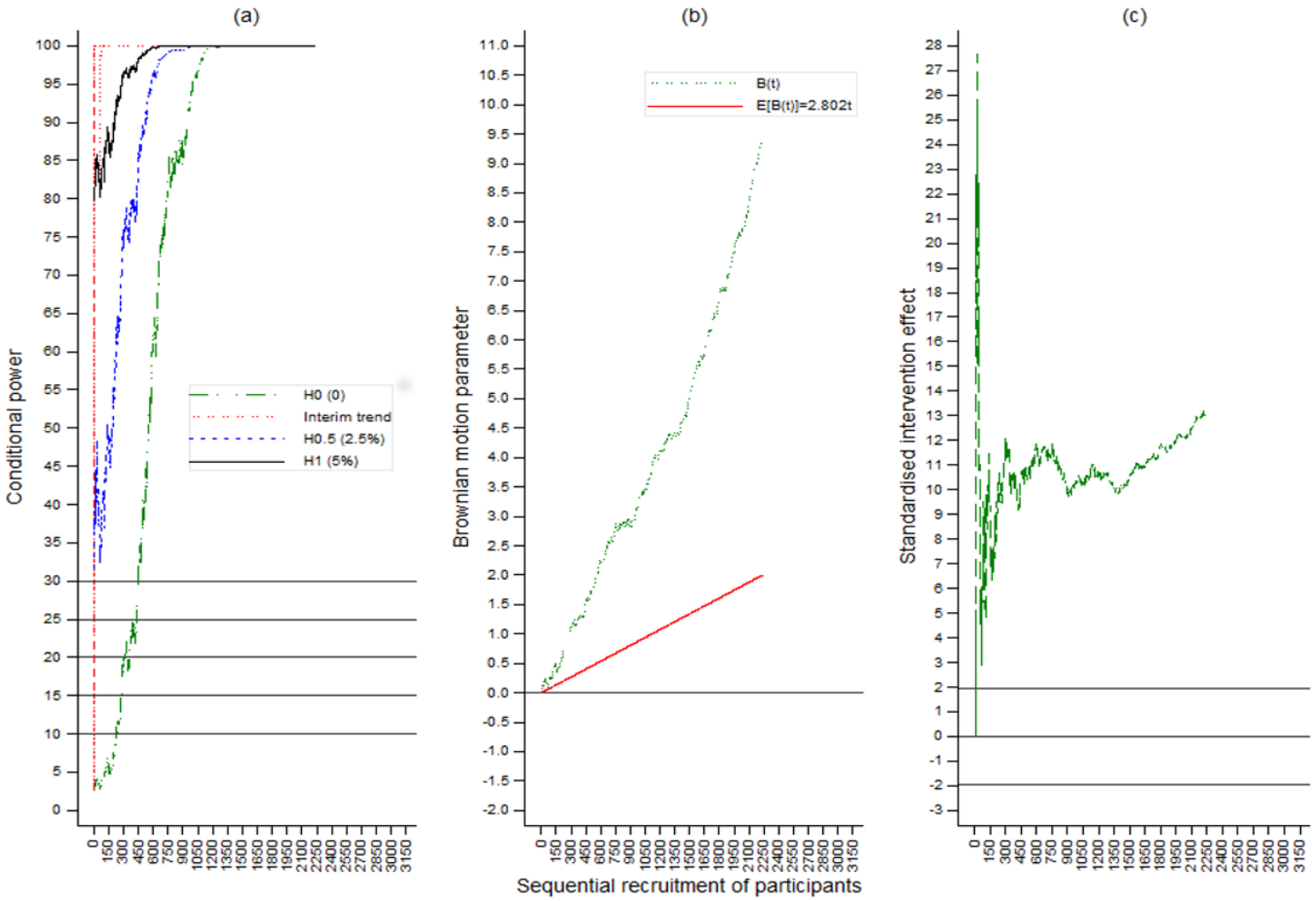

Figure 7.7. Trends of conditional power and intervention effect for RATPAC trial.

7.5.2.3 Booster Trial

As highlighted in Section 7.3.4, the Booster trial was designed to recruit a total 600 participants (200 per arm). However, the research team managed to recruit 282 (47%) of the targeted recruitment. Furthermore, the trial retained 160 (57%) participants with valid primary outcome data meeting the intention-to-treat criteria; combined booster interventions (n=99) and control (n=61). Even though valid data constituted only 27% of the planned, this section illustrates how stochastic curtailment futility analysis can enhance public funders' decision-making regarding whether additional funding requests are necessary.

Here, Figure 7.8 (a) displays the pattern of CP and intervention effect for the first 160 participants to the point when investigators requested further research funding from public funders. The intervention effect waned drastically after 60 participants. Figure 7.8 (b) and Figure 7.8 (c) show that the intervention effect is even in favour of the control arm. After 80 participants, the CP assuming the future intervention effect is as observed at interim or 50% of the assumed H_1effect is less than 25%. Furthermore, the CP assuming the observed interim effect is close to 0% after 100 participants. However, if the intervention effect under $H_1$is assumed for the remaining data, the CP is slightly above 45%. This illustrates the difficulties in stopping early under $H_1$when the CP is performed early on in the trial, even when the interim intervention effect is very small as highlighted in Section 2.6.7 of Chapter 2.

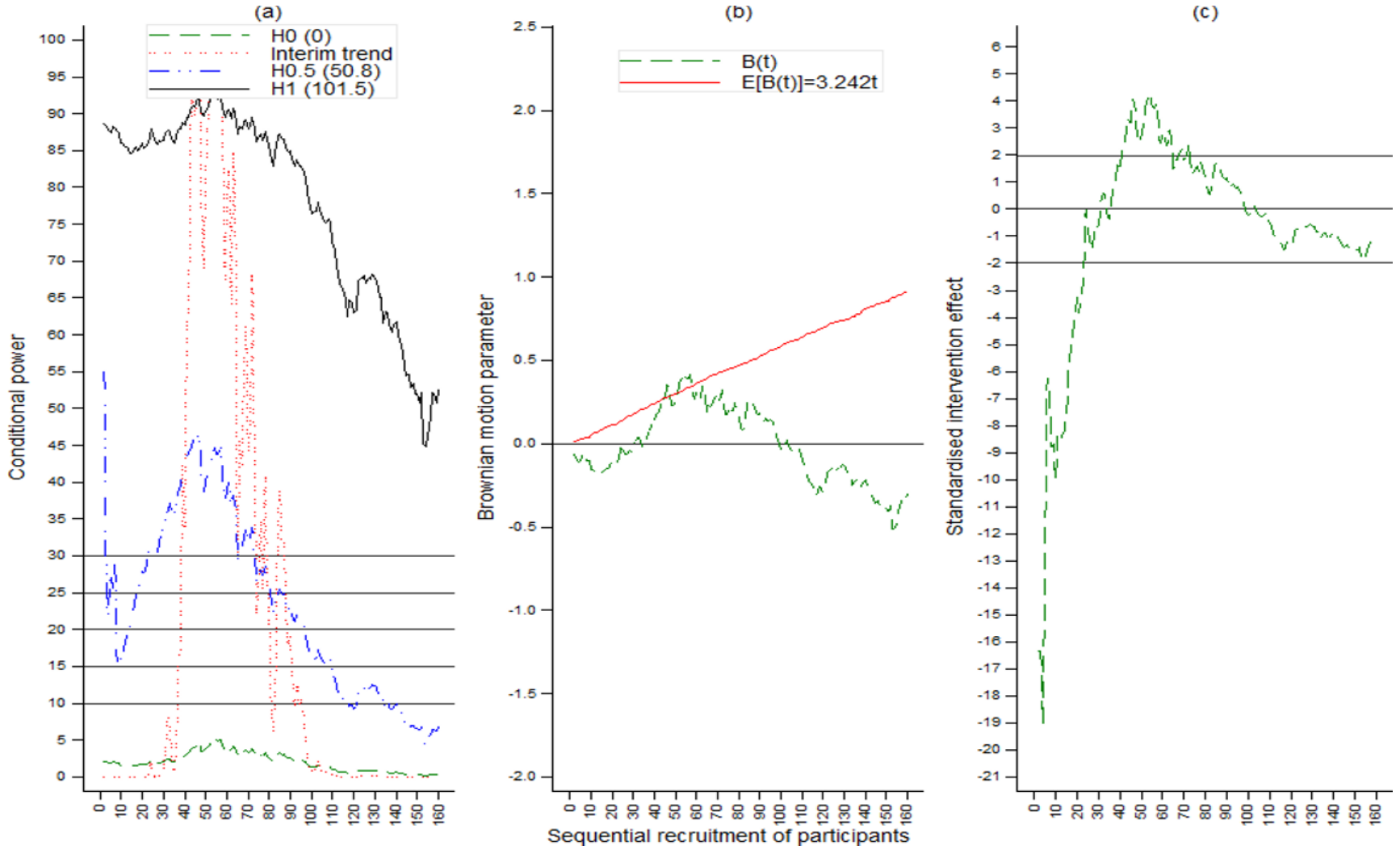

Figure 7.8. Trends of conditional power and intervention effect for Booster trial.

7.5.3 Group Sequential Design

This section demonstrates the design, implementation, and interim decision-making for a GSD to allow for early stopping as described in Section 2.7 of Chapter 2. The RATPAC and 3CPO case studies are utilised as they give contrasting perspectives as highlighted in Section 7.4.2. In addition to the wishes of the investigators, results from Chapter 2 and lessons learned in Chapter 6 guided the choice and timing of interim analyses.

7.5.3.1 RATPAC Trial

Sutton et al (2012) previously reanalysed RATPAC based on a sequential approach using Whitehead's triangular approach described in Appendix 2.2. In addition, Sutton and colleagues computed the stopping boundaries conditional on the planned fixed sample size. Here, a different approach is illustrated using a GSD with delayed analyses.

7.5.3.1.1 The Design

A two-sided GSD allowing for either futility or efficacy early stopping is considered although in practice the choice is trial depended. Stopping boundaries and the timing of interim analyses were chosen balancing the benefits of stopping early and concerns raised in Chapters 3 and 4. These concerns include those relating to the robustness of ADs in decision-making and credibility of results to change practice, when trials are stopped early. As illustrated in Sections 7.5.1 and 7.5.2, conducting interim analyses too early may result in unreliable decisions because of unstable estimates of outcome variability. As reflected in Chapter 2, it is important to avoid interim analyses in regions where learning effects are most likely. In addition, relatively fair representation of participants across centres and reliability of estimates of the intervention effects are important considerations in delaying the interim analyses. As a result, two interim analyses were planned at 50% and 70% of the target recruitment, in addition to the final analysis at 100%. Efficacy stopping boundaries were constructed using LD spending functions that mimic OBF boundaries. The inner wedge 'non-binding' futility boundaries were constructed using LD spending functions extended by Pampallona and Tsiatis to allow for futility early stopping.

The investigators wanted an 80% powered trial to detect a 5% increase in successful hospital discharge from the 50% assumed in the SC arm (OR=1.22). Assuming the three interim analyses at 50%, 70% and 100%, and a two-sided efficacy and 'non-binding' futility boundaries at 5% two-sided type I error, the trial would need a maximum total sample size of 3348 (1674 per arm); about 218 more participants compared to the fixed sample size design. If the intervention effect assumed under H_0 or H_1 is true, the trial would be stopped with an expected

total sample size of 2398 and 2608, respectively. The 1st and 2nd interim analyses would be performed when 1674 and 2344 participants were enrolled, respectively. Figure 7.9 displays the stopping boundaries of the design.

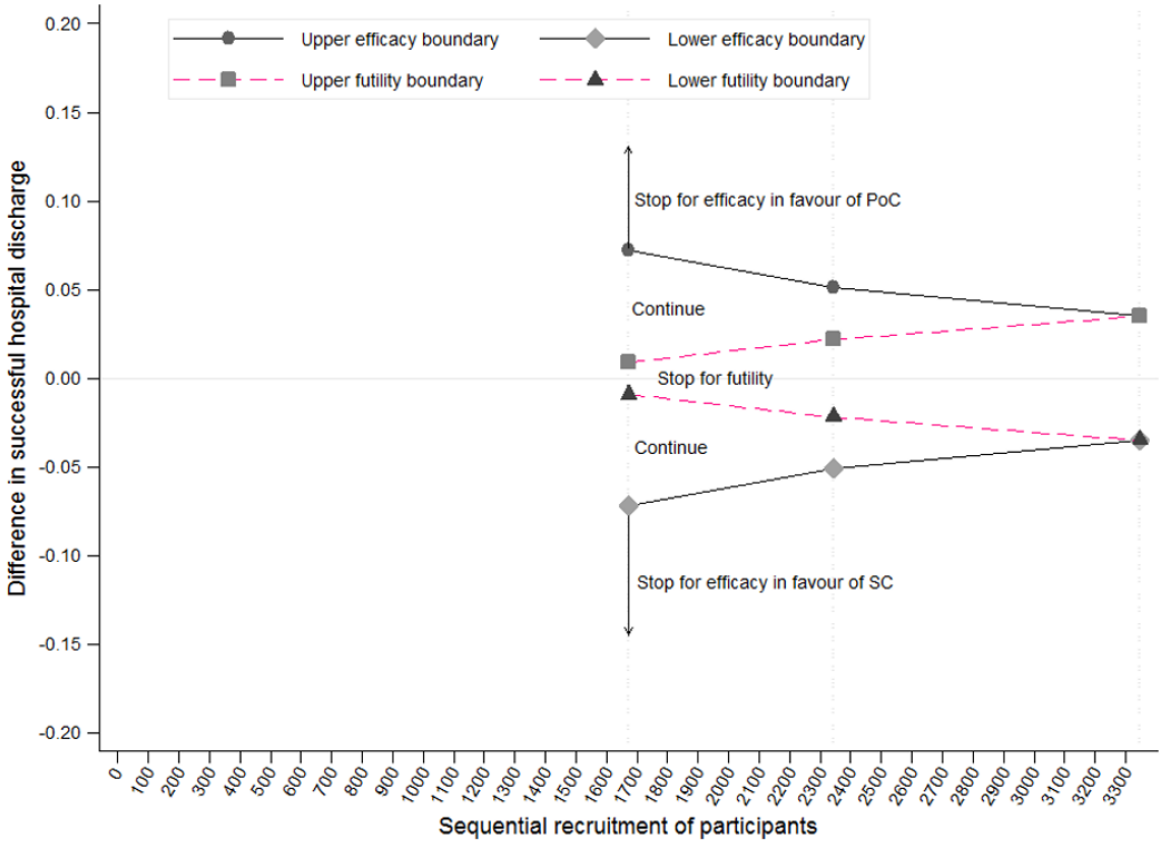

Figure 7.9. Stopping boundaries for RATPAC two-sided group sequential test.

It is important to understand the performance of the planned design under various assumptions regarding hospital discharge estimate for the SC arm and the intervention effect. For example, the research team may be interested to understand what would happen when close to 25% hospital discharge is observed rather than the assumed 50%. Simulation work to answer such questions would help the research team to plan in advance against the unexpected. Table 7.3 summarises statistical properties of the design based on 100,000 simulated trials under selected scenarios, conditional on the maximum sample size of 3348 using ADDPLAN software. Here, the SC hospital discharge (50%, 40%, 30%, 25%, and 20%) and absolute increase in successful hospital discharge (0%, 5%, and 10%) are considered. In practice, simulation scenarios could be extended to include decrease in successful discharge rates in favour of the SC arm. These were excluded here to manage the content of this thesis in view of

the preliminary results presented in Section 7.5.2.2, although such results should be provided in practice. The chances of stopping at the 1st and 2nd interim analyses, and final analysis for any reasons are presented together with corresponding chances for efficacy or futility. For example, under the design assumptions (row 2), the proportion of simulated trials stopping at the 1st and 2nd interim analyses, and final analysis are 21.8%, 37.0%, and 41.2%. Simulated trials that would be stopped at the 1st interim analysis for efficacy in favour of the PoC or futility made up 82.3% or 17.7% of the undertaken trials, respectively.

As highlighted in Section 7.5.1.1, the assumed successful hospital discharge in the SC arm was significantly overestimated. In Table 7.3 (rows 5 and 9), a 25% successful hospital discharge in the SC arm instead of 50% is assumed. Furthermore, on row 9, an overwhelming increase in successful hospital discharge of 10% rather than 5% is assumed. Under these assumptions, 93.4% of simulated trials would be stopped for efficacy at the 1st interim analysis. As the assumed successful hospital discharge in the SC arm is decreased from 50% and intervention effect sought increased from 5% to 10%, the power of the study increases drastically to 100%.

Table 7.3. Statistical properties of a retrospective group sequential design for RATPAC trial.

Simulation parameters	Increase in successful hospital discharge	Interim information fraction	Probability of stopping for futility	Proportion stopping for efficacy in favour of:		Probability of stopping for any reasons	Statistical Power
				PoC (upper boundary)	SC (lower boundary)		
$n_c = 50\%$	0%	0.50 (1674)	98.9%	0.6%	0.5%	28.5%	~4.7%
$n_t = 50\%$		0.70 (2344)	97.4%	1.3%	1.3%	45.3%	
		1.00 (3348)	87.8%	6.0%	6.2%	26.2%	
$n_c = 50\%$		0.50 (1674)	17.7%	82.3%	-	21.8%	~80.2%
$n_t = 55\%$	5%	0.70 (2344)	16.1%	83.9%	-	37.0%	
		1.00 (3348)	24.2%	75.8%	-	41.2%	
$n_c = 40\%$		0.50 (1674)	17.0%	83.0%	-	22.8%	~80.5%
$n_t = 45\%$	5%	0.70 (2344)	15.4%	84.6%	-	37.1%	
		1.00 (3348)	24.7%	75.3%	-	40.2%	
$n_c = 30\%$		0.50 (1674)	12.9%	87.1%	-	25.1%	~84.6%
$n_t = 35\%$	5%	0.70 (2344)	11.9%	88.1%	-	37.8%	
		1.00 (3348)	20.8%	79.2%	-	37.1%	
$n_c = 25\%$		0.50 (1674)	8.7%	91.3%	-	27.6%	
$n_t = 30\%$	5%	0.70 (2344)	9.2%	90.8%	-	38.6%	~87.8%
		1.00 (3348)	18.5%	81.5%	-	33.7%	
$n_c = 50\%$		0.50 (1674)	-	100.0%	-	87.3%	
$n_t = 60\%$	10%	0.70 (2344)	-	100.0%	-	12.0%	~100.0%
		1.00 (3348)	-	100.0%	-	0.8%	
$n_c = 40\%$		0.50 (1674)	-	100.0%	-	87.3%	
$n_t = 50\%$	10%	0.70 (2344)	-	100.0%	-	11.9%	~100.0%
		1.00 (3348)	0.8%	99.2%	-	0.8%	
$n_c = 30\%$		0.50 (1674)	-	100.0%	-	90.7%	
$n_t = 40\%$	10%	0.70 (2344)	-	100.0%	-	8.7%	~100.0%
		1.00 (3348)	0.5%	99.5%	-	0.4%	
$n_c = 25\%$		0.50 (1674)	-	100.0%	-	93.4%	
$n_t = 35\%$	10%	0.70 (2344)	-	100.0%	-	6.3%	100.0%
		1.00 (3348)	-	100.0%	-	0.2%	
$n_c = 20\%$		0.50 (1674)	-	100.0%	-	96.2%	
$n_t = 30\%$	10%	0.70 (2344)	-	100.0%	-	3.7%	100.0%
		1.00 (3348)	-	100.0%	-	0.1%	

PoC: Point-of-care; SC: Standard Care; "-" represents 0%; 100,000 simulations performed.

Table 7.4 summarises the stopping boundary values for early stopping decision-making criteria expressed in terms of difference in proportions, p-value, and Z statistic scales. For example, the trial would be stopped for efficacy at the 1st interim analysis in favour of PoC or SC when the difference in hospital discharge is above 7.2% or below -7.2%, respectively. That is, when p-value is ≤ 0.002. Furthermore, futility would be declared when the 1st interim difference in hospital discharge fell between -0.9% and 0.9%.

Table 7.4. Stopping boundary values for a retrospective group sequential design for RATPAC trial.

Boundary Scale	Information fraction (n)	Cumulative $\boldsymbol{\alpha}$ spent	Cumulative $\boldsymbol{\beta}$ spent	Stopping boundary values			
				Efficacy		Futility	
				Lower	Upper	Lower	Upper
Difference in proportions	0.50 (1674)	0.003	0.040	-0.072	0.072	-0.009	0.009
	0.70 (2344)	0.015	0.099	-0.051	0.051	-0.022	0.022
	1.00 (3348)	0.050	0.200	-0.035	0.035	-0.035	0.035
P-value	0.50 (1674)	0.003	0.040	0.002	0.002	0.703	
	0.70 (2344)	0.015	0.099	0.007	0.007	0.287	
	1.00 (3348)	0.050	0.200	0.023	0.023	0.045	
Z statistic	0.50 (1674)	0.003	0.040	-2.963	2.963	-0.381	0.381
	0.70 (2344)	0.015	0.099	-2.462	2.462	-1.065	1.065
	1.00 (3348)	0.050	0.200	-2.002	2.002	-2.002	2.002

7.5.3.1.2 Interim Monitoring and Decision-Making Process

At the 1st interim analysis, 1674 participants with primary outcome data would have been recruited; PoC (n=842) and SC (n=832), assuming there were no participants with delayed responses. Successful hospital discharge was 258(30.6%) and 119(14.3%) in the PoC and SC arms, respectively. That is, an increase in successful hospital discharge rate (95% CI) of 16.3% (12.4% to 20.2%). The stagewise adjusted and naïve results are the same. This is equivalent to an OR (95% CI) of 2.65(2.08 to 3.38); p-value <0.0001. For consistency with the presentation of the original study, interim monitoring was performed using a logistic regression model accounting for centre effect. This produced an OR (95% CI) of 3.00(2.32 to 3.90) in favour of the PoC arm in increasing successful hospital discharge. Figure 7.10 displays the intervention effect trend with 95% CI up to the time of the 1st interim analysis and the decision-making criteria based on the difference in proportion stopping boundary scale summarised in Table 7.4. Since the intervention effect (16.3%) is overwhelmingly in favour of PoC arm (above the upper boundary of 7.2%), the trial could have been stopped early for efficacy after the enrolment of 1674 participants. It is important to note that the trend of the intervention effect should be provided only at the point of early stopping.

Figure 7.10. Interim monitoring for a retrospective group sequential design for RATPAC trial.

Table 7.5 summarises the approximate benefits of early stopping in terms of savings in recruitment duration and participant numbers under three design scenarios. For instance, stopping early at the 1st interim analysis would have averted further recruitment of 1456(46.5%) participants and resulted in an approximate reduction in recruitment duration of 10.3 months compared to the planned fixed sample size design.

Table 7.5. Benefits of efficacy early stopping for a retrospective group sequential design for RATPAC trial.

Scenario	Sample size	Proportion of sample size used	Participants savings	Reduction in recruitment duration (months)
Planned fixed sample size	3130	53.5%	1456	10.3
Planned GSD	3348	50.0%	1674	11.8
Achieved recruitment	2263	74.0%	589	4.2

GSD: Group sequential design

7.5.3.2 3CPO Trial

7.5.3.2.1 The Design

In this case study, it is demonstrated that a trial can be designed to stop early for futility only at the earliest interim and either for futility or efficacy at subsequent interims. These properties may reflect the wishes

of some conservative investigators favouring a delay in early stopping for efficacy. Hence, a two-sided GSD with two interim analyses at 50% and 65% of the planned enrolment is considered. The rationale for this timing is similar to that provided for RATPAC in Section 7.5.3.1. However, the spacing of the interims is at the discretion of the research team taking into account aspects such as logistics and added benefits.

The investigators wished to detect a 5% reduction in mortality for superiority in favour of the CPAP or NIPPV arm from the 15% mortality assumed in the SOT arm. To preserve at least 80% power and a 5% two-sided type I error, the GSD considered here would require a total of 1168 participants with an allocation ratio of 2 to 1: CPAP (n≈389), NIPPV (n≈389), and SOT (n≈389). Table 7.6 summarises the properties of the design based on 100,000 simulations for a fixed sample size of 1168 under various scenarios about the assumed SOT mortality (15%, 12%, 10%, and 8%) and observed mortality reduction (0%, 3%, and 6%). The exclusion of the efficacy stopping option at the 1st interim slightly increases the power to 82.1% and reduces the type I error to 4.6%. The interpretation of Table 7.6 is similar to Table 7.3.

Table 7.6. Design properties of a retrospective group sequential design for 3CPO trial.

Simulation parameters	Mortality reduction	Interim information fraction	Probability of stopping for futility	Proportion stopping for efficacy in favour of:		Probability of stopping for any reasons	Statistical Power
				SOT (upper boundary)	CPAP or NIPPV (lower boundary)		
$n_c = 15\%$	0%	0.50 (584)	100.0%	NA	NA	28.9%	4.6%
$n_t = 15\%$		0.65 (759)	96.9%	2.0%	1.1%	36.0%	
		1.00 (1168)	90.2%	5.2%	4.7%	35.1%	
$n_c = 15\%$		0.50 (584)	100.0%	NA	NA	18.0%	26.2%
$n_t = 12\%$	3%	0.65 (759)	75.4%	0.1%	24.6%	27.8%	
		1.00 (1168)	64.2%	0.1%	35.7%	54.2%	
$n_c = 15\%$		0.50 (584)	100.0%	NA	NA	3.3%	82.1%
$n_t = 9\%$	6%	0.65 (759)	8.0%	-	92.0%	44.5%	
		1.00 (1168)	21.1%	-	78.9%	52.2%	
$n_c = 12\%$		0.50 (584)	100.0%	NA	NA	15.7%	31.6%
$n_t = 9\%$	3%	0.65 (759)	69.0%	0.1%	31.0%	26.9%	
		1.00 (1168)	59.5%	~0.0%	40.5%	57.4%	
$n_c = 12\%$		0.50 (584)	100.0%	NA	NA	1.7%	
$n_t = 6\%$	6%	0.65 (759)	3.2%	-	96.8%	55.6%	90.6%
		1.00 (1168)	13.9%	-	86.1%	42.7%	
$n_c = 10\%$		0.50 (584)	100.0%	NA	NA	14.0%	
$n_t = 7\%$	3%	0.65 (759)	61.3%	~0.0%	38.7%	26.3%	37.1%
		1.00 (1168)	54.9%	~0.0%	45.1%	59.7%	
$n_c = 10\%$		0.50 (584)	100.0%	NA	NA	0.7%	
$n_t = 4\%$	6%	0.65 (759)	1.2%	-	98.8%	67.1%	95.9%
		1.00 (1168)	8.0%	-	92.0%	32.1%	
$n_c = 8\%$		0.50 (584)	100.0%	NA	NA	11.1%	
$n_t = 5\%$	3%	0.65 (759)	49.9%	~0.0%	50.1%	26.2%	45.7%
		1.00 (1168)	48.1%	~0.0%	51.9%	62.7%	
$n_c = 8\%$		0.50 (584)	100.0%	NA	NA	0.2%	
$n_t = 2\%$	6%	0.65 (759)	0.2%	-	99.8%	83.8%	99.3%
		1.00 (1168)	2.6%	-	97.4%	16.1%	

NA: Not applicable; SOT: standard oxygen therapy; CPAP: continuous positive airway pressure; NIPPV: non-invasive intermittent positive-pressure ventilation; "-" represents 0.0%; 100,000 trial simulations performed.

Underestimation of the SOT mortality of 15% causes an increase in statistical power assuming the 5% mortality difference is still clinically relevant to detect. Simulated trial results presented in Table 7.6 can be extended to include the possibility for overestimation in SOT mortality, such as 18%, 20%, and 22%. However, these results are excluded here in view of the plausible estimates presented in Appendix 7.1.

Table 7.7 summarises the stopping boundary values for the design presented on three boundary scales. For instance, at the 1st interim analysis based on the outcomes of the 584 participants, the trial would be terminated for futility when the difference in mortality lies between -1.1% and 1.1%. On the other hand, the trial would be stopped for superiority of CPAP or NIPPV when the difference in mortality is -6.6% at the 2nd interim analysis based on outcomes of 759 participants.

Table 7.7. Stopping boundary values of a retrospective group sequential design for 3CPO trial.

Boundary Scale	Information fraction (n)	Cumulative α spent	Cumulative β spent	Stopping boundary values			
				Efficacy		Futility	
				Lower	Upper	Lower	Upper
Mortality difference	0.50 (584)	0.000	0.040	NA	NA	-1.1%	1.1%
	0.65 (759)	0.011	0.083	-6.6%	6.6%	-2.2%	2.2%
	1.00 (1168)	0.050	0.200	-4.1%	4.1%	-4.1%	4.1%
P-value	0.50 (584)	0.000	0.040	NA	NA	0.708	
	0.65 (759)	0.011	0.083	0.005	0.005	0.394	
	1.00 (1168)	0.050	0.200	0.023	0.023	0.047	
Z statistic	0.50 (584)	0.000	0.040	NA	NA	-0.375	0.375
	0.65 (759)	0.011	0.083	-2.546	2.546	-0.853	0.853
	1.00 (1168)	0.050	0.200	-1.990	1.990	-1.990	1.909

NA: Not Applicable. Note: The lower and upper boundary p-values for declaring futility are the same because the futility region is the same as shown in Figure 7.11.

Graphical representation of stopping regions superimposed on the interim results is presented in Figure 7.11.

7.5.3.2.2 Monitoring and Decision-Making Process

At the 1st interim analysis, the observed mortality in the SOT and CPAP or NIPPV arms was 20/194(10.3%) and 28/390(7.2%) respectively; a mortality difference of -3.1% in favour of the CPAP or NIPPV arms. The 95% naïve CI and RCI were (-8.1% to 1.9%) and (-10.7% to 4.4%), respectively. In addition, the trial had a CP of 35.2% assuming the future unobserved trend is the same as the interim results. The trial would proceed to the 2nd interim analysis since the observed mortality difference falls inside the futility limits of -1.1% and 1.1%.

After observing outcomes of 759 participants at the 2nd interim analysis, the observed mortality in the SOT and CPAP or NIPPV arms were 24/255(9.4%) and 44/504(8.7%), respectively. That is, a mortality difference of just -0.7% and naïve 95% CI of (-5% to 3.7%; p-value=0.760). Median unbiased results could not be estimated

due to computational limitations of East 6.3 and ADDPLAN 6.1 when futility boundaries are crossed. Figure 7.11 displays the trend of results up to the 2nd interim analysis and the design's stopping boundary regions.

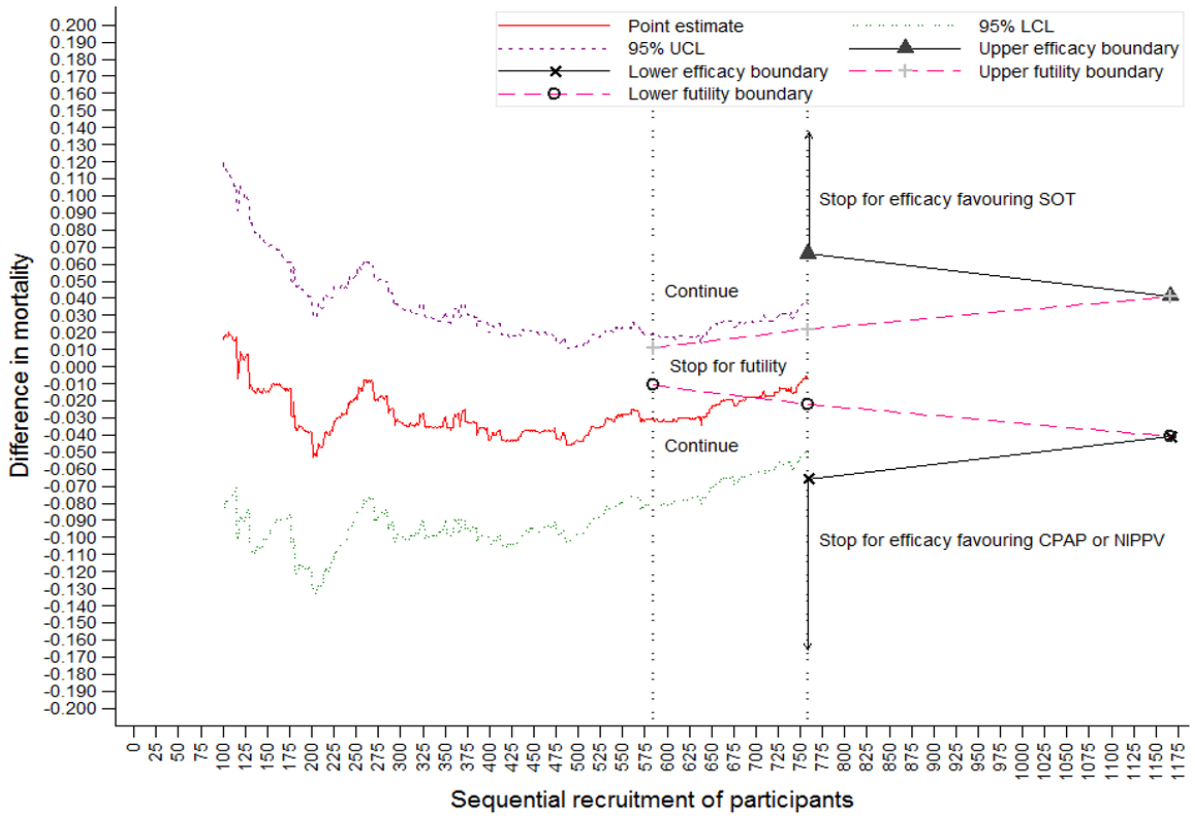

Figure 7.11. Interim monitoring for a retrospective group sequential design for 3CPO trial.

As evident in Figure 7.11, a -0.7% observed mortality difference falls within the futility limits of -2.2% and 2.2% at the 2nd interim analysis. Furthermore, the results are getting worse and trending towards H_0 effect. As a result, a recommendation to terminate the trial for futility could be made. However, since the futility boundaries are 'non-binding', investigators may choose to ignore such a recommendation without undermining the type I error and power, when there are justifiable reasons to do so.

The interpretation of results may be challenging because the naïve 95% CI is wide with both limits falling within the continuation zones. To aid decision-making, 100,000 trials were simulated to estimate the chances of 'accepting' and rejecting H_0 favouring CPAP or NIPPV at the final analysis, assuming the trial is continued to the planned end. Simulated trial results are summarised in Table 7.8. For instance; given the 2nd interim results, planned stopping boundaries, intended maximum sample size of 1168, and assuming the results after the 2nd

interim analysis are close to those observed (~1% difference in favour of CPAP or NIPPV), 99.7% of the simulated trials would fail to reject H_0. Even assuming an unlikely 6% mortality difference expected under H_1 for the remaining 408 participants, only 15.4% of simulated trials would yield results in favour of the CPAP or NIPPV arm. In this regard, the chances of trend reversal are very unlikely; a convincing rationale to recommend early stopping for futility.

Table 7.8. Conditional simulation results for 3CPO trial after the 2nd interim analysis.

Conditional simulation parameters	Assumed mortality reduction after 2nd interim analysis	Probability of stopping for futility	Probability of stopping for efficacy in favour of CPAP or NIPPV
$n_c = 10\%$ $n_t = 9\%$	1%	99.7%	0.3%
$n_c = 10\%$ $n_t = 8\%$	2%	99.1%	0.9%
$n_c = 10\%$ $n_t = 7\%$	3%	98.0%	2.0%
$n_c = 10\%$ $n_t = 6\%$	4%	95.7%	4.3%
$n_c = 10\%$ $n_t = 5\%$	5%	91.5%	8.5%
$n_c = 10\%$ $n_t = 4\%$	6%	84.5%	15.4%

CPAP: continuous positive airway pressure; NIPPV: non-invasive intermittent positive-pressure ventilation; 100,000 trial simulations performed.

The 3CPO trial recruited a total of 1069 participants within a period of approximately 46 months; a uniform recruitment rate of about 23.2 participants per month. Table 7.9 summarises potential benefits in terms of reduction in recruitment duration and participants savings under three design scenarios. Here, it is assumed that the trial was stopped early for futility at the 2nd interim analysis; after recruitment of 759 participants. Early stopping would have reduced recruitment duration by approximately 13.4 months and saved 310 participants compared to the achieved recruitment of 1069.

Table 7.9. Benefits of futility early stopping at the 2nd interim analysis for 3CPO trial.

Scenario	Sample size	Proportion of sample size used	Participants savings	Reduction in recruitment duration (months)
Planned fixed sample size [a]	1036	73.3%	227	9.8
Planned GSD	1168	65.0%	409	17.6
Achieved recruitment	1069	71.0%	310	13.4

[a] estimated based on pooled event rate; GSD: Group Sequential Design.

7.5.3.3 Assessing Robustness of Adaptive Designs Interim Decision-Making

This section briefly describes how the interim results and decision-making process of the presented case studies compare with the final findings. For the final RATPAC results, 358/1125(31.8%) were successfully discharged compared to 146/1118(13.1%) in the SC, an increase in successful hospital discharge (95% CI) of 18.8% (15.4% to 22.1%) compared to 16.3% (12.4% to 20.2%) at early trial stopping (1st interim analysis). The final results showed an adjusted OR (95% CI) of 3.81(3.01 to 4.82; p-value< 0.0001) compared to 3.00(2.32 to 3.90); p-value<0.0001) at the 50% interim analysis.

For the 3CPO trial, the final results showed a 7-day mortality of 66/695(9.5%) in the CPAP or NIPPV and 26/367(9.8%) in the SOT arm; mortality difference (95% CI) of -0.3% (-4.1% to 3.4; p-value=0.870) compared to -0.7% (-5% to 3.7%; p-value=0.760) at the 2nd interim analysis.

In summary, the final published results for the RATPAC and 3CPO trials are consistent with the interim results assuming these trials were stopped early at 50% and 65% of the planned recruitment, respectively. Thus, there was no additional information gained to address the primary trial objectives by recruiting participants beyond the considered interim analyses. However, it should be noted that the magnitude of the interim intervention effects is slightly lower than the observed effects when trials recruit to the planned end.

7.5.4 Information Based Group Sequential Design for RATPAC trial

In this section, the implementation of an information based GSD highlighted in Section 2.8 is illustrated using the RATPAC trial. This case study was selected because the assumed nuisance design parameters were inaccurate. Statistical implementation was performed using East software.

7.5.4.1 The Design

The trial design characteristics described in Section 7.3.1 including stopping boundaries and number and timing of interim analyses described in Section 7.5.3.1.1 were preserved. For a fixed sample size design assuming a 5% clinically relevant difference, 50% successful hospital discharge rate in the SC arm, 80% power and 5% two-sided type I error, the maximum information required is given by:

$$I_1 = \left(\frac{Z_{\alpha/2} + Z_{\beta}}{\theta_{\delta}}\right)^2 = \left(\frac{Z_{0.025} + Z_{0.8}}{0.05}\right)^2 = \left(\frac{-1.959964 - 0.84162123}{0.05}\right)^2 \approx 3139.55$$

Thus, a fixed sample size study without interim analysis would aim to recruit until the information of 3139.55 is reached. This information fraction is approximately equivalent to a total sample size of 3124 (1562 per arm) assuming the 50% SC successful hospital discharge is correct.

Now, for an information based GSD with three interim analyses at 50%, 70%, and 100% of the planned recruitment, LD OBF equivalent efficacy and non-binding futility stopping boundaries, the trial would require a maximum information of approximately 3364.48. Table 7.10 summarises the approximate equivalent total sample size under a number of scenarios about the observed SC hospital discharge rate for a fixed 5% clinically relevant difference. For example, if the observed SC hospital discharge is close to the assumed rate of 50% (between 45% and 50%), a total of 3348 participants (1674 per arm) would be required to reach the maximum information fraction. However, if the SC observed discharge rate is around 15% (close to the actual observed rate), the trial would only require approximately 1935 participants to reach the same desired maximum information fraction. In addition, the trial would be stopped early under H_0 or H_1 with an average total sample size of 1386 or 1507, respectively.

Table 7.10. Conversion of design information for RATPAC information based group sequential designs.

Maximum information	Observed hospital discharge in the SC arm	Translated equivalent total sample size	Average sample size under:	
			H_0	H_1
3364.48	50%	3348	2398	2608
	45%	3348	2398	2608
	40%	3281	2350	2556
	35%	3146	2254	2451
	30%	2944	2109	2294
	25%	2675	1916	2084
	20%	2339	1676	1822
	15%	1935	1386	1507

A 50% hospital discharge is assumed at the design stage to facilitate the initial planning. However, as shown in Table 7.10, the actual sample size required may change depending on the observed SC hospital discharge rate. Now, there is a need to convert the 50% (1682.24) and 70% (2355.14) interim information fractions to the sample sizes to facilitate the timing of interim analyses. The interim analyses are expected to be undertaken at corresponding total sample sizes of approximately 1562 and 2188 (rounded upwards), respectively. Table 7.11 summarises the stopping boundaries of the design presented on three scales.

Table 7.11. Stopping boundaries for RATPAC information based group sequential design.

Boundary Scale	Information fraction (n)	Cumulative α spent	Cumulative β spent	Stopping boundary values			
				Efficacy		Futility	
				Lower	Upper	Lower	Upper
Absolute difference	0.50 (1562)	0.003	0.040	-7.2%	7.2%	-0.9%	0.9%
	0.70 (2188)	0.015	0.099	-5.1%	5.1%	-2.2%	2.2%
	1.00 (3348)	0.500	0.200	-3.5%	3.5%	-3.5%	3.5%
P-value	0.50 (1562)	0.003	0.040	0.002	0.002	0.703	
	0.70 (2188)	0.015	0.099	0.007	0.007	0.287	
	1.00 (3348)	0.500	0.200	0.023	0.023	0.045	
Z statistic	0.50 (1562)	0.003	0.040	-2.963	2.963	-0.381	0.381
	0.70 (2188)	0.015	0.099	-2.462	2.462	-1.065	1.065
	1.00 (3348)	0.500	0.200	-2.002	2.002	-2.002	2.002

Note: The lower and upper boundary p-values for declaring futility are the same because the futility region is the same.

7.5.4.2 Interim Monitoring and Decision-Making

By the time of the 1st interim analysis, the cumulative information fraction is expected to be 1682.24 (sample size of 1562) assuming the 50% SC hospital discharge rate is correct. After the recruitment of 1562 participants, the observed discharge rates in the SC and PoC arms were 111/780(14.23%) and 236/782(30.18%), translating to an absolute difference of 15.95% in favour of the PoC arm. The standard error of this interim absolute difference is approximately 0.0206382. By invoking equation (2:41), the interim formation fraction is approximately 2347.77, which is far ahead of the expected information fraction at this point of 1682.24. In fact, the interim information fraction is 69.8% of the maximum than the expected 50% at this point. This reflects a marked overestimation of the SC discharge rate at the design stage; 50% against the observed ~14% hospital discharge rate at the 1st interim analysis.

Since the interim information fraction of 2347.77 is less than the required maximum of 3364.48, further recruitment would be required if the interim stopping boundaries were not crossed as shown in Figure 2.5. When conducting interim analysis, the 1st interim (50%) stopping boundaries shown in Table 7.11 are no longer valid. The new stopping boundaries are recalculated to mirror the observed interim information fraction of 2347.77. For instance, the new upper and lower efficacy boundaries on a Z statistic scale become 2.443 and -2.443, respectively. In addition, 1.130 and -1.130 would be the corresponding new upper and lower futility stopping boundaries.

Since the observed interim test statistic of 7.728 is far greater than the upper efficacy stopping boundary of 2.443, the trial would be stopped early to declare efficacy of the PoC arm. The unadjusted absolute difference (95% CI) and the adjusted one using stepwise ordering of the sample space are similar; 15.9% (11.9% to 20.0%; p-value<0.0001). The exact p-value is 1.096^{-14}. The intervention effect (95% CI) on an OR scale adjusted for centre effect is 2.97(2.27 to 3.90).

In the circumstance that the stopping boundaries are not crossed at the 1st interim analysis, the sample size would need to be recalculated given the observed and expected maximum information fraction. The revised projected sample size required to reach the planned maximum information fraction would be given by:

$$n(t_1) \times \frac{I_{max}}{I(t_1)} = 1562 \times \frac{3364.48}{2347.77} \approx 2238.43$$

A total of 2239 (rounded upwards) participants would now be required to reach the maximum information fraction rather than the 3348 planned assuming an inaccurate 50% SC hospital discharge. The revised sample size under this design is consistent with the average sample size using the blinded SSR method presented in Section 7.5.1.1, assuming the trial recruited to the scheduled end. However, in this case, it would be stopped at the 1st interim analysis after recruiting 1562, a saving of 677 participants under the revised sample size.

7.6 Discussion

This section highlights the main findings from the case studies used and discusses implications for the application of the methods. Lessons learned are discussed and some limitations highlighted to help the planning of future trials with similar characteristics.

7.6.1 Lessons Learned from Sample Size Re-estimation

Retrospective application of blinded SSR revealed marked inaccuracy in the design assumptions made regarding the pooled event rate. Of the three case studies considered, only the 3Mg trial made reasonably accurate assumptions about the pooled event rate and the others potentially overestimated the desired sample size by a significant amount. This may highlight some limitations of the 'restricted' SSR in cases where the planned sample size is found to be significantly larger than re-estimated since the design does not allow reduction in sample size. Although the case studies used an 'unrestricted' SSR, other considerations should be made before reducing the

sample size. For instance, it could be argued that since the two case studies were planned with 80% power, investigators may choose to increase the power to 85% say before they terminate recruitment.

Across all case studies, the trial estimates of pooled event rate were inaccurate and associated with marked uncertainty during the recruitment of approximately of the first 250 to 300 participants in total. Teare et al (2014) showed diminishing gain in precision around the event rate after enrolment of 200 participants. The authors considered event rates ranging from 10% to 90%. Therefore, the application of SSR after the minimum recruitment of 300 participants in total when the estimation of the pooled event rate is of interest appears to be a logical and conservative approach. For confirmatory trials often involving large numbers of participants, the greater the number of participants above 300 at the interim analysis, the better the performance of the SSR procedure.

Most importantly, all the case studies failed to meet recruitment targets based on design assumptions within the planned duration. Despite the overestimation in the sample size by a significant amount in two of the case studies, the research team requested additional funding from the Public Funders to meet the planned sample sizes based on inaccurate assumptions. For example, the RATPAC trial required an average of 2253 participants against a planned total of 3130. At the time of the additional funding request, the research team had already recruited 2243 participants, which was sufficient to address the primary research questions based on the observed hospital discharge. In essence, there would have been no need for the additional funding request if SSR was planned and factored into the design.

Blinded SSR may not produce the most accurate estimates compared to unblinded SSR when the intervention effect is overwhelmingly huge which was the case for RATPAC trial. However, the method is still better than doing nothing about inaccurate estimates of design parameters. In addition, it is simpler to implement and minimises operational bias.

Inaccurate assumptions regarding the pooled or control event rate may raise questions regarding the relevance of the trial and/or the clinically relevant effect sought. The performance of SSR and the decision-making process is sensitive to the characterisation or scale of the effect size sought in the presence of huge uncertainty. It is therefore important for research teams to carefully consider the scale of the clinical effect relevant to detect, taking into account the clinical interpretation and influence of uncertainty around design parameters. Furthermore, whenever SSR is considered, the SSR approach and scale of the effect size sought should be pre-specified in some study related document in order to preserve the scientific integrity of the trial. The case studies reinforce the need

for robust sample size estimation using evidence based design assumptions such as from systematic reviews of similar studies. The importance of SSR is to address residual uncertainty, and is not a remedy for poor sample size estimation at the design stage.

SSR may raise some fundamental questions regarding the value for money and efficient use of trial participants in cases when an interim decision recommended an increase in the sample size. Such a decision would be logical if the interim results are promising. Otherwise, such an increase would be a waste of resources, participants, and time. It could be helpful to consider SSR designs that allow for sample size increase only when the results are promising. An information based GSD only with futility stopping boundaries could be a potential solution to this problem. Some authors in the private sector considered alternative SSR methods using this promising results idea (Chen et al., 2004; Chen, Li, et al., 2015b; Jennison and Turnbull, 2015; Mehta and Pocock, 2011). However, the advocates of this approach argue for its use in situations where the clinical effect sought is not well defined. That is, when there is uncertainty around the clinical effect size assumed under H_1. On the other hand, there is a potential risk of increasing the sample size based on promising interim CP resulting in a trial with a highly statistically significant p-value, but small and irrelevant clinical effect unless the increase in sample size is appropriately capped.

7.6.2 Lessons Learned from Stochastic Curtailment Futility Analysis

The application of one futility analysis based on stochastic curtailment using CP described in Section 2.6 of Chapter 2 showed promise as a technique in public sector trials. For the 3CPO trial, this approach could have averted the unnecessary recruitment of a significant number of participants and hence saved resources and time. Furthermore, additional funding to achieve the planned recruitment would have been unwarranted given that it was determined that the trial was unlikely to yield clinically relevant and statistically significant results.

For the Booster trial, the research team struggled with recruitment and requested additional research funding. The CP assuming the observed interim effect or 50% of H_1 effect for the remaining data suggests that the trial is very unlikely to produce clinically relevant and statistically significant results. The trial revealed some challenges in decision-making using CP as highlighted in Section 2.6.11 of Chapter 2. For instance, the CP assuming H_1 suggested that the trial had moderate chance to detect a clinically relevant and statistically significant result, contradicting decisions assuming the observed interim effect or 50% of H_1 effect. In reality, the Booster trial faced challenges with recruitment and retention of participants with valid primary outcome data. The research

team requested additional funding and Public Funders asked for an independent futility analysis. Taking other considerations into account, Public Funders used the low value of the CP assuming the observed interim trend for the remaining data to complement their decision-making and declined an additional research funding request.

In contrast, the RATPAC trial highlighted an overwhelming intervention effect larger than expected under H_1, with CP reaching 100% after 38% or approximately 53% of the target recruitment under the design or revised sample size, respectively. Since early stopping for efficacy based on stochastic curtailment is discouraged, the use of a GSD to make efficacy early stopping decisions has been illustrated. However, when the RATPAC research team requested an additional funding extension to complete planned recruitment, Public Funders requested independent calculation of the CP. Based on the observed CP, the Funders declined additional funding on the grounds that the trial had already addressed the intended objectives. In other words, the CP close to 100% revealed to the Public Funders that the observed intervention effect was overwhelmingly larger than assumed under H_1. Again, this demonstrates that stochastic curtailment methods can help Public Funders in decision-making when faced with practical realities of additional funding requests.

The case studies highlighted the instability of the CP and inconsistencies in decision-making under various assumptions of future trend of unobserved data during the early course of the trials. This is a region associated with huge uncertainty around the intervention effect, learning effects, and poor representation of participants across centres. More so, early stopping under H_1 is difficult unless the futility analysis is delayed.

In summary, in the context of the case studies employed, the use of CP appeared useful and consistent in decision-making after 50% of target recruitment. In this regard, if the use of CP is contemplated in similar trials, it seems reasonable to conduct futility analysis between 50% and 75% of target recruitment if possible. These results are consistent with findings from retrospective analyses of 10 case studies using survival outcomes (Jitlal et al., 2012). Furthermore, Public Funders may adopt a similar mandatory rule for all trials requesting additional funding. There is a need to develop processes and procedures to guide the conduct and communication of interim results during futility analysis among key stakeholders to preserve the integrity of the trials.

7.6.3 Lessons Learned from Group Sequential Design

The design, implementation, and interim decision-making of group sequentially designed RCTs has been demonstrated. Simulations were performed to facilitate understanding of design properties and to aid the interim

decision-making process. Trends of interim results prior to the interim analysis at the point of early stopping were presented superimposed with the planned GSD to aid interpretation.

The two trials considered illustrated missed opportunities for early stopping for either futility or efficacy. The RATPAC trial could have been stopped early due to overwhelming benefit of the investigative PoC intervention and the 3CPO for no difference in mortality. Similar conclusions were reached for RATPAC trial using a different approach (Sutton et al., 2012). Both case studies illustrated that further recruitment was unnecessary since early stopping decisions were consistent with the observed results based on all recruited participants. Resources and time could have been saved and decision-making expedited. For instance, the SC arm for RATPAC trial could have been withdrawn from practice earlier and saved many lives. More so, early stopping of trials may mitigate unnecessary research funding requests for recruitment extensions, which was the case for these case studies.

The interim results and decision-making for both case studies were consistent with the observed final results. Although these findings may not be generalisable to all trials, the results provide some reassurance of the robustness of ADs in decision-making, which is one of the major concern raised in Chapters 3 and 4. The trends in the intervention effects appeared unstable during the recruitment of approximately the first 300 to 400 participants in total. It is therefore important to delay the 1st interim analysis to avoid premature early stopping. This also permits waning of leaning effects and better representation of participants across centres for generalisability. Furthermore, results from Chapter 6 suggest that trials are most likely to be stopped early with median (IQR) of 65% (50% to 85%) of the planned sample size (or number of events). This suggests that performing interim analysis after 50% of the panned sample size (or number of events) may be reasonable.

The presentation of trends of interim results prior to early stopping in trial related publications at the point of or after trial termination is important. It may facilitate effective communication and alleviate concerns about robustness of ADs in decision-making and credibility to change practice, when trials are stopped early.

7.6.4 Lessons Learned from an Information Based Group Sequential Design

In the case of the RATPAC trial, the interim results and decision-making of the standard and information based GSDs are similar. This could be due to the fact that the intervention effect was overwhelmingly huge. However, given the overestimation of the SC hospital discharge rate, the information based GSD offered an added advantage of self-correction – a form of ‘unrestricted’ SSR within a group sequential test. As highlighted in

Section 7.5.1.1 in light of the observed SC hospital discharge rate, the fixed sample size for the RATPAC trial was significantly overestimated. In fact, the maximum information fraction would have been reached after the recruitment of approximately a total of 2239 participants. In cases where there is little information to inform the design and SSR and early stopping are adaptations of interest, the information based GSD offers added advantages over a standard GSD without SSR in the presence of marked uncertainty around design nuisance parameters.

7.6.5 Reflection on Limitations

The case studies considered had immediate to short term primary endpoints where delayed responses are not an issue since recruitment can be paused at interim analyses without major impact on the conduct and duration of the trial. As a result, the handling of delayed responses has not been illustrated. However, the characteristics of the case studies considered are typical of trials where ADs have huge application potential as highlighted throughout this thesis.

The case studies presented were based on binary primary outcomes because of the limitations of available data. As a result, ADs methods applicable to continuous outcomes could not be demonstrated and explored. Furthermore, only selected types of ADs, which are relatively simple to implement and perceived to have huge potential in public sector confirmatory trials have been considered. It is important to further illustrate the application of more complex ADs, which have not been considered here in order to bridge the gap between theory and practice. The case studies here are specific examples and therefore findings such as missed opportunities and robustness in decision-making may not be generalisable. However, the case studies were for illustrative purposes and lessons have been learned to guide the planning of future adaptive trials with similar characteristics.

7.6.6 Direction of the Remainder of the Thesis

Chapter 2 reviewed different types of confirmatory ADs from a statistical and practical perspective. Chapters 3 and 4 investigated roadblocks and facilitators to the use of confirmatory ADs using in-depth interviews and surveys, respectively. Building on these findings, Chapters 5 and 6 reviewed case studies of ADs used in clinical trials practice and investigated the state of their reporting, respectively. This chapter has demonstrated the design aspects, interim monitoring, decision-making process, and potential missed opportunities using retrospective planned case studies. Important lessons have been learned to guide the planning of adaptive trials with similar characteristics. Building on this, the next chapter illustrates the prospective planning of ADs using

two actual grant applications submitted for funding aided with simulation work and exemplars to help Clinical Trialists. Discussions with the research teams are presented to help reflect on some of the challenges and considerations during the planning of ADs highlighted in interviews and surveys in Chapters 3 and 4.

Chapter 8. Design and Planning of Prospective Case Studies

8.1 Introduction

In the previous chapter, retrospective case studies were used to demonstrate the statistical design and implementation of certain types of ADs. Potential lost opportunities, robustness of and some limitations of considered ADs were highlighted. Equally importantly, lessons were learned to inform the design of future trials with similar characteristics. For these case studies, the original trials were completed fixed sample size designs with published results. The trials were however redesigned and reanalysed as if they were ADs. Building on Chapter 7 and lessons learned, this chapter demonstrates the design and planning of ADs using two prospective case studies. Both case studies were actual grant applications submitted to the NIHR HTA programme (References 13/55/43 and 13/115/101) for funding which I was involved in as a co-applicant during the course of this thesis.

As pointed out in the discussion section of Chapter 3, one way to improve the practical use of ADs is for Clinical Trialists to put forward AD-related grant proposals for consideration whenever they are appropriate to address the research questions. In addition, findings from Chapters 3 and 4 highlighted Public Funders' receptiveness to consider the funding of appropriate AD-related grant proposals. These case studies were submitted to Public Funders for consideration, with the aim of bridging the gap in applied knowledge and to improve the appropriate application of confirmatory ADs. Here, the case studies are described and discussed with focus on mitigating some of the highlighted obstacles in Chapters 3 and 4.

This chapter acknowledges the research team and co-applicants on these case studies, particularly the Chief Investigators (Prof Robert Storey and Mr Sabapathy Balasubramanian), Sheffield CTU proposal developer (Dr Judith Cohen), and Senior Statistician (Prof Steven Julious, my supervisor) for their contributions during proposal development. The Chief Investigators gave consent to use the grant applications as case studies.

8.2 Aims and Objectives

Building on Chapter 7, this chapter aims to describe the design and planning of two adaptive grant proposals that were submitted to Funders. Specifically, an aim is to help facilitate communication of design related aspects in future grant applications and the planning process of ADs. In addition, a focus is to highlight the discussions which took place as these can help future research teams during the planning process and to provide

an exemplar of how to communicate the statistical properties of ADs in grant proposals. Finally, an aim is to reflect on lessons learned during the grant application process of the case studies in the context of roadblocks to the use of ADs presented in Chapters 3 and 4.

8.3 PENNYWISE Study

This section describes the design and discussions held with the research team during the development of a grant application for the PENNYWISE trial. The application was submitted to the NIHR HTA programme during the course of this thesis, but was unsuccessful because it failed to meet the clinical research priorities set by the Funder. However, there was no feedback highlighting any shortcomings of the proposed AD and the study presented a good example of the prospective application of an AD.

8.3.1 Brief Background

The current standard of care for a significant number of patients in the UK who present with large heart attacks known as STEMIs (ST-elevation myocardial infarctions), which can be diagnosed by electrical tracing of the heart, is via insertion of a wire catheter to place a stent and thereby recanalise the artery. This intervention is followed by balloon treatment to disperse blood clot and insertion of a metal mesh or stent into the wall of the artery to keep it open. This intervention, known as primary percutaneous coronary intervention (PPCI), is often highly effective at treating the heart attack. However, anticlotting drugs must be given intravenously and orally at the time of the procedure and afterwards to prevent further blood clots forming in the artery, which sometimes blocks off the stent and can be fatal.

The best combination of intravenous and oral anticlotting drugs is unknown and there are numerous possibilities now available. However, all the available options are limited by suboptimal effectiveness in preventing stent blockage or increased risk of serious bleeding or excessive cost or a combination of two or more of these. New oral anticlotting drugs, prasugrel and ticagrelor, reduce the risk of stent blockage but can sometimes take up to 8 hours to reach their full effect in PPCI patients due to slow absorption and so ideally intravenous anticlotting therapy should cover this critical period after stent insertion. Enoxaparin is a relatively cheaper anticlotting drug that has shown promising results in PPCI when given as a single bolus injection. The research team for the PENNYWISE study wanted to assess whether a novel regimen of enoxaparin, given as a bolus

followed by prolonged intravenous infusion for up to 6 hours, offers the best combination of efficacy, safety and cost-effectiveness in PPCI patients compared to standard of care anticlotting strategies.

8.3.2 Study Design and Primary Endpoint

The study was designed as a two arm, multi-centre, non-inferiority, open-label, parallel group, group sequential RCT with 1:1 allocation ratio. The severity of the medical condition (STEMI) and the need to expedite the decision-making process to approve the clinically and cost effective intervention were the main reasons why the research team were keen to consider early stopping for futility or efficacy as a design feature. In addition, the Chief Investigator argued that early stopping would save resources and time in the view of large number of participants required for this trial, as highlighted in Section 8.3.3. The research team were interested in a composite primary endpoint (recurrent MI, stroke, death or definite stent thrombosis) at 30 days. The realisation of the primary endpoint within 30 days was ideal for an adaptive trial. The event rate of the composite primary endpoint in the standard care (SC) bivalirudin therapy arm in approximately 55% of the centres in the UK was assumed to be around 8%. The research team justified that the investigative prolonged enoxaparin (PE) is deemed non-inferior to SC when the associated event rate is less than 10%. That is, a non-inferiority margin (NIM) of 2%.

8.3.3 Primary Hypothesis and Sample Size Estimates

The absolute risk difference in event rates of the composite primary endpoint between the two arms was the intended primary analysis. The H_0 and H_1 are configured as follows for a non-inferiority test with a 2% NIM.

H_0: PE is inferior to the SC arm ($p_t - p_c \geq 2\%$)

H_1: PE is non-inferior to the SC arm ($p_t - p_c < 2\%$)

Assuming an 8% SC event rate, NIM of 2%, 1:1 allocation ratio, one-sided type I error rate of 2.5% and power of 90%; a fixed trial design with only one analysis at the scheduled end would require a total of 7734 participants (3867 per group). Six GSDs allowing for stopping early either for futility (claiming inferiority) or efficacy (claiming non-inferiority) with two interim analyses at 50% and 75% of the planned recruitment, and final analysis at the scheduled end were considered and presented to the research team. In consideration of lessons learned from Chapter 7, the Chief Investigator agreed the timing of interim analyses as reflected in Section 8.3.4. The detailed properties of these designs are summarised in Table 8.1 to preserve 90% power and 2.5% one-sided type I error. For instance, considering more conservative LD α and β spending boundaries for efficacy and futility

equivalent to the OBF, respectively, the expected maximum total sample size required to preserve a 90% power and one-sided type I error is 8374 (4187 per arm). If H_0 is true (PE inferior), the average total sample size is 5136; a potential saving of 2598 participants compared to a fixed sample design. Similarly, if H_1 is true (PE non-inferior), the average total sample size needed is 6124 with a saving of 1610 participants. With 100,000 simulated trials, the average proportions of trials stopping early at the 1st and 2nd interim analyses and final analysis are 30%, 47% and 23%, respectively. Of the trials that stop early at these interims, 94%, 91% and 81% would stop for non-inferiority, respectively. To claim non-inferiority at the 1st and 2nd interim analyses, and final analysis, the one-sided p-value would need to be less than 0.0015, 0.0092 and 0.0220, respectively.

The design with LD α and β spending boundaries for efficacy and futility equivalent to Pocock type is less conservative as demonstrated by significantly larger numbers of simulated trials stopping early (76%) at the 1st interim analysis and has the largest maximum total expected sample size. This design was presented to the investigators for completeness.

The design properties with Gamma ($\gamma = -2$) and Rho ($\rho = 2$) family stopping boundaries are similar to one another and require a higher level of evidence to claim non-inferiority at the scheduled end than the planned 2.5% nominal level. The WT ($\partial = 0.2$) and WT ($\partial = 0.25$) are variants of the same design but with different shape parameters used to compute the stopping boundaries. Their properties are quite similar but the WT ($\partial = 0.2$) requires 198 fewer participants than the WT ($\partial = 0.25$). Results presented in Table 8.2 are replicated from Table 8.1, assuming a study power of 85% rather than 90%.

Table 8.1. Statistical properties of six group sequential designs for the PENNYWISE trial at 90% power.

Stopping rules or boundaries		Fixed design sample size	Average total sample size				Interim analyses (sample size)	Rejection P-value		Average probability of stopping early for any reasons	Average probability of stopping early for:	
			Max	H_0	H_1	$H_{0.5}$		Reject H_0	Reject H_1		Non-inferiority	Futility
Extremely conservative	LD (OBF)	7734	8374	5136	6124	6559	0.50 (4187)	0.0015	0.3701	30%	94%	6%
							0.75 (6281)	0.0092	0.0982	47%	91%	9%
							1.00 (8374)	0.0220	0.0220	23%	81%	19%
	Gamma family ($\gamma = -2$)	7734	8576	5127	5769	6450	0.50 (4287)	0.0067	0.3139	50%	95%	5%
							0.75 (6431)	0.0100	0.1058	30%	91%	9%
							1.00 (8576)	0.0186	0.0186	20%	78%	22%
	Rho family ($\rho = 2$)	7734	8594	5149	5783	6458	0.50 (4297)	0.0062	0.3241	49%	95%	5%
							0.75 (6446)	0.0109	0.0989	33%	90%	10%
							1.00 (8594)	0.0184	0.0184	19%	76%	24%
	WT ($\partial = 0.2$)	7734	8794	5043	5753	6265	0.50 (4397)	0.0063	0.2472	53%	92%	8%
							0.75 (6595)	0.0136	0.0757	34%	89%	11%
							1.00 (8794)	0.0214	0.0214	14%	79%	21%
	WT ($\partial = 0.25$)	7734	8962	5077	5743	6262	0.50 (4481)	0.0075	0.2269	56%	93%	7%
							0.75 (6721)	0.0140	0.0709	32%	88%	12%
							1.00 (8962)	0.0205	0.0205	12%	82%	18%
Less conservative (Extremely liberal)	LD (Pocock)	7734	10472	5597	6059	6612	0.50 (5236)	0.0155	0.1294	76%	92%	8%
							0.75 (7854)	0.0104	0.0499	17%	88%	12%
							1.00 (10472)	0.0100	0.0100	7%	75%	25%

SC: Standard Care; LD: Lan and DeMets; WT: Wang and Tsiatis; OBF: O'Brien and Fleming; Max: maximum; NIM: non-inferiority margin; 90% power, 2.5% one-sided type I error, 2% NIM and 8% SC event rate.

Table 8.2. Statistical properties of six group sequential designs for the PENNYWISE trial at 85% power.

Stopping rules or boundaries		Fixed design sample size	Average total sample size				Interim analyses (sample size)	Rejection P-value		Average probability of stopping early for any reasons	Average probability of stopping early for:	
			Max	H_0	H_1	$H_{0.5}$		Reject H_0	Reject H_1		Non-inferiority	Futility
Extremely conservative	LD (OBF)	6610	7320	4354	5457	5526	0.50 (3660) 0.75 (5489) 1.00 (7320)	0.0015 0.0092 0.0220	0.3088 0.0877 0.0220	28% 47% 26%	86% 89% 79%	14% 11% 21%
	Gamma family ($\gamma = -2$)	6610	7330	4385	5125	5470	0.50 (3665) 0.75 (5497) 1.00 (7330)	0.0067 0.0100 0.0186	0.3140 0.1073 0.0186	45% 31% 24%	91% 87% 72%	9% 13% 27%
	Rho family ($\rho = 2$)	6610	7346	4405	5136	5479	0.50 (3823) 0.75 (5734) 1.00 (7346)	0.0062 0.0109 0.0184	0.3252 0.1001 0.0184	43% 33% 24%	92% 86% 71%	8% 14% 29%
	WT ($\partial = 0.2$)	6610	7646	4308	5090	5273	0.50 (3823) 0.75 (5734) 1.00 (7646)	0.0066 0.0141 0.0220	0.2146 0.0703 0.0220	49% 36% 15%	86% 86% 78%	14% 14% 22%
	WT ($\partial = 0.25$)	6610	7794	4347	5083	5279	0.50 (3897) 0.75 (5846) 1.00 (7794)	0.0078 0.0145 0.0210	0.1981 0.0661 0.0210	53% 33% 14%	87% 86% 78%	13% 14% 22%
Less conservative (Extremely liberal)	LD (Pocock)	6610	9024	4809	5353	5624	0.50 (4512) 0.75 (6768) 1.00 (9024)	0.0155 0.0104 0.0100	0.1243 0.0490 0.0100	72% 19% 9%	87% 83% 72%	13% 17% 27%

SC: Standard Care; LD: Lan and DeMets; WT: Wang and Tsiatis; OBF: O'Brien and Fleming; Max: maximum; NIM: non-inferiority margin; 2.5% one-sided type I error, 2% NIM, and 8% SC event rate.

8.3.4 Selection of the Desired Design

Discussions were held with investigators in order to make a decision regarding the desired design. The investigators favoured a delay in conducting the 1st interim analysis for a number of reasons. First, to avoid premature early stopping and provide convincing evidence to change practice. Second, the health economics team member believed that such a timing would provide adequate information for health economics evaluation, which is important to support the non-inferiority decision-making process. Third, the Chief Investigator highlighted that fair representation of participants across centres at the time of the 1st interim analysis would be important for generalisability of findings. Finally, the Chief Investigator wanted to avoid terminating the trial too early at a stage when learning effects are high. As a result, the investigators suggested conducting the 1st interim analysis at 50% of the targeted recruitment. The investigators also wanted a design with a reasonable chance of stopping early for either futility or non-inferiority, if the interim evidence is overwhelming. The Gamma ($\gamma = -2$), Rho ($\rho = 2$), WT ($\partial = 0.2$) and WT ($\partial = 0.25$) families were potential choices with reasonable chances of stopping early. In view of the feasible maximum total sample size to recruit across eligible centres in the UK, the investigators selected WT ($\partial = 0.25$) as a desirable design. A stratified block randomisation procedure with variable block size was considered to ensure fair distribution across centres per intervention arm at the interim analyses.

8.3.5 Sensitivity Analysis of the Statistical Properties of the WT ($\partial = 0.25$) Design

As highlighted in Chapter 7, it is important to understand the statistical properties of the proposed design under a range of plausible scenarios. Here, 100,000 trials were simulated for each scenario and statistical properties of the design and their implications investigated. Table 8.3 summarises the results of sensitivity analyses around the assumed SC event rate and its impact on the overall trial power, probability of stopping early at interims, and associated chances of stopping early for either non-inferiority or futility. The power, one-sided type I error and NIM were fixed at 90%, 2.5% and 2%, respectively. For instance, in the event that the SC event rate is 10% instead of the assumed 8%, the trial would have an 82.5% power. The trial would have less than 80% power when the observed SC event rate is above 10%. A trial powered at 90% provides a safeguard of at most 9% in case of underestimation of the SC event rate. In contrast, overestimation of the SC event rate yields a trial that has more power than pre-planned. The chances of stopping early either for non-inferiority or futility at the 1st interim analysis increases as the assumed SC event rate gets smaller for a fixed NIM of 2%.

Table 8.3. Study properties for WT ($\partial = 0.25$) for varying SC event rate scenarios.

SC event rate	NIM	Expected power	Interim analyses	Probability of stopping early for any reason	Average probability of stopping early for:	
					Non-inferiority	Futility
10%	2%	82.5%	0.50	48%	84%	16%
			0.75	36%	85%	15%
			1.00	16%	80%	20%
9%	2%	86.3%	0.50	54%	88%	12%
			0.75	33%	87%	13%
			1.00	13%	77%	23%
8%	2%	90.0%	0.50	56%	93%	7%
			0.75	32%	88%	12%
			1.00	12%	82%	18%
7%	2%	93.5%	0.50	60%	96%	4%
			0.75	29%	93%	7%
			1.00	10%	80%	20%
6%	2%	96.4%	0.50	66%	98%	2%
			0.75	26%	95%	5%
			1.00	8%	83%	17%

NIM: non-inferiority margin; SC: Standard Care.

Table 8.4 summarises additional results of the sensitivity analyses based on 100,000 simulated trials assuming the observed event rate in the investigative PE arm ranges from 10% to 6%. Here, the SC event rate, maximum sample size and NIM are fixed as planned. The impact on the proportion of trials stopping early at interims for various reasons are presented.

Table 8.4. Study properties of a WT ($\partial = 0.25$) design for varying effectiveness of Prolonged Enoxaparin.

Effectiveness of PE	SC event rate	NIM	Actual observed PE rate	Interim analyses	Probability of stopping early for any reason	Average probability of stopping early for:	
						Non-inferiority	Futility
PE little worse	8%	2%	10%	0.50	99.9%	3%	97%
				0.75	0.1%	32%	68%
				1.00	-	-	-
	8%	2%	9%	0.50	73.8%	20%	80%
				0.75	20.5%	59%	41%
				1.00	5.8%	59%	41%
	8%	2%	8%	0.50	55.5%	93%	7%
				0.75	32.5%	88%	12%
				1.00	12.0%	82%	18%
	8%	2%	7%	0.50	90.6%	100%	-
				0.75	8.5%	100%	-
				1.00	0.9%	95%	5%
PE little better	8%	2%	6%	0.50	99.7%	100%	-
				0.75	0.3%	100%	-
				1.00	-	100%	-

PE: prolonged enoxaparin; NIM: non-inferiority margin; "-" represents 0.0%.

8.3.6 Grant Submission Exemplar of the Design and Sample Size Estimates

So far, the detailed properties of the selected design have been investigated. In practice, grant applications have limited space and so cannot accommodate a lot of detail. This section therefore presents an exemplar of a concise description of the sample size estimates and properties of the design as communicated to the Funder in the grant application.

The primary outcome is a composite endpoint of whether a patient had a stroke, definite stent thrombosis, recurrent MI or died. We anticipate an 8% event rate on SC and less than 10% on PE for an assumed non-inferiority limit of 2%. For 90% power and a one-sided type I error of 2.5% the fixed sample size would be 7734 patients in total (3867 per arm). This sample size is calculated under the assumption that there will be only one analysis of the data at the end. The trial will be analysed as a group sequential trial with 3 scheduled interim analyses after 50%, 75% and 100% of patients are enrolled. Wang-Tsiatis stopping rules will be applied with a delta set at 0.25. The one-sided type I error will be maintained at 2.5%, but the maximum sample size will increase to 8962 in total, however, we would anticipate the expected trial sample size to be smaller than this.

The trial will stop for non-inferiority of PE over control (PE is not worse than SC) if the one-sided P-value is less than 0.0075, 0.0140 and 0.0205 at the 1st, 2nd, and 3rd analyses, respectively. If we observed 8% and 8% on PE and SC interventions respectively (no difference between arms), the expected sample size would be 5743 patients in total.

The trial will stop for futility (PE and standard care equivocal) if the one sided P-value is greater than 0.2269, 0.0709 and 0.0205 at the 1st, 2nd, and 3rd interim analyses, respectively. If we observe 10% on PE and 8% on control respectively (interventions are different) the expected sample size is 5077 patients in total. To investigate the properties of the design, 100,000 simulations were performed using East - a specialist adaptive design software package. Assuming SC event rate of 8%, simulations were used to investigate the proportion of trials which would be anticipated to stop at the 1st interim analysis – after 3867 patients - if the effect on PE was 6%, 7%, 8%, 9% or 10% (ranging from a little better than SC to a little worse). From simulations it is anticipated that 99.7%, 90.6%, 55.5%, 73.8% and 99.9% of trials would stop at the 1st interim analysis for PE event rates of 6%, 7%, 8%, 9%, and 10%, respectively. For trials that stop early at this interim 0%, 0%, 7%, 80% and 97% would stop for futility, respectively.

The sample size of 8962 is therefore the maximum sample size. On average, a smaller sample size than this is expected, and one smaller than a fixed sample size design, if the effect is: as anticipated under the alternative hypothesis; as anticipated under the null hypothesis; bigger than expected; smaller than expected.

8.3.7 Costing of the Grant Application

In addition to standard costing for fixed sample size designs, the total grant costs were estimated under different recruitment scenarios, allowing for the same amount of time for trial set-up and close-out, and 2 months to convene IDMC meetings to make early stopping recommendations. Furthermore, a 1 to 2 month recruitment pause was factored in to allow for delayed responses, data cleaning and analysis, and the decision-making process.

Assuming the trial recruited the maximum sample size of 8962 patients, the recruitment was expected to take 40 months with a trial duration of 63 months. The maximum total cost would be £2,282,298. If the trial were terminated at 75%, after recruiting 6721 participants, the total trial duration would be 56 months with a recruitment period of 30 months. This corresponds to a total cost of £1,915,285; £367,013 less than the cost of a trial recruiting the maximum sample size. Finally, a 48-month study with 21 months recruitment would be expected if the trial is stopped at 50% - after recruiting 4481 participants; for a total cost of £1,555,242. This translates to total cost savings of £727,055 and trial duration of 15 months compared to a scenario of maximum sample size recruitment. The costing scenarios were presented in the grant application.

The lessons learned during the development of the PENNWISE grant proposal are discussed in Section 8.5.1.

8.4 NERVE BLOCK Study

This case study describes the design that was put forward during the development of another grant application for which I was a co-applicant. The proposal was submitted to the NIHR HTA programme, although it failed to meet the clinical research priorities of the funding panel. Nevertheless, the feedback received from the funding panel did not raise any concerns relating to the design. Like with PENNYWISE the case study is used for illustrative purposes as a case study in prospective AD planning.

8.4.1 Brief Background

In England, over 11,000 thyroid and 3000 parathyroid related procedures are performed every year. The investigators for the NERVE BLOCK study argued that there are no established standards on postoperative pain relief following these procedures. Although all patients are provided with regular oral pain relief following surgery, some patients may suffer significant pain needing parenteral opiates - up to 90% in some cases. Parenteral

opiates are associated with postoperative nausea and vomiting in over 50% of patients. In addition to opioid and non-opioid pain medications, the use of local anaesthetic agents to infiltrate the wound or to block the superficial cervical plexus that innervates the area of surgery has been described. However, clinical practice across England with regard to the use of these techniques varies considerably. Some centres use no local anaesthetic at all; some employ local wound infiltration (LWI); some provide a nerve block (NB); and others use a combination of these techniques. LWI and bilateral superficial cervical plexus block (BSCPB) may alleviate pain and reduce nausea and vomiting, but concerns about safety prevent widespread use.

The research team for the trial argued that a recent summary of studies on BSCPB showed some benefit in pain control and the procedure was shown to be safe (Warschkow et al., 2012). However, no recommendations were possible as the evidence was limited and the authors suggested further trials to evaluate the appropriate dose and effects on nausea and vomiting given that LWI provides effective pain relief. Although BSCPB may be as effective as, if not superior to LWI, the combination has not been adequately assessed in an appropriately sized trial. Hence, the research team aimed to evaluate the effectiveness and safety of Bupivacaine (a loco-regional anaesthetic agent) as an agent for LWI or BSCPB, or both in comparison to a placebo (saline) in reducing postoperative pain, nausea and vomiting following thyroid and parathyroid surgery.

8.4.2 Design Issues and Adaptive Aspects

For this study, the feasible sample size across centres considered by the NERVE BLOCK research team was limited to around 500 participants. The research team wanted to address a number of questions using that limited participant pool. In addition, they wanted a design which:

- Allows simultaneous evaluation of both BSCPB and LWI against the placebo;
- Allows additional evaluation of the effectiveness of the combination of BSCPB and LWI against the placebo;
- Increases the proportion of patients receiving active interventions;
- Enables a more informative and efficient process to expedite clinical decision-making.

As a result, an adaptive factorial design was considered, with four intervention arms: LWI only, BSCPB only, LWI and BSCPB, and placebo. The research team did not expect an interaction between LWI and BSCPB. Adaptive features were planned to allow for a pre-planned change in strategy depending on interim results, enabling a flexible and efficient trial design that addresses clinical questions in order of importance as displayed

in Figure 8.1. The research team wanted the primary comparison to be difference in pain relief between BSCPB and placebo on the first postoperative day. The secondary and least important comparisons were LWI versus placebo, and LWI and BSCPB versus placebo, respectively. A hierarchical testing procedure was adopted to control for multiple testing of comparisons with respect to the order of clinical importance (Dmitrienko et al., 2010). The hierarchical testing strategy shown in Figure 8.1 within a GSD controls for multiple testing due to interim analyses and multiple comparisons. Additional data were required for health economics evaluation performed only if either BSCPB or LWI is superior to a placebo.

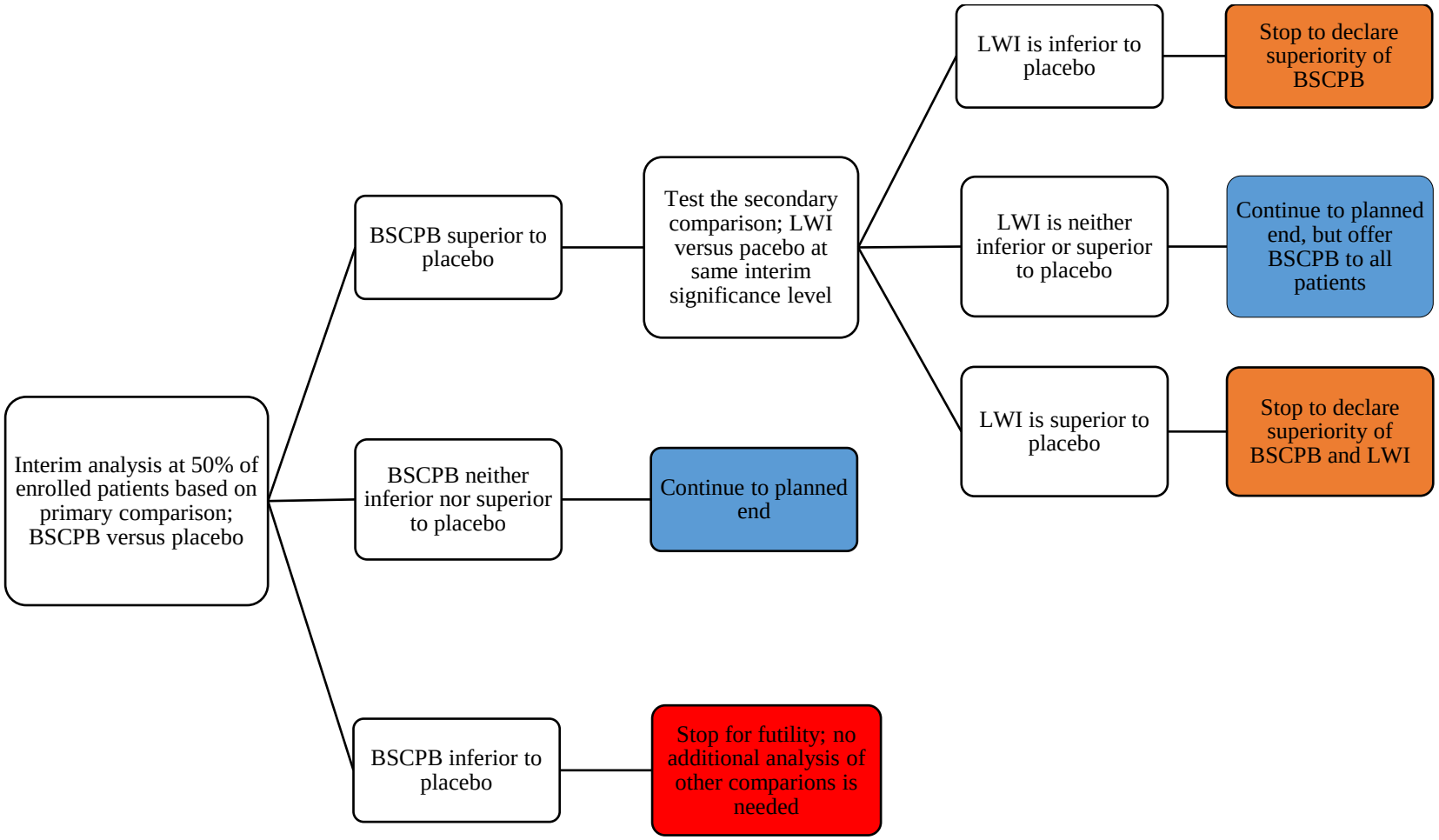

Figure 8.1. Adaptive hierarchical testing strategy within a group sequential test for the NERVE BLOCK trial.

8.4.3 Sample Size Estimates and Planned Analysis

The research team wanted to detect a 0.5 difference in VAS for pain between BSCPB and placebo. A 1.5 SD was assumed based on available literature. In addition to the final analysis at the scheduled end, the study team wanted one interim analysis after 50% of patients were enrolled, consistent with results of Chapter 7. The timing of the 1st interim was chosen based on similar reasons to those highlighted in Section 8.3.4 and also due to the fact that the trial would have a power close to 60%. Due to the constraint in the feasible sample size, a GSD with Pampallona and Tsiatis ($\delta_0 = 0.35, \delta_1 = 0.35$) stopping rules was considered to allow for early stopping for futility and efficacy. 'Binding' futility stopping boundaries were considered as advised by the research team. Assuming a 90% power, a one-sided type I error of 2.5%, and equal allocation, the maximum sample size is 436.2; rounded up to 440 in total to allow for even distribution of patients between arms. This was inflated to 464 to allow for a 5% dropout or withdrawal rate.

The trial was intended to stop for superiority of BSCPB over placebo if the one-sided p-value was < 0.0117 or 0.0206 at the 1st interim or final analysis, respectively. In addition, the trial would be stopped for futility at the 1st interim analysis if the associated p-value fell between 0.192 and 0.808. The trial would have a power of 57.5% at the 1st interim analysis. The assessment of safety was also an important consideration. With 220 patients receiving BSCPB and an adverse event risk of 2 in 463, there is a 40.3% chance of observing at least one adverse event.

ADDPLAN 6.1 software was used to simulate 100,000 trials to estimate the proportion of trials stopped at the interim analysis (after 220 patients). Eighty-three percent, 51.4%, 63.9%, 93.8%, and 99.6% of trials would stop for intervention effects of 0.00, 0.25, 0.50, 0.75, and 1.00, respectively. For an effect size of 0.00, 0.25, and 0.50, respectively, 82.8%, 37.3% and 6.5% of trials would stop for futility. Based on this, a smaller sample size was anticipated than the maximum sample size of 440, and one smaller than a fixed sample size design of 382 patients, if the effect: as anticipated under H_1 is bigger than expected; or as anticipated under H_0 is smaller than expected.

The primary outcome – pain scores in the morning after surgery measured by the VAS was intended to be analysed using an analysis of covariance (ANCOVA) regression model, adjusting for gender, age, type of surgery, and baseline response. An intention-to-treat primary analysis was planned. Furthermore, median unbiased results were intended to be obtained using the stagewise ordering approach.

8.4.4 Exemplar for Rationale and Costing of the Grant Application

Justification of full research costs were provided for when the study stops at either 50% or 100% recruitment. The costs accounted for expected recruitment duration; the same set-up and close up time, 2 months for analysis, convening of the IDMC and decision-making process; and other factors. Below is an extract from the grant application.

> The trial would stop for superiority (either BSCPB, LWI or both over placebo), or for futility (BSCPB and placebo are equivocal) at planned interim analyses with pre-specified limits. The design optimises the ability to answer the research question in an efficient manner with time and cost savings if it stops early, but requires some additional statistical support compared to a standard design.
>
> The maximum research cost assuming recruitment of 464 patients was estimated at £982,153; a 24-month recruitment period and total trial duration of 41 months. If the trial is stopped at 50% after recruiting 232 patients, the total research cost would be £691,889; a saving of £290,264 compared to maximum recruitment. This corresponds to a total study duration of 31 months with a 12-month recruitment period.

The lessons learned during the development of the NERVE BLOCK grant proposal are discussed in the context of previous thesis findings in Section 8.5.1.

8.5 Discussion

The NERVE BLOCK and PENNYWISE case studies illustrated the statistical design of adaptive trials, the thought process, practical considerations, and provided exemplars on how to communicate the rationale for and important aspects of ADs in grant applications. Lessons learned from previous chapters guided the design considerations for the two case studies. Although these grant applications were unsuccessful because of failure to meet the clinical priorities of the funding panels, the case studies provide some insights to help Trialists wishing to utilise ADs appropriately. The rationale for the case studies presented was motivated by the need to expedite the clinical decision-making process; optimise the evaluation of a number of clinical questions with a limited patient pool; minimise the number of participants in the control arm; save research participant pool; reduce trial duration; and save research resources.

8.5.1 Reflection on Lessons Learned in the Context of Previous Thesis Findings

Chapters 3 and 4 highlighted a number of obstacles perceived to be hampering the use of ADs in publicly funded trials. These include the lack of funding support to aid the design work, lack of time, lack of practical knowledge, limited access to case studies, and challenges in marketing ADs to key stakeholders. During the planning of the presented case studies, it has been learned that the process is truly more involved and time consuming compared to that of a fixed sample size design. The design may require statistical simulation work under a number of plausible scenarios in order to understand the design and its implications. It is important to maintain close engagement with the research team in an iterative process. For instance, the simulation scenarios should be discussed with the clinical investigators and results communicated to the research team in a way that is easy for them to understand.

The experience of the investigators is an important factor in the planning of ADs. The PENNYWISE Chief Investigator was more experienced in the conduct of multicentre trials and had knowledge about aspects of interim analysis and other considerations. In addition, the design, planning, and communication of simulation results for the PENNYWISE trial, which used a standard GSD was much easier than for the NERVE BLOCK trial, which used a factorial AD using group sequential methods. The investigators of the two trials were receptive to the idea of using ADs and the availability of design and planning support from the Sheffield CTRU made it easier for them to buy into the idea. This supports the findings of Chapter 3 that highlighted the receptiveness of Clinical Investigators to using ADs depended on the availability of the support and how they are communicated to them by Clinical Trialists.

The involvement of an experienced proposal developer with some basic understanding of ADs and related considerations is helpful. For instance, the Sheffield CTRU proposal developer worked out variable costs of the proposal with my help for variable sample size scenarios including IDMC considerations. The variable costs were presented to the Funders in the grant applications. This highlights the importance of capacity building and basic practical training of proposal developers within CTUs on AD-related aspects. This complements related findings from Chapter 3. Importantly, the Sheffield CTRU did not raise any concerns regarding the impact of early stopping on staff research contracts. This could be because the CTRU is receiving NIHR support infrastructure funding and/or has a sizable portfolio of trials running or in the pipeline.

The news of proposal rejection by Public Funders can be demoralising, particularly after spending so much time and effort on the statistical design and simulation work. More so, given that the design and planning

time of myself and Prof Steven Julious was not funded. The time commitment was driven by the desire to facilitate the appropriate use of ADs in publicly funded trials betting on uncertain success of the grant applications. In this regard, the concern about the lack of a business case for and bridge funding to support planning of ADs raised by CTU Leaders in Chapters 3 and 4 is understandable. Despite the fact that the proposals were rejected, the lack of feedback relating to the design by the Public Funders was somewhat reassuring. In addition, two Trial Statisticians (myself and Prof Steven Julious) were costed as co-applicants on both grant applications because of the additional statistical support that would be required for the study. This was explained in the grant applications and no objections were raised by the Public Funders or Reviewers.

8.5.2 Reflection on Limitations

Even though the presented case studies were actual submitted grant applications, the Public Funders rejected the proposals on grounds not related to the proposed ADs. As a result, there was no opportunity to implement the proposed ADs in practice and gain some experience to facilitate reflection on issues arising during the conduct of adaptive trials. Lastly, the investigators claimed to have enough data to inform the design – hence an information based group sequential approach was not utilised. However, the approach adopted may not be optimal if such claims are proved to be inaccurate.

8.5.3 Direction of the Remainder of the Thesis

This chapter has shared personal practical experiences of preparation of grant applications for adaptive trials. The work was motivated by findings from Chapters 3 and 4, and the need to facilitate applied learning and alleviate some of the uncovered roadblocks. Lessons learned were presented and reflected on the context of the perceived barriers presented in Chapters 3 and 4. The next chapter concludes the thesis with a discussion in the context of the aims and objectives of this work. Recommendations for best practice to improve the appropriate use of ADs are provided based on thesis findings. Finally, potential areas for future related research beyond this thesis are highlighted.

Chapter 9. Discussion and Recommendations

9.1 Introduction

Well conducted RCTs play a fundamental role in the evaluation and approval process of investigative interventions. Fixed sample size designs are currently the mainstream approach for conducting RCTs, however, some limitations of the approach are prompting a paradigm shift towards alternative designs. The increasing need to maximise value for money in clinical trials research, speed the evaluation of investigative interventions, make efficient use of trial participants, and improve design efficiency and conduct of RCTs demands innovative approaches. Appropriately used and well executed ADs have the potential to mitigate some of the shortcomings of fixed sample size designs. However, the application of ADs is limited despite their promising benefits and availability of related statistical literature.

This thesis investigated why ADs are underused in confirmatory trials in the UK, particularly those funded by the public sector. Importantly, it explored facilitators to overcome some of the uncovered obstacles in order to improve the appropriate uptake of ADs. Chapter 2 set the scene by reviewing confirmatory ADs from a statistical and practical perspective, hence it gave a foundation of the understanding of ADs referred to throughout the thesis. The reviewed literature also guided the application of ADs in subsequent chapters using retrospective and prospective case studies.

Chapter 3 provided a platform to address the main thesis aim by exploring roadblocks and facilitators to the use of ADs using in-depth qualitative interviews of multidisciplinary key stakeholders in clinical trials research. This approach provided rich information which guided the design of follow-up quantitative surveys. Importantly, the in-depth interviews enhanced understanding of the obstacles and facilitators from the perspectives of those involved in clinical trials research and the related decision-making process. Building on this work, Chapter 4 investigated wider perceptions of roadblocks and key facilitators using cross-sector quantitative surveys involving CTUs, the private sector, and Public Funders. The results of Chapters 3 and 4 informed the rest of the work of the thesis and potential areas of future research.

Chapter 5 reviewed case studies of undertaken confirmatory ADs through clinical trials registers – an important resource to facilitate practical learning and mitigate some of the uncovered barriers. Chapter 6 then investigated the state of reporting of the most commonly used AD and the shortcomings of the reporting guidance framework for ADs. Chapter 7 illustrated the design and statistical execution of ADs using retrospectively planned

case studies of completed RCTs aimed to improve applied knowledge. In addition, Chapter 8 demonstrated the design and planning of prospectively planned ADs and highlighted the issues involved using two real world grant applications submitted to the NIHR HTA programme. The lessons learned can inform the design of future related trials.

In conclusion, this chapter discusses the findings in the context of the overall thesis aims, implications to practice, and how the findings relate to recent related work. Recommendations to facilitate the appropriate uptake of ADs are summarised. In addition, a concise summary of key general considerations Clinical Trialists need to think about when contemplating the use of ADs in the confirmatory setting is provided. The aim is to enhance the thought process of researchers at the design stage to facilitate proper planning for the successful implementation of the ADs considered. Finally, potential areas of future related research beyond this thesis are highlighted.

9.2 The Main Thesis Findings

It is important to emphasise that ADs are not appropriate for every trial. Therefore, this thesis is not advocating the application of ADs when they are inappropriate. When contemplating the use of ADs, a number of considerations highlighted in Section 9.3 should be examined on a trial-to-trial basis, depending on the research question(s) and trial objective(s). The findings discussed here assume that the use of an AD is deemed appropriate.

9.2.1 The Perspective of UK CTUs on Roadblocks

The thesis established multifaceted barriers and concerns hindering the routine use of confirmatory ADs. The leading barrier unique to the public sector is the lack of funding support accessible to CTUs to help with the design developmental work of time consuming and complex ADs. In addition to interviews and survey results, the lessons learned in Chapters 7 and 8 highlighted that the design and planning of ADs is generally time-consuming and more involved than mainstream fixed sample size designs. However, the time and effort commitment depends on the complexity of the type and scope of the proposed AD, underlying knowledge of the methods, and availability of resources such as statistical software or user-written code. As a result, a few CTUs with the practical knowledge, experience and capacity to support ADs are the ones applying them more often.

The lack of practical knowledge and related experience are the leading obstacles hampering the routine uptake of ADs across sector. This is strongly connected to the lack of hands-on applied training and limited access

to case studies of undertaken ADs to facilitate practical learning and problem solving. Unsurprisingly, there is a strong preference for mainstream fixed sample size designs which are well accepted and simpler to execute compared to ADs. In view of the immense pressure to deliver on existing competing priorities of fixed sample size designs, Clinical Trialists are less likely to support time-consuming and complex ADs, even when they are appropriate. In contrast, despite the potential underreporting and poor indexing of adaptive trials, the thesis found that even simpler types of ADs appear to be underused although they are easier to plan and execute. This may be explained partly by the strong preference for mainstream fixed sample size designs by both researchers and Funders.

The thesis found that cross-sector Trialists face difficulties in marketing ADs to key stakeholders such as Funders/Sponsors, Clinical Collaborators, Investigators, and Regulators. The importance of convincing key stakeholders regarding the appropriateness and potential advantages of the proposed AD compared to competing designs for a given trial situation cannot be overemphasised. This is enhanced through better communication and presentation of design scenarios and potential benefits. The planning process of ADs is demanding requiring more close collaboration and engagement with key stakeholders compared to mainstream fixed sample size designs.

9.2.2 Cross-sector Differences in Perceptions on Roadblocks

The perceptions on most of the barriers appeared consistent across sector. Nonetheless, there are exceptions, reflecting differences in the organisational research funding structures, nature of investigative interventions and related regulatory framework, and underlying practical experiences in the conduct of ADs. For example, the additional practical complexities associated with the implementation of ADs and inadequate data management infrastructure to support the demands of adaptive trials were among leading obstacles perceived as more important by the private sector than by the public sector. Perhaps this is because of the differences in practical experiences influencing the underlying knowledge of the practical implementation demands of ADs. In addition, the lack of awareness of acceptable scope of confirmatory ADs and the associated fear of risking regulatory approval were rated slightly higher in the private sector compared to the public sector. This could partly be explained by the differences in scope of investigative interventions and associated regulatory demands governing trial conduct and approval or the commissioning process of effective interventions. The lack of funding support to aid the design work of complex ADs and worry about staff employment contracts when trials are stopped early were highly and moderately rated in the public sector compared to the private sector. This reflects differences in the organisational research funding structure. Despite a few differences, consistency in cross-sector perceptions

on barriers highlights the strong need for cross-sector collaboration to mitigate some of the commonly shared obstacles.

9.2.3 The Perspective of UK Public Funders on Roadblocks

Public Funders raised a number of obstacles, most of which are linked to the inadequate description of the rationale of the proposed AD rather than a competing mainstream design. Furthermore, the inadequate description of the type of proposed AD, decision-making criteria to guide the adaptation process, and to some extent variable costs, hinder the review process and lower the chances of success of grant applications. It is important to note that Public Funders consider a number of aspects when recommending and approving grant applications for funding. Some of the considerations include the importance of the research question(s) based on clinical priorities of the Funder, the design and scientific merits of the proposal, the quality of the research team and their experience to deliver the research, and value for money. Therefore, the obstacles raised by the Public Funders are not the only contributors to unsuccessful AD-related grant proposals. A grant proposal may address all aspects of the proposed AD but be rejected on the basis of failure to meet clinical priorities set.

There is a lack of reviewing and commissioning experience among Public Funders, mainly because of the small number of AD-related proposals being put forward for consideration. This lack of experience could be a contributing factor influencing Public Funders' preference for mainstream designs and risk averse attitude to fund ADs perceived to be associated with marked financial uncertainty. It is therefore important to encourage researchers to submit AD-related proposals for funding considerations provided that the proposed AD is appropriate to address the research question(s). The lack of capacity of Reviewers with AD-related expertise to help Funders during the review process of grant applications can only be addressed through training and capacity building. There are current initiatives by the NIHR and MRC to address the ADs skills gap through training fellowships at different levels such as MSc, PhD, and Post-Doctoral Career Development. In return, the trained fellows should contribute to the peer review process of AD-related grant applications to help the Funders in the decision-making process.

9.2.4 Paradigm Shift in Perceptions Towards Adaptive Designs

Despite the challenges discussed so far, there is a growing cross-disciplinary interest and receptiveness towards the appropriate use of ADs. The inferred change in Public Funders' attitudes and receptiveness towards

the use ADs when appropriate is driven by the desire to use efficient designs to address research questions and maximise the value for money in clinical trials research. This is supported by a number of initiatives supporting the funding of ADs-related activities: fellowships such as this PhD research, research methods grants (MRC NHTMR, 2014), outreach events (Lamb, 2014), research grant calls (NIHR HTA, 2014c), and an ADWG, which I am a member of.

There appears to be a positive will among clinical investigators to utilise ADs when appropriate. This is motivated by the desire to improve clinical trial design to address research question(s) efficiently. Nonetheless, this positive desire depends mostly on how ADs are marketed to them by Clinical Trialists and the availability of additional support for successful trial planning and conduct. The Clinical Investigators' positive desire somewhat contradicts findings from a related study that investigated the use of innovative designs in early phase trials (Jaki, 2013). Jaki concludes that Clinical Investigators insist on the application of certain methods, contradicting inferred findings of this thesis in confirmatory trials. This could partly be explained by the differences in research methods employed. For example, Jaki surveyed Statisticians and did not engage clinical investigators, thus the results may be biased against the latter. In contrast, this thesis engaged a number of key stakeholders at various stages to enhance the robustness of the findings.

The thesis inferred improving regulatory awareness and receptiveness towards ADs, and increasing numbers of AD-related proposals and approvals are reflected in the most recent related literature in the EU and USA (Elsäßer et al., 2014; FDA, 2015; Lin et al., 2015). Importantly, the improving receptiveness depends on strong caveats, which are reflected in Section 9.3.7. Detailed regulatory considerations are found in related documents (CHMP, 2007; FDA, 2010, 2015).

9.3 Recommendations for Best Practice

This section summarises key considerations, potential facilitators, and recommendations to improve the appropriate uptake of ADs in publicly funded confirmatory trials. These are applicable to particular key stakeholders in clinical trials research including Clinical Trialists and Public Funders.

9.3.1 Description of Rationale, Type and Scope of the Proposed Adaptive Design

Clinical Trialists must provide a concise explanation of the rationale for considering the proposed AD instead of competing mainstream fixed sample size designs. Furthermore, it is important to describe and highlight

the potential benefits of the proposed AD. This may include benefits in terms of patient savings, reduction in trial duration, savings in research resources, and ability to address considered research question(s) efficiently. Simulation or modelling work may help to illustrate some of the potential opportunities under wide range of scenarios. Furthermore, Clinical Trialists must provide a detailed description of the type and scope of the proposed AD to key stakeholders such as Funders and Regulators within the context of the rationale put forward. The scope relates to the design features which are intended to be modified. The inclusion of this information in trial related publications such as grant applications, protocol, and reports of the main trial results is imperative.

9.3.2 Adaptation by Design, Managed Scope and Design Properties

Clinical Trialists must use prospectively planned ADs, clearly described and documented upfront such as in the study protocol or appropriate related trial document. The intended adaptation scenarios and agreed decision-making criteria should be laid in advance at the planning stages and documented with an audit trail. The importance of doing so to maintain credibility and integrity of trial results cannot be overemphasised. In addition, it enables adequate exploration of the statistical properties of the design and influence on the decision-making process.

In line with the objectives of confirmatory trials, Clinical Trialists should minimise undertaking too many adaptations within the same trial, unless there is a strong rationale to do so. Otherwise, too many adaptations increase complexity, complicate the interpretation of findings, and impact on trial credibility. Furthermore, it is imperative to provide assurance that the statistical properties of the design are controlled. Evidence should be provided through referencing published literature. For complex ADs requiring simulation work, evidence of adequate simulations accompanied with simulation protocol, implementation code, and simulation report should be provided. Since it is impossible to cover all simulation scenarios, it is important to provide justification of the scenarios considered for simulations in an accessible simulation protocol. Discussions with the research team and Regulators, where appropriate, at the design stage may be helpful.

9.3.3 The Choice of Decision-Making Criteria

It is important to conduct some consultations with key stakeholders at the planning stage in order to draw acceptable decision-making criteria to guide the adaptation process where appropriate. This process can be informed by available literature and perceptions of key stakeholders such as Regulators, patient groups, and

clinical research groups. Such an approach is vital to enhance credibility and acceptability of the findings from an adaptive trial.

9.3.4 Suitability of the Primary Endpoints and Practical Aspects

Consideration must be given to whether the realisation of the primary endpoint data relative to the expected recruitment rate fit in with the practicalities for smooth implementation of the proposed AD. For example, research questions investigated based on immediate to short-term endpoints are the most suitable candidates for the application of ADs compared to those with long-term endpoints. For instance, if the primary endpoint is long-term (such as 2 years) and recruitment rate is too fast relative to accrual of the outcome data coupled with short intervention exposure, then by the 1st interim analysis the trial would have recruited almost all required participants already exposed to study interventions. In such circumstances, there will be little benefits to adapt the trial. However, it is important to note that ADs can still be applied in trials with long term endpoints as long as there is a clear rationale and desire to speed up the decision-making process to benefit patients outside the trial rather than saving the number of recruited trial participants. This is often the case for some severe health conditions such as in oncology. In summary, operational feasibility regarding the implementation of the proposed AD must be carefully considered at the planning stage.

9.3.5 Consideration for Key Secondary Objectives

It is important to carefully consider the impact of the proposed AD on other key secondary trial objectives where appropriate. These objectives may include the evaluation of safety, health economics, and centre effects. Such secondary objectives may influence the timing of the interim analyses or the suitability of an AD. For instance, early trial stopping based on the primary outcome may yield insufficient data to address other key secondary objectives. It is therefore imperative to balance the benefits of early stopping and the maturity of data to address key secondary objectives which are trial dependent. The delay of the 1st interim analysis as much as possible may be necessary.

9.3.6 Data Management and Information Sharing Platform

In general, ADs require improved data management systems depending on the nature of the proposed AD. This is necessitated by the need to minimise potential operational bias during the conduct of the trial.

Considerations should be given to the nature of information that should be disclosed and to whom, how the information should be transferred, who has access to what, and clarity on who is doing what. Importantly, this should be backed by audit trails to enhance trial credibility and integrity. Furthermore, turnaround time of data management processes to provide clean and robust data to inform the adaptation is paramount. This involves real-time data capturing, cleaning, and processing.

9.3.7 Appropriate Regulatory Engagement

For trials requiring regulatory approval, it is paramount for Clinical Trialists to engage Regulators through scientific advice meetings and adhere to regulatory guidance when considering appropriate ADs from trial planning to completion. Regulators appear receptive to such engagements to facilitate appropriate use of ADs (Elsäßer et al., 2014; Lin et al., 2015). These authors highlighted the following key regulatory considerations, all of which have been found by the thesis to be imperative:

- Rationale for considering an AD rather than the competing mainstream design(s) as described in Section 9.3.1;
- The appropriateness of the AD to address research question(s) in line with trial objectives;
- Documented pre-specification and adequate description of the adaptive features of the design, its scope, and decision-making criteria;
- Adequate control of the type I error;
- Consideration of the steps to minimise or avoid operational bias in the conduct of the trial;
- Feasibility considerations in the implementation of the proposed design;
- Use of appropriate statistical inference to minimise estimation bias.

Questions can be asked regarding the adequacy of the quality control framework for ADs in the public sector where the majority of trials do not require regulatory approval. The current framework in the public sector appears to depend on the ability of the scientific and ethics committees and Reviewers. It is therefore important to empower these stakeholders with adequate knowledge on ADs to enhance high quality control of adaptive trials. Otherwise, an increase in poorly designed and conducted ADs may be witnessed in coming years in the public sector.

9.3.8 Engaging Clinical Investigators

Clinical Trialists should explore ways to engage clinical investigators involved in clinical trials research to raise awareness of opportunities to use ADs. In addition, this offers an opportunity to hold discussions on AD-related challenges, limitations, when ADs are appropriate, scope of trial adaptation, and support available to those wishing to apply ADs. The use of practical case studies to illustrate these issues is paramount.

9.3.9 Addressing Funding and Support Accessible to UK CTUs

Clinical Trialists are encouraged to use ADs which are simpler and less time-consuming to implement, within the existing scope of public funding models for fixed sample size designs. Such ADs include SSR and futility analysis. A number of approaches can be adopted in the case of complex ADs. Public funders such as NIHR or MRC are encouraged to draw small grants for design developmental work of complex ADs provided that the proposed research meet the scientific merits and clinical priorities for funding. Furthermore, researchers should provide a clear rationale on why such a small design developmental grant is needed. Public funders may draw a contract requiring researchers to make design-related outputs publicly accessible, such as statistical implementation code or software in order to enhance the planning of future related trials. Alternatively, Funders may team up to fund a group of experts with practical knowledge to support CTUs with the planning and conduct of ADs. These experts should provide practical training to CTUs on AD-related issues with the aid of case studies. The ADWG of the MRC NHTMR, which I am a member of, recently initiated outreach events accessible to UK CTUs on request. The events are aimed to enhance applied practical knowledge and improve the appropriate use of ADs across all trial phases. In addition, the group is piloting an initiative to support and collaborate with CTUs willing to undertake ADs. Here, the members of the group team up with researchers as collaborators on grant proposals and provide implementation support costed in the grant application. However, their involvement still depends on the uncertain future success of the grant applications put forward.

9.3.10 Pertaining to Public Funders

There is a need for outreach activities targeting Public Funders prior to their boards or panel meetings aimed to raise awareness and improve their knowledge of AD-related issues. Such events can be provided by fellows and experts trained or funded by the Public Funders. It is important for Funders to continue with outreach events and activities to highlight their receptiveness to fund AD-related grant applications provided they are

appropriate to answer research questions. Such activities may include webinars, presentations at fellowship or grant application related events, seminars, YouTube videos, and funding calls for AD-related research. In addition, Funders should provide suggestions as part of their feedback to researchers recommending the use of ADs, when they feel a certain type of trial adaptation is appropriate. This process depends on their knowledge of ADs.

Public Funders are encouraged to develop flexible contracts similar to those for 'internal pilot' trials or programme grants that are acceptable to key collaborators to allow for ADs with early stopping options. Such contracts should incentivise those CTUs implementing ADs when appropriate rather than penalise them. Importantly, knowledge sharing among Public Funders aimed to learn lessons and address challenges raised by the use of ADs is paramount. It is likely that Public Funders have varying degrees of experience regarding the use of ADs – hence, learning from pacesetters is important.

Most of the specialised AD software or implementation codes are commercial or produced by researchers in the private sector. For example, East (Cytel, 2015) is one of the most specialised and user-friendly ADs software, however, its annual subscription is very expensive. It is therefore important to fund initiatives for the development of publicly accessible and free-to-use software. Similarly, Public Funders should make it mandatory for methodological researchers funded by them to make outputs such as implementation codes and software publicly accessible. This may also help to reduce unnecessary duplication of research and reduce waste. In addition, Public Funders are encouraged to fund research on ADs embedded within mainstream trial designs to build a knowledge base.

9.3.11 Bridging the Practical Knowledge Gap

Clinical Trialists are encouraged to publish open access case studies of 'successful' and 'unsuccessful' undertaken ADs and related materials. Such publications should include aspects addressing the rationale and design; statistical and practical challenges, and how they are resolved; implementation resources; lessons learned; regulatory, data management, and communication hurdles, and how these are resolved; among other facilitators to successful implementation. The case studies should also help to raise awareness of benefits and pitfalls of ADs, and when ADs are appropriate. Some of these case studies are starting to come though (Baraniuk et al., 2014b; Bratton et al., 2013; Brinton et al., 2015; Carreras et al., 2015; Patra et al., 2015; Pritchett et al., 2011; Sydes et al., 2009b).

There is a strong need for practical education tailored for Trialists on ADs through educational seminars, webinars, YouTube type videos, and practice-oriented workshops to facilitate translational knowledge sharing. The content of such activities should cover practical, statistical and logistical aspects that need addressing when planning and conducting adaptive trials aided with actual case studies where possible. Furthermore, as highlighted in Section 9.3.9, there is need for a focal group of practical experts publicly funded to support and partner CTUs with little practical expertise wishing to use ADs. These experts should also provide practical training accessible to CTUs. Following the dissemination of the findings of this thesis (Dimairo, Boote, Julious, Nicholl, et al., 2015), the ADWG of the MRC NHTMR (2016) has since initiated outreach events accessible to all UK CTUs. The events are aimed to raise awareness of opportunities to use ADs across trial phases and implementation resources, present case studies, and collaborate with CTUs which wish to put forward AD-related grant applications.

Researchers receiving public funding for AD-related methodological research are strongly encouraged to produce open access resources such as free-to-use software or code to facilitate the application of the methods developed. In addition, CTUs receiving AD-related bridge or research funding should form a compendium of case studies for open-access publications such as in monographs. This resource may be important in reducing research waste and improving the appropriate conduct of ADs, and would be helpful for applied knowledge transfer.

There is a strong need for a standardised, well crafted, consensus guidance document tailored for ADs in the public sector similar to the guidance for the development and evaluation of complex interventions (Craig et al., 2008, 2013). Such a guidance document may help researchers to assess the appropriate scope, benefits, statistical and practical considerations for successful application of ADs in confirmatory trials. An extension could be made to cover all trial phases aided with case studies.

There is scope for the development of a troubleshooting toolkit addressing important general and design-specific questions Clinical Trialists should ask themselves when considering the use of ADs at the planning stage. Such toolkit should cover aspects such as practical, statistical, tips for successful implementation (‘do’s and don’ts’), and implementation resources. Some design recommendations for certain types of ADs are starting to come through, which is a welcome development. For instance, Pritchett et al (2015) give statistical considerations and practical guidance for SSR in confirmatory trials. Sydes et al (2009a) articulate issues regarding the implementation of a platform MAMS design for the STAMPEDE case study. Wason et al (2013) also give some considerations and recommendations for a MAMS design.

9.3.12 Adaptive Trials Monitoring Capacity

Adaptive trials require the involvement of TSC and IDMC members with knowledge of AD-related aspects and awareness of measures to protect confidentiality during trial conduct. More so, the IDMC members must have basic understanding of the theoretical concepts underpinning the proposed AD. The training of and discussions with the IDMC members regarding the proposed AD, decision-making criteria, and communication processes and procedures are vital prior to the beginning of the trial. Furthermore, it is important to build capacity of IDMC members with AD-related knowledge. This could be achieved through training workshops, seminars, webinars, and inexperienced members shadowing IDMC meetings for ongoing adaptive trials. Those trained through public funded fellowships such as MRC and NIHR should be encouraged to undertake such training as part of professional development and capacity building.

9.3.13 Transparency and Reporting Framework of Adaptive Trials

Glasziou et al (2014) highlight the importance of replicable and reproducible science, which is enhanced through adequate reporting in order to reduce research waste . In general, the more complex the design and conduct of the trial is the greater the demands on reporting. The use of interim data to make decisions is bound to raise anxiety of some key stakeholders in clinical trials research. It is therefore, important to provide reassurance of the scientific rigour and conduct of ADs through transparent and adequate reporting. Clinical Trialists are encouraged to make AD-related trial material publicly accessible. These include protocols with related amendments, simulation protocols, simulation code and reports, open and closed IDMC minutes and interim results reports. The publication of trials reports in monographs, such as through the NIHR HTA is a welcome initiative for researchers to provide adequate details about the design and conduct of the trial. In addition, most journals are now accepting the publication of supplementary material to support the main trial results.

To optimise the potential benefits of clinical trials registers, such as ClinicalTrials.gov, in adaptive clinical trials research, it would be helpful for registers to contain a section dedicated to the type of AD and scope of the adaptation, including stopping rules, if this is a feature of the design. Clinical Trialists are encouraged to include the term ‘Adaptive Design’ in the Title or the brief summary or design section or abstract when registering and reporting adaptive trials. Better indexing of undertaken ADs is vital for easier retrieval of case studies during searches.

The access to AD case studies can only be useful if they are adequately reported. There is therefore, a strong need for a multidisciplinary and cross-sector approach to draw recommendations for an adaptive CONSORT extension statement to enhance transparent and adequate reporting of the conduct of ADs. Journals would then be encouraged to adopt such an adaptive CONSORT statement as part of their policy. Building some form of consensus through a Delphi process would be needed (Hasson et al., 2000). In addition to existing checklists for mainstream designs, the following aspects (among others) should be considered during the development of an adaptive CONSORT statement:

1) Inclusion of the term 'Adaptive Design' in the Title and/or Abstract;
2) A clear rationale on why an AD was considered rather than a competing mainstream design;
3) A clear description of the type of the AD considered and adaptive aspects of the trial;
4) A clear description of the decision-making criteria guiding the adaptation and decision-making process;
5) The inclusion of a simulation protocol, report, and statistical programs or code used for the design or to aid the decision-making process, where appropriate;
6) Description of the systems, procedures, and processes put in place to minimise the operational bias in the conduct of the trial due to the knowledge of the interim results;
7) Description and explanation of any deviations from the planned adaptation;
8) Provision of prior interim results, where appropriate. A figure showing a trend of results (point estimates and CIs) up to the interim stopping should be considered where appropriate;
9) A clear description of the statistical methods used to obtain unbiased or bias-adjusted results (point estimate, CI and p-value);
10) A discussion of lessons learned from using the considered AD to help the planning of future related trials;
11) A discussion of the generalisability of the results from the adaptive trial and to whom the results pertain to.

9.3.14 Addressing Credibility of Findings from Adaptive Trials

It is important for the consumers of research findings to be able to judge the credibility of the results of a trial in front of them. As highlighted in Section 9.3.13, although there are many factors which may influence acceptance of findings from ADs, transparent and adequate reporting is paramount. Increased publicity of ADs

using case studies of undertaken ADs, which have managed to change practice and made 'high impact' in medical practice may help improve acceptability of ADs in clinical trials research. The retrospective design and analysis of case studies whose results are known may help to illustrate lost opportunities, learn positive and negative lessons, and to provide reassurance of the robustness of ADs in decision-making.

9.3.15 Promising Adaptive Designs in the Public Sector

Clinical Trialists are encouraged to use simple ADs in confirmatory trials more often where possible. Blinded SSR to validate design assumptions is a simple and well accepted method by Regulators and the research community. The method is easily implemented by in-house Trial Statisticians and well known to have 'negligible' impact on type I error and inference when conducted with relatively large amounts of information. As a result, no adjustments to the type I error and results are necessary. Using case studies with binary outcomes, blinded SSR performed well after the enrolment of the first 300 participants in total. Since the sample sizes for confirmatory trials are often large, blinded SSR can be delayed and performed at any point after the recruitment of this total sample size to improve statistical efficiency. More retrospective analysis of case studies is needed to guide the planning of blinded SSR for trials with other outcome measures. This would add more information to the available theoretical knowledge.

The conduct of one futility analysis is something that should be considered by both Funders and Clinical Trialists whenever possible. All fixed designed trials requesting funding extensions should be subject to one futility analysis through stochastic curtailment. Public Funders are encouraged to make this mandatory as part of their policy to reduce unnecessary research waste given that a significant proportion of trials fail to recruit within the planned period. Furthermore, most investigative interventions do not meet efficacy requirements to translate into or remain in clinical practice. Importantly, this thesis found a cross-sector and multidisciplinary receptiveness towards futility analysis. Such a stochastic curtailment approach could be planned and implemented between 50% and 70% of the target recruitment. This rule does not apply to all trials since some may request funding extensions before reaching such recruitment thresholds. However, futility analysis would still help Funders in decision-making even if the method is likely to produce inconclusive results when conducted too early. In cases where evidence from a trial is required to withdraw an intervention from practice, such trials should be designed with futility and/or efficacy stopping options using GSDs where possible.

In circumstances where stopping early for futility and/or efficacy and SSR are of primary interest in a single trial, an information based GSD is an efficient approach which should be adopted by Clinical Trialists. It appears sensible to perform interim analysis within the region of 50% to 85% of the planned information fraction. This region appears to be associated with high probability of early stopping and such an information fraction may provide fair representation of participants across centres and be associated with less statistical variability. The choice of the number of interim analyses should be guided by balancing benefits and practicalities of performing an additional analysis. In practice, as found by this thesis, the number of interim analyses is commonly either one or two and rarely more than five.

Operational seamless design has the potential to speed up the evaluation and approval of investigative interventions into practice. The design is simple and does not require sophisticated statistical methods because the data from stages are analysed separately. Furthermore, findings from earlier stage(s) can be disseminated either before or during the conduct of the later stage(s). However, more lessons need to be learned on its performance through actual case studies, particularly from a public sector perspective, along the lines of private sector case studies highlighted by Cuffe et al (2014).

The MAMS design has promising potential to speed up the evaluation of multiple competing interventions in a single trial rather than in a series of multiple two arm trials. The thesis found cross-sector and multidisciplinary receptiveness towards this approach. In theory, the design appears to be efficient and could save resources and patients, and reduce the evaluation process of interventions. However, there are a lot of lessons to be learned regarding practical aspects, efficiency, and some unanswered statistical questions such as on statistical inference. In addition, there is a need for free-to-use statistical implementation resources for this design.

9.4 Main Thesis Strengths and Dissemination Achievements

Based on the best available knowledge, the thesis appears the only research exploring concerns about, barriers and facilitators to the appropriate use of confirmatory ADs in clinical trials research in the UK public sector. Importantly, it appears to be the first to use both in-depth qualitative interviews and quantitative surveys in chronological order to understand why ADs are underused in UK confirmatory clinical trials research. In-depth interviews were purposively targeted to represent key stakeholders and decision-makers in clinical trials research. Hence, the interview findings provided robust information to aid understanding of the subject and informed the design of subsequent quantitative surveys. Importantly, the thesis seems to be the only research to formally

investigate the perceptions of UK Public Funders towards the use of ADs. The surveys also utilised an efficient rating scale model to rank barriers and concerns in order of importance for prioritisation.

Equally importantly, the research helped to explore facilitators to the appropriate use of ADs from the perspective of those involved in clinical trials research and related decision-making. Although this thesis focused on publicly funded trials, a cross-sector approach used helped to provide an in-depth understanding of barriers – a platform for close collaboration between the private and public sectors. The thesis addressed some of the barriers by reviewing case studies of adaptive trials and investigating their reporting. Important lessons were learned from prospective and retrospective case studies to help the planning of future related adaptive trials.

Four reports based on the findings of Chapters 3 to 6 have been published in Trials and PLOS ONE open access peer reviewed journals (Dimairo, Boote, Julious, Nicholl, et al., 2015; Dimairo, Julious, Todd, Nicholl, et al., 2015; Hatfield et al., 2016; Stevely et al., 2015). In addition, the results have been disseminated at a number of clinical trials research related conferences: 3rd ICTMC (Dimairo, Julious, Todd and Nicholl, 2015; Dimairo, Stevely, Todd, Julious, et al., 2015); 36th Annual Meeting of the SCT (Dimairo, Stevely, Julious, Todd, et al., 2015); JSM (Julious et al., 2015); PSI (Dimairo, Todd, Julious and Nicholl, 2015); and Evidence Live 2015. Importantly, survey results of Chapter 4 were presented at the Bi-annual Statisticians Operational Group Meeting of the UK CRC registered CTUs Network held in Sheffield on the 5th October 2015 and shared. Furthermore, the results have been shared with the ADWG of the MRC NHTMR to influence the implementation of some of the proposed facilitators. The findings were also disseminated to the Funders through some Chairs and members of funding panels and boards.

9.5 Key Limitations and Interpretation of Findings

Due to the wide scope of the research, exhaustive literature review was impractical and the types of ADs considered focused on those perceived to have huge potential in confirmatory trials. As a result, complex ADs such as biomarker/population enrichment or subgroup selection, adaptive response randomisation and adding new arms to ongoing trials were not considered. Nonetheless there is recent related research focusing on these areas (Antoniou et al., 2016; Cohen et al., 2015; Renfro et al., 2016).

The moderate response rates observed from quantitative surveys limited the exploration of barriers and concerns. There is paucity of related research which surveyed Public Funders with which to compare response rates. Recent research that used a similar UK CTU sampling frame observed low to moderate response rates

ranging from 25% to 67% (Bower et al., 2014; Tudur Smith, Hickey, et al., 2014). Consequently, survey non-responders are more likely to be different to responders in some way that may influence the interpretation of results. For example, non-responders are more likely to be ADs non-enthusiasts or lack basic knowledge on ADs than responders. This may partly explain why some barriers which were more pronounced during in-depth qualitative interviews such as the lack of awareness of opportunities and acceptable scope of ADs were rated as less important during quantitative surveys. Furthermore, a significant number of respondents to the CTU and private sector surveys were designated Senior Statisticians. Hence, some results on barriers and concerns may be biased towards this responder group. For example, this may partly explain why lack of statistical expertise was among least ranked barriers.

The private sector organisations invited to take part in surveys were those who were contactable, because of confidentiality barriers. The differences in perceptions and experiences between contactable and uncontactable private organisations are unclear. However, the number of survey responders was similar to previous research in the private sector, predominantly in the USA (Morgan et al., 2014; Quinlan et al., 2010). It should be noted that the private sector survey results were complementary.

One of the raised concerns from in-depth interviews is the contrived perception by Journal Editors and Reviewers that early stopping of a trial is a failure. The related wider perceptions have not been investigated among Journal Editors and Reviewers through quantitative surveys due to time limitations. This is important to investigate and come up with remedial strategies if found to be true or demystify the fear when proved otherwise. This has been noted as an area of future research. Furthermore, this thesis did not examine the ethical implications of ADs. Recent related work publicly funded in the USA investigated ethical implications of certain types of ADs using mixed methods (Legocki et al., 2015). Further research on ethical implications of ADs is required.

Although the thesis focused on the use of ADs in the publicly funded setting in the UK in line with the interest of NIHR as the research funder, a cross-sector approach adopted facilitated the exploration of some AD-related themes in the private sector. Most of the findings on barriers appear to be consistent across-sector in the UK setting with some exceptions.

The demonstration of the application of ADs using retrospective case studies was limited by the nature of available trial data. For example, the case studies had binary outcomes. Hence, illustration of the application of ADs for continuous outcomes was not explored. In addition, the findings from retrospective case studies may be unique to the trial situation considered, thus generalisability is questionable. Nevertheless, the objective was

for illustration purposes. Finally, lessons regarding the conduct of PENNYWISE and NERVE BLOCK could not be learned because the grant proposals were unsuccessful.

9.6 Areas of Future Related Research Beyond This Thesis

This section proposes areas of future research aimed to improve practical knowledge and appropriate conduct of ADs in clinical trials research. The research propositions which are presented can be achieved in the short, medium, and long term. It is imperative to highlight the importance of collaboration across sectors with the involvement of multidisciplinary key stakeholders where possible.

The importance of accessible publication of retrospective case studies in the form of simple tutorial papers to enhance practical knowledge and to learn negative and positive lessons cannot be overemphasised. This encompasses the sharing of case studies used in Chapters 7 and 8 with other researchers. My intention is also to lead and collaborate in this area beyond the case studies and ADs utilised in this thesis. This may require collaboration with researchers within other CTUs to redesign and reanalyse a large number of retrospective case studies of completed fixed sample size design trials. Lessons learned will facilitate the design and conduct of future related adaptive trials.

The wish is to collaborate with other researchers to develop a troubleshooting toolkit on important general and design-specific questions Clinical Trialists should ask themselves when considering ADs at the planning stage. This should consider ADs which are beyond this thesis. In addition, the intention is to examine the perceptions of Journal Editors and Reviewers towards ADs and to engage leading medical journals to enhance platforms for adequate reporting of adaptive trials.

The bigger picture is to engage the MRC NHTMR ADWG and CONSORT groups to initiate a development process for an AD tailored checklist using a comprehensive Delphi process to draw recommendations based on consensus among key stakeholders to enhance transparent adequate reporting of ADs. Furthermore, there is scope to collaborate on initiatives to develop some form of standardised AD consensus guidance document tailored for the public sector addressing aspects such as the appropriate scope, statistical and practical considerations for successful implementation. These two activities are time consuming and require extensive cross-sector and multidisciplinary collaboration.

There is still more methodological work required on ADs, particularly for the more recent and complex types. For example, even though the MAMS design has attracted cross-disciplinary and cross-sector interest, there

are still unanswered methodological questions, particularly regarding inference following treatment selection. There is scope for methodological research to investigate statistical methods to obtain unbiased or bias-adjusted results. Public Funders are encouraged to consider funding of such research because the design has significant potential to improve efficiency, reduce the time required to evaluate multiple interventions, and maximise value for money in trials research. Furthermore, more resources should be dedicated to the development of user-friendly open access software or code to implement the design.

The availability of software for the application of an information based GSD is currently limited to an expensive commercial software (East) despite the efficiency appeal of the design. Limited application of the design could be because of the lack of practical knowledge and unavailability of open access or cheaper statistical software. It is therefore important to market this design and produce related open access implementation resources for various outcomes.

There is copious statistical literature on individually randomised adaptive trials. However, there is a scarcity of statistical literature to help the design of cluster randomised ADs. There is scope for further methodological research to extend statistical methods to accommodate cluster randomised adaptive trials and explore related challenges and considerations. Furthermore, the influence of population drift on the decision-making process of ADs requires further investigation.

Finally, there is a need for research exploring wider implications of ADs on trial related aspects such as the ethics and consent process, and health economic evaluation. For instance, the thesis findings on concerns regarding the impact of ADs on important secondary trial objectives informed Laura Flight's Doctoral Research Fellowship (Grant Number: DRF-2015-08-013), which has been funded by the NIHR. This project aims to explore the impact of ADs on health economics evaluation and propose recommendations.

9.7 Overall Conclusions

There is scope to utilise ADs more often in the conduct of publicly funded confirmatory trials. However, there are considerable, multifaceted individual and organisational obstacles which are hampering the appropriate use of ADs in confirmatory trials in the public sector. The lack of funding accessible to UK CTUs wishing to support developmental design work seems to be an important obstacle requiring redress. Most of the obstacles are connected to the lack of practical knowledge and hands-on experience of ADs. Cross-sector collaboration and paradigm shift towards translational applied training, and access to adequately reported case studies are important

drivers to address the dearth of knowledge and experience. The perceptions of key stakeholders on roadblocks are largely consistent across sectors, with a few exceptions reflecting differences in organisational funding structures, experiences, and the nature of study interventions and related regulatory involvement.

The degree of multidisciplinary conservatism towards ADs appears to be influenced by researchers' inadequate description of the rationale, scope, decision-making criteria, and appropriateness of the proposed AD to address the research question(s), measures to minimise operational bias, and use of appropriate statistical methods, among others. Importantly, some key facilitators have been highlighted as areas of future collaborative research to improve the appropriate use of ADs. These encompass a troubleshooting toolkit of key general and design-specific questions researchers need to ask themselves when considering ADs, a CONSORT statement to enhance transparent and adequate reporting of the conduct of adaptive trials, a multidisciplinary consensus guidance document on the acceptable scope of ADs in confirmatory trials, and retrospective case studies to learn positive and negative lessons.

Despite a number of uncovered roadblocks, there are some positives which may facilitate and improve the appropriate use of ADs. Widespread interest and UK Public Funders' positive changes in attitudes and receptiveness towards ADs when appropriate are supportive, and provide a platform for the future use of ADs in the public sector. Clinical investigators appear to have the desire to use ADs when appropriate, depending on how they are marketed to them and on the availability of implementation support. These are encouraging opportunities which should be exploited by Clinical Trialists. Furthermore, case studies on undertaken ADs are starting to come through and the hope is that this trend will continue to rise. Lastly, more lessons on ADs need to be learned and this can only be achieved if their appropriate application in routine practice is improved.

Chapter 10. References

Akobeng, A.K. (2005), "Understanding randomised controlled trials.", *Archives of Disease in Childhood*, Vol. 90 No. 8, pp. 840–4.

Allen, R., Sharma, U. and Barlas, S. (2014), "Clinical experience with desvenlafaxine in treatment of pain associated with diabetic peripheral neuropathy.", *Journal of Pain Research*, Vol. 7, pp. 339–51.

Alling, D.W. (1963), "Early Decision in the Wilcoxon Two-Sample Test", *Journal of the American Statistical Association*, Taylor & Francis, Ltd. on behalf of the American Statistical Association, Vol. 58 No. 303, pp. 713–720.

Altman, D.G. (1991), *Practical Statistics For Medical Research*, 1st ed., CHAPMAN & HALL/CRC, London,UK.

Altman, D.G., Moher, D. and Schulz, K.F. (2012), "Improving the reporting of randomised trials: the CONSORT Statement and beyond.", *Statistics in Medicine*, Vol. 31 No. 25, pp. 2985–97.

Anderson, K. (2015), "Package ' gsDesign '".

Andrich, D. (1978), "A rating formulation for ordered response categories", *Psychometrika*, Springer-Verlag, Vol. 43 No. 4, pp. 561–573.

Anscombe, F.J. (1953), "Sequential Estimation", *Journal of the Royal Statistical Society. Series B (Methodological)*, Wiley for the Royal Statistical Society, Vol. 15 No. 1, pp. 1–29.

Antoniou, M., Jorgensen, A.L. and Kolamunnage-Dona, R. (2016), "Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review.", *PloS One*, Public Library of Science, Vol. 11 No. 2, p. e0149803.

Armitage, P. (1957), "Restricted sequential procedures", *Biometrika*, Vol. 44 No. 1, pp. 9–26.

Armitage, P. (1958), "Numerical studies in the sequential estimation of a binomial parameter", *Biometrika*, Vol. 45 No. 1, pp. 1–15.

Armitage, P. (2014), "The evolution of ways of deciding when clinical trials should stop recruiting. Interview by Iain Chalmers.", *Journal of the Royal Society of Medicine*, SAGE Publications, Vol. 107 No. 1, pp. 34–9.

Armitage, P., McPherson, C. and Rowe, B. (1969), "Repeated significance tests on accumulating data", *Journal*

of the Royal Statistical …, Vol. 132 No. 2, pp. 235–244.

Arnold, D.M., Burns, K.E.A., Adhikari, N.K.J., Kho, M.E., Meade, M.O. and Cook, D.J. (2009), "The design and interpretation of pilot trials in clinical research in critical care.", *Critical Care Medicine*, Vol. 37 No. 1 Suppl, pp. S69–74.

Baraniuk, S., Tilley, B.C., del Junco, D.J., Fox, E.E., van Belle, G., Wade, C.E., Podbielski, J.M., et al. (2014a), "Pragmatic Randomized Optimal Platelet and Plasma Ratios (PROPPR) Trial: Design, rationale and implementation.", *Injury*, Elsevier, Vol. 45 No. 9, pp. 1287–95.

Baraniuk, S., Tilley, B.C., del Junco, D.J., Fox, E.E., van Belle, G., Wade, C.E., Podbielski, J.M., et al. (2014b), "Pragmatic Randomized Optimal Platelet and Plasma Ratios (PROPPR) Trial: Design, rationale and implementation.", *Injury*, Vol. 45 No. 9, pp. 1287–95.

Bassler, D., Briel, M., Montori, V.M., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansdell, D., et al. (2010), "Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis.", *JAMA*, American Medical Association, Vol. 303 No. 12, pp. 1180–7.

Bassler, D., Montori, V.M., Briel, M., Glasziou, P. and Guyatt, G. (2008), "Early stopping of randomized clinical trials for overt efficacy is problematic.", *Journal of Clinical Epidemiology*, Vol. 61 No. 3, pp. 241–6.

Bauer, P., Bretz, F., Dragalin, V., König, F. and Wassmer, G. (2015), "Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls", *Statistics in Medicine*, p. n/a–n/a.

Bauer, P. and Einfalt, J. (2006), "Application of Adaptive Designs – a Review", *Biometrical Journal*, Vol. 48 No. 4, pp. 493–506.

Bauer, P. and Kieser, M. (1999), "Combining different phases in the development of medical treatments within a single trial.", *Statistics in Medicine*, Vol. 18 No. 14, pp. 1833–48.

Bauer, P. and Kohne, K. (1994), "Evaluation of experiments with adaptive interim analyses", *Biometrics*, Vol. 50 No. 4, pp. 1029–1041.

Bechhofer, R.E., Dunnett, C.W. and Sobel, M. (1954), "A Two-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with a Common Unknown Variance", *Biometrika*, Biometrika Trust, Vol. 41 No. 1/2, pp. 170–176.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., et al. (1996), "Improving the quality

of reporting of randomized controlled trials. The CONSORT statement.", *JAMA*, Vol. 276 No. 8, pp. 637–9.

Benavente, O.R., White, C.L., Pearce, L., Pergola, P., Roldan, A., Benavente, M.-F., Coffey, C., et al. (2011), "The Secondary Prevention of Small Subcortical Strokes (SPS3) study.", *International Journal of Stroke : Official Journal of the International Stroke Society*, Vol. 6 No. 2, pp. 164–75.

Bennett, C., Khangura, S., Brehaut, J.C., Graham, I.D., Moher, D., Potter, B.K. and Grimshaw, J. (2011), "Reporting guidelines for survey research: An analysis of published guidance and reporting practices", *PLoS Medicine*, Vol. 8 No. 8, pp. 1–11.

Berry, D.A. (2012), "Adaptive clinical trials in oncology.", *Nature Reviews. Clinical Oncology*, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., Vol. 9 No. 4, pp. 199–207.

Berry, S.M., Connor, J.T. and Lewis, R.J. (2015), "The platform trial: an efficient strategy for evaluating multiple treatments.", *JAMA*, American Medical Association, Vol. 313 No. 16, pp. 1619–20.

Betensky, R. (2000), "Alternative derivations of a rule for early stopping in favor of H 0", *The American Statistician*, Vol. 54 No. 1, pp. 35–39.

Betensky, R.A. and Tierney, C. (1997), "An examination of methods for sample size recalculation during an experiment.", *Statistics in Medicine*, Vol. 16 No. 22, pp. 2587–98.

Birkett, M.A. and Day, S.J. (1994), "Internal pilot studies for estimating sample size.", *Statistics in Medicine*, Vol. 13 No. 23-24, pp. 2455–63.

Bland, J.M. and Altman, D.G. (1995), "Statistics notes: Multiple significance tests: the Bonferroni method", *BMJ*, Vol. 310 No. 6973, pp. 170–170.

Bowalekar, S. (2011), "Adaptive designs in clinical trials.", *Perspectives in Clinical Research*, Vol. 2 No. 1, pp. 23–7.

Bowden, J. and Mander, A. (2014), "A review and re-interpretation of a group-sequential approach to sample size re-estimation in two-stage trials.", *Pharmaceutical Statistics*, Vol. 13 No. 3, pp. 163–72.

Bower, P., Brueton, V., Gamble, C., Treweek, S., Smith, C.T., Young, B. and Williamson, P. (2014), "Interventions to improve recruitment and retention in clinical trials: a survey and workshop to assess current practice and future priorities.", *Trials*, Vol. 15 No. 1, p. 399.

Brannath, W., Mehta, C.R. and Posch, M. (2009), "Exact confidence bounds following adaptive group sequential tests.", *Biometrics*, Vol. 65 No. 2, pp. 539–46.

Bratton, D.J., Choodari-Oskooei, B., Phillips, P.P.J., Sydes, M.R. and Parmar, M.K.B. (2015), "Comments on 'A modest proposal for dropping poor arms in clinical trials' by Proschan and Dodd.", *Statistics in Medicine*, Vol. 34 No. 18, pp. 2678–2679.

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*, Stata Press, College Station, TX, Vol. 15 No. 2, pp. 350–368.

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*, Vol. 13, p. 139.

Braun, V. and Clarke, V. (2006), "Using thematic analysis in psychology", *Qualitative Research in Psychology*, Routledge, Vol. 3 No. 2, pp. 77–101.

Bretz, F., Koenig, F., Brannath, W., Glimm, E. and Posch, M. (2009), "Adaptive designs for confirmatory clinical trials", *Statistics in Medicine*, Vol. 28 No. 8, pp. 1181–1217.

Bretz, F., Schmidli, H., König, F., Racine, A. and Maurer, W. (2006), "Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts.", *Biometrical Journal. Biometrische Zeitschrift*, Vol. 48 No. 4, pp. 623–34.

Brinton, J.T., Ringham, B.M. and Glueck, D.H. (2015), "An internal pilot design for prospective cancer screening trials with unknown disease prevalence.", *Trials*, Vol. 16, p. 458.

Browne, R.H. (1995), "On the use of a pilot sample for sample size determination.", *Statistics in Medicine*, Vol. 14 No. 17, pp. 1933–40.

BSC. (n.d.). "Statistical Material for Download", available at: http://www.bsc.gwu.edu/bsc/webpage.php?no=6&rnd=5 (accessed 10 November 2014).

Buchbinder, S.P., Mehrotra, D. V, Duerr, A., Fitzgerald, D.W., Mogg, R., Li, D., Gilbert, P.B., et al. (2008), "Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial.", *Lancet (London, England)*, Vol. 372 No. 9653, pp. 1881–93.

Burnham, N., Quinlan, J., He, W., Marshall, M., Nicholls, G., Patel, N., Parke, T., et al. (2014), "Effective Drug Supply for Adaptive Clinical Trials: Recommendations by the DIA Adaptive Design Scientific Working Group Drug Supply Subteam", *Therapeutic Innovation & Regulatory Science*, Vol. 49 No. 1, pp. 100–107.

Burnham, N., Quinlan, J., He, W., Marshall, M., Nicholls, G., Patel, N., Parke, T., et al. (2015), "Effective Drug Supply for Adaptive Clinical Trials: Recommendations by the DIA Adaptive Design Scientific Working Group Drug Supply Subteam ", *Therapeutic Innovation & Regulatory Science* , Vol. 49 No. 1 , pp. 100–107.

Butts, C., Socinski, M. a, Mitchell, P.L., Thatcher, N., Havel, L., Krzakowski, M., Nawrocki, S., et al. (2014), "Tecemotide (L-BLP25) versus placebo after chemoradiotherapy for stage III non-small-cell lung cancer (START): a randomised, double-blind, phase 3 trial.", *The Lancet. Oncology*, Vol. 15 No. 1, pp. 59–68.

Cabana, M.D., Rand, C.S., Powe, N.R., Wu, A.W., Wilson, M.H., Abboud, P.A. and Rubin, H.R. (1999), "Why don't physicians follow clinical practice guidelines? A framework for improvement.", *JAMA*, Vol. 282 No. 15, pp. 1458–65.

Camm, C.F., Chen, Y., Sunderland, N., Nagendran, M., Maruthappu, M. and Camm, A.J. (2013), "An assessment of the reporting quality of randomised controlled trials relating to anti-arrhythmic agents (2002-2011).", *International Journal of Cardiology*, Elsevier, Vol. 168 No. 2, pp. 1393–6.

Campbell, M.K., Piaggio, G., Elbourne, D.R. and Altman, D.G. (2012), "Consort 2010 statement: extension to cluster randomised trials.", *BMJ (Clinical Research Ed.)*, Vol. 345 No. sep04_1, p. e5661.

Carreras, M., Gutjahr, G. and Brannath, W. (2015), "Adaptive seamless designs with interim treatment selection: a case study in oncology.", *Statistics in Medicine*, Vol. 34 No. 8, pp. 1317–33.

Chalmers, T.C., Levin, H., Sacks, H.S., Reitman, D., Berrier, J. and Nagalingam, R. (1987), "Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials.", *Statistics in Medicine*, Vol. 6 No. 3, pp. 315–28.

Chang, M. (1989), "Confidence intervals for a normal mean following a group sequential test", *Biometrics*, Vol. 45 No. 1, pp. 247–254.

Chang, M., Chow, S.-C. and Pong, A. (2006), "Adaptive design in clinical research: issues, opportunities, and recommendations.", *Journal of Biopharmaceutical Statistics*, Vol. 16 No. 3, pp. 299–309; discussion 311–2.

Chang, W.H. and Chuang-Stein, C. (2004), “Type I error and power in trials with one interim futility analysis”, *Pharmaceutical Statistics*, Vol. 3 No. 1, pp. 51–59.

Chapman, K.R., Rennard, S.I., Dogra, A., Owen, R., Lassen, C. and Kramer, B. (2011), “Long-term safety and efficacy of indacaterol, a long-acting β_2-agonist, in subjects with COPD: a randomized, placebo-controlled study.”, *Chest*, American College of Chest Physicians, Vol. 140 No. 1, pp. 68–75.

Charles, P., Giraudeau, B., Dechartres, A., Baron, G. and Ravaud, P. (2009), “Reporting of sample size calculation in randomised controlled trials: review.”, *BMJ (Clinical Research Ed.)*, Vol. 338 No. may12_1, p. b1732.

Charlesworth, G., Burnell, K., Hoe, J., Orrell, M. and Russell, I. (2013), “Acceptance checklist for clinical effectiveness pilot trials: a systematic approach.”, *BMC Medical Research Methodology*, Vol. 13, p. 78.

Chen, Y.H.J., DeMets, D.L. and Lan, K.K.G. (2004), “Increasing the sample size when the unblinded interim result is promising.”, *Statistics in Medicine*, Vol. 23 No. 7, pp. 1023–38.

Chen, Y.H.J., Gesser, R. and Luxembourg, A. (2015), “A seamless phase IIB/III adaptive outcome trial: design rationale and implementation challenges.”, *Clinical Trials (London, England)*, Vol. 12 No. 1, pp. 84–90.

Chen, Y.H.J., Li, C. and Lan, K.K.G. (2015a), “Sample size adjustment based on promising interim results and its application in confirmatory clinical trials”, *Clinical Trials* , available at:http://doi.org/10.1177/1740774515594378.

Chen, Y.J., Li, C. and Lan, K.G. (2015b), “Sample size adjustment based on promising interim results and its application in confirmatory clinical trials”, *Clinical Trials*, available at:http://doi.org/10.1177/1740774515594378.

Chen, Z., Zhao, Y., Cui, Y. and Kowalski, J. (2012), “Methodology and Application of Adaptive and Sequential Approaches in Contemporary Clinical Trials”, *Journal of Probability and Statistics*, Vol. 2012, pp. 1–20.

Cheng, A.-L., Kang, Y.-K., Lin, D.-Y., Park, J.-W., Kudo, M., Qin, S., Chung, H.-C., et al. (2013), “Sunitinib versus sorafenib in advanced hepatocellular cancer: results of a randomized phase III trial.”, *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, Vol. 31 No. 32, pp. 4067–75.

Chew, E.Y., Clemons, T.E., Bressler, S.B., Elman, M.J., Danis, R.P., Domalpally, A., Heier, J.S., et al. (2014), “Randomized trial of a home monitoring system for early detection of choroidal neovascularization home monitoring of the Eye (HOME) study.”, *Ophthalmology*, Vol. 121 No. 2, pp. 535–44.

Chi, L., Hung, H.M., Wang, S.J., Cui, L., Hung, H.M. and Wang, S.J. (1999), "Modification of sample size in group sequential clinical trials", *Biometrics*, Vol. 55 No. 3, pp. 853–857.

Chiang-Stein, C., Anderson, K., Gallo, P. and Collins, S. (2006), "Sample size reestimation: a review and recommendations", *Drug Information Journal*, Vol. 40, pp. 475–484.

CHMP. (2007), *Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design*.

Chow, S.-C. (2014), "Adaptive clinical trial design.", *Annual Review of Medicine*, Annual Reviews, Vol. 65, pp. 405–15.

Chow, S.-C. and Chang, M. (2008), "Adaptive design methods in clinical trials - a review.", *Orphanet Journal of Rare Diseases*, Vol. 3, p. 11.

Chow, S.-C. and Chang, M. (2011), *Adaptive Design Methods in Clinical Trials*, Second Edi., Chapman and Hall/CRC, available at:http://doi.org/doi:10.1201/b11505-1.

Chow, S.-C. and Chang, M. (2012), "Adaptive Sample Size Adjustment", *Adaptive Design Methods in Clinical Trials*, 2nd ed., CHAPMAN & HALL/CRC Press, pp. 143–165.

Chow, S.-C. and Corey, R. (2011), "Benefits, challenges and obstacles of adaptive clinical trial designs.", *Orphanet Journal of Rare Diseases*, Vol. 6, p. 79.

Chow, S.-C., Shao, J. and Wang, H. (2003), *Sample Size Calculations in Clinical Research*, CPC Press.

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: review.", *BMJ (Clinical Research Ed.)*, Vol. 346 No. mar21_1, p. f1135.

Cocks, K. and Torgerson, D.J. (2013), "Sample size calculations for pilot randomized trials: a confidence interval approach.", *Journal of Clinical Epidemiology*, Vol. 66 No. 2, pp. 197–201.

Coffey, C.S. and Kairalla, J.A. (2008), "Adaptive clinical trials: progress and challenges.", *Drugs in R&D*, Vol. 9 No. 4, pp. 229–42.

Coffey, C.S., Levin, B., Clark, C., Timmerman, C., Wittes, J., Gilbert, P. and Harris, S. (2012), "Overview, hurdles, and future work in adaptive designs: perspectives from a National Institutes of Health-funded workshop.", *Clinical Trials (London, England)*, Vol. 9 No. 6, pp. 671–80.

Cohen, D.R., Todd, S., Gregory, W.M. and Brown, J.M. (2015), "Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice.", *Trials*, Vol. 16 No. 1, p. 179.

Collier, R. (2009), "Rapidly rising clinical trial costs worry researchers.", *CMAJ : Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne*, Vol. 180 No. 3, pp. 277–8.

Collinson, F.J., Gregory, W.M., McCabe, C., Howard, H., Lowe, C., Potrata, D., Tubeuf, S., et al. (2012), "The STAR trial protocol: a randomised multi-stage phase II/III study of Sunitinib comparing temporary cessation with allowing continuation, at the time of maximal radiological response, in the first-line treatment of locally advanced/metastatic renal canc", *BMC Cancer*, Vol. 12 No. 1, p. 598.

Cook, R. and Farewell, V. (1994), "Guidelines for monitoring efficacy and toxicity responses in clinical trials", *Biometrics*, Vol. 50 No. 4, pp. 1146–1152.

Cox, D.R. (1952), "Estimation by Double Sampling", *Biometrika*, Biometrika Trust, Vol. 39 No. 3/4, pp. 217–227.

Coyne, I. (1997), "Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries?", *Journal of Advanced Nursing*, Vol. 26 No. 3, pp. 623–630.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I. and Petticrew, M. (2008), "Developing and evaluating complex interventions: the new Medical Research Council guidance.", *BMJ (Clinical Research Ed.)*, Vol. 337, p. a1655.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I. and Petticrew, M. (2013), "Developing and evaluating complex interventions: the new Medical Research Council guidance.", *International Journal of Nursing Studies*, Vol. 50 No. 5, pp. 587–92.

Creswell, J.W. (2007), *Qualitative Inquiry and Research Design: Choosing Among Five Traditions*, Second Edi., SAGE Publications, Inc.

Cuffe, R.L., Lawrence, D., Stone, A. and Vandemeulebroecke, M. (2014), "When is a seamless study desirable? Case studies from different pharmaceutical sponsors.", *Pharmaceutical Statistics*, Vol. 13 No. 4, pp. 229–37.

Cui, L., Hung, H. and Wang, S. (1999), "Modification of Sample Size in Group Sequential Clinical Trials", *Biometrics*, Vol. 55 No. 3, pp. 853–857.

Cytel. (2015), “East”, Cytel.

Dantzig, G.B. (1940), “On the Non-Existence of Tests of ‘Student’s’ Hypothesis Having Power Functions Independent of σ”, *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, Vol. 11 No. 2, pp. 186–192.

Davis, B. and Hardy, R. (1990), “Upper bounds for type I and type II error rates in conditional power calculations”, *Communications in Statistics-Theory and …*, Vol. 19 No. 10, pp. 3571–3584.

Davis, B. and Hardy, R. (1994), “Data monitoring in clinical trials: the case for stochastic curtailment”, *Journal of Clinical Epidemiology*, Vol. 47 No. 9, pp. 1033–1042.

Day, S. (2000), “Operational Difficulties with Internal Pilot Studies to Update Sample Size”, *Drug Information Journal*, Vol. 34 No. 2, pp. 461–468.

Day, S.J. and Altman, D.G. (2000), “Statistics notes: blinding in clinical trials and other studies.”, *BMJ (Clinical Research Ed.)*, Vol. 321 No. 7259, p. 504.

Demets, D. and Ware, J. (1980), “Group sequential methods for clinical trials with a one-sided hypothesis”, *Biometrika*, Vol. 67 No. 3, pp. 651–660.

Demets, D.L. (2006), “Futility approaches to interim monitoring by data monitoring committees.”, *Clinical Trials (London, England)*, Vol. 3 No. 6, pp. 522–9.

DeMets, D.L., Hardy, R., Friedman, L.M. and Lan, K.K. (1984), “Statistical aspects of early termination in the beta-blocker heart attack trial.”, *Controlled Clinical Trials*, Vol. 5 No. 4, pp. 362–72.

DeMets, D.L. and Lan, K.K. (1994), “Interim analysis: the alpha spending function approach.”, *Statistics in Medicine*, Vol. 13 No. 13-14, pp. 1341–52; discussion 1353–6.

DeMets, D.L., Pocock, S.J. and Julian, D.G. (1999), “The agonising negative trend in monitoring of clinical trials.”, *Lancet (London, England)*, Vol. 354 No. 9194, pp. 1983–8.

Denne, J.S. (2001), “Sample size recalculation using conditional power.”, *Statistics in Medicine*, Vol. 20 No. 17-18, pp. 2645–60.

Denne, J.S. and Jennison, C. (1999), “Estimating the sample size for a t-test using an internal pilot.”, *Statistics in Medicine*, Vol. 18 No. 13, pp. 1575–85.

Dent, L. and Raftery, J. (2011), “Treatment success in pragmatic randomised controlled trials: a review of trials

funded by the UK Health Technology Assessment programme.", *Trials*, BioMed Central Ltd, Vol. 12 No. 1, p. 109.

Detry, M., Lewis, R., Broglio, K. and Connor, J. (2012), *Standards for the Design, Conduct, and Evaluation of Adaptive Randomized Clinical Trials*.

Dickersin, K., Chan, S., Chalmers, T.C., Sacks, H.S. and Smith, H. (1987), "Publication bias and clinical trials.", *Controlled Clinical Trials*, Vol. 8 No. 4, pp. 343–53.

Dimairo, M., Boote, J., Julious, S.A., Nicholl, J.P. and Todd, S. (2015), "Missing steps in a staircase: a qualitative study of the perspectives of key stakeholders on the use of adaptive designs in confirmatory trials.", *Trials*, BioMed Central Ltd, Vol. 16 No. 1, p. 430.

Dimairo, M., Julious, S., Todd, S. and Nicholl, J. (2015), "Meandering journey towards routine trial adaptation: survey results on barriers to use of adaptive designs in confirmatory trials", *Trials*, BioMed Central Ltd, Vol. 16 No. Suppl 2, p. O20.

Dimairo, M., Julious, S.A., Todd, S., Nicholl, J.P. and Boote, J. (2015), "Cross-sector surveys assessing perceptions of key stakeholders towards barriers, concerns and facilitators to the appropriate use of adaptive designs in confirmatory trials.", *Trials*, BioMed Central Ltd, Vol. 16 No. 1, p. 585.

Dimairo, M., Stevely, A., Julious, S., Todd, S., Cooper, C., Hind, D. and Nicholl, J. (2015), "Differential Reporting of Group Sequential Randomised Controlled Trials (GS RCT): Shortcomings of the CONSORT 2010 Statement", *36th Annual Meeting of the Society for Clinical Trials*, p. 91.

Dimairo, M., Stevely, A., Todd, S., Julious, S., Cooper, C., Hind, D. and Nicholl, J. (2014), "Reporting issues in group sequential randomised controlled trials: a systematic review protocol of published journal reports.", Sheffield.

Dimairo, M., Stevely, A., Todd, S., Julious, S., Nicholl, J., Hind, D. and Cooper, C. (2015), "Investigation of the shortcomings of the consort 2010 statement for the reporting of group sequential randomised controlled trials", *Trials*, BioMed Central Ltd, Vol. 16 No. Suppl 2, p. O53.

Dimairo, M., Todd, S., Julious, S. and Nicholl, J. (2015), "Meandering journey towards routine trial adaptation: Survey results on the use of adaptive designs in confirmatory trials", *PSI*, London.

DiMasi, J.A., Feldman, L., Seckler, A. and Wilson, A. (2010), "Trends in risks associated with new drug development: success rates for investigational drugs.", *Clinical Pharmacology and Therapeutics*, Vol. 87

No. 3, pp. 272–7.

DiMasi, J.A., Hansen, R.W. and Grabowski, H.G. (2003), "The price of innovation: new estimates of drug development costs.", *Journal of Health Economics*, Vol. 22 No. 2, pp. 151–85.

Djulbegovic, B., Kumar, A., Miladinovic, B., Reljic, T., Galeb, S., Mhaskar, A., Mhaskar, R., et al. (2013), "Treatment success in cancer: industry compared to publicly sponsored randomized controlled trials.", *PloS One*, Vol. 8 No. 3, p. e58711.

Dmitrienko, A., Bretz, F., Westfall, P.H., Troendle, J., Wiens, B.L., Tamhane, A.C. and Hsu, J.C. (2010), "Multiple Testing Methodology", in Dmitrienko, A., Tamhane, A.C. and Bretz, F. (Eds.), *Multiple Testing in Pharmaceutical Statistics*, Chapman and Hall/CRC, Boca Raton, pp. 35–95.

Donner, A. (1984), "Approaches to sample size estimation in the design of clinical trials—a review", *Statistics in Medicine*, Vol. 3 No. 3, pp. 199–214.

Donner, B.A. and Makuch, R. (1985), "Approaches to sample size estimation in the design of clinical trials—a review", *Statistics in Medicine*, Vol. 4 No. 2, pp. 247–247.

Donohue, J.F., Fogarty, C., Lötvall, J., Mahler, D.A., Worth, H., Yorgancioglu, A., Iqbal, A., et al. (2010), "Once-daily bronchodilators for chronic obstructive pulmonary disease: indacaterol versus tiotropium.", *American Journal of Respiratory and Critical Care Medicine*, Vol. 182 No. 2, pp. 155–62.

Dragalin, V. (2006), "Adaptive Designs: Terminology and Classification", *Drug Information Journal*, Vol. 40, pp. 425–435.

Dunnett, C. (1955), "A multiple comparison procedure for comparing several treatments with a control", *Journal of the American Statistical Association*, Vol. 50 No. 272, pp. 1096–1121.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics*, Institute of Mathematical Statistics, Vol. 7 No. 1, pp. 1–26.

Egger, M., Jüni, P. and Bartlett, C. (2001), "Value of flow diagrams in reports of randomized controlled trials.", *JAMA*, Vol. 285 No. 15, pp. 1996–9.

Ellenberg, S.S., Fleming, T.R. and DeMets, D.L. (2003), "Statistical, Philosophical and Ethical Issues in Data Monitoring", *Data Monitoring Committees in Clinical Trials: A Practical Perspective*, John Wiley & Sons Ltd, Chichester, England, pp. 119–148.

von Elm, E., Röllin, A., Blümle, A., Huwiler, K., Witschi, M. and Egger, M. (2008), "Publication and non-publication of clinical trials: longitudinal study of applications submitted to a research ethics committee.", *Swiss Medical Weekly*, Vol. 138 No. 13-14, pp. 197–203.

Elsäßer, A., Regnstrom, J., Vetter, T., Koenig, F., Hemmings, R.J., Greco, M., Papaluca-Amati, M., et al. (2014), "Adaptive clinical trial designs for European marketing authorization: a survey of scientific advice letters from the European Medicines Agency.", *Trials*, Vol. 15 No. 1, p. 383.

Emerson, S. and Fleming, T. (1989), "Symmetric group sequential test designs", *Biometrics*, Vol. 45 No. 3, pp. 905–923.

Emerson, S. and Fleming, T. (1990), "Parameter estimation following group sequential hypothesis testing", *Biometrika*, Vol. 77 No. 4, pp. 875–892.

Emerson, S. and Kittelson, J. (1997), "A computationally simpler algorithm for the UMVUE of a normal mean following a group sequential trial", *Biometrics*, Vol. 53 No. 1, pp. 365–369.

Emerson, S., Kittelson, J. and Gillen, D. (2007), "Frequentist evaluation of group sequential clinical trial designs", *Statistics in Medicine*, Vol. 26, pp. 5047–5080.

Emerson, S.S. (1993), "Computation of the uniform minimum variance unbiased estimator of a normal mean following a group sequential trial.", *Computers and Biomedical Research, an International Journal*, Vol. 26 No. 1, pp. 68–73.

Emerson, S.S. (2014), "RCTdesign".

Eng, J. (2003), "Sample size estimation: how many individuals should be studied?", *Radiology*, Vol. 227 No. 2, pp. 309–13.

Englander, M. (2012), "The Interview: Data Collection in Descriptive Phenomenological Human Scientific Research*", *Journal of Phenomenological Psychology*, Brill, Vol. 43 No. 1, pp. 13–35.

EU CTR. (2004), "EU Clinical Trials Register", available at: https://www.clinicaltrialsregister.eu/ (accessed 14 May 2014).

Fairbanks, K. and Madsen, R. (1982), "P values for tests using a repeated significance test design", *Biometrika*, Vol. 69 No. 1, pp. 69–74.

Fan, X.F., DeMets, D.L. and Lan, K.K.G. (2004), "Conditional bias of point estimates following a group

sequential test.", *Journal of Biopharmaceutical Statistics*, Vol. 14 No. 2, pp. 505–530.

Farrell, B., Kenyon, S. and Shakur, H. (2010), "Managing clinical trials.", *Trials*, Vol. 11 No. 1, p. 78.

FDA. (2010), *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*.

FDA. (2013), *Good Review Practice: Clinical Review of Investigational New Drug Applications*.

FDA. (2015), *Adaptive Designs for Medical Device Clinical Studies: Draft Guidance for Industry and Food and Drug Administration Staff*.

Ferebee, B. (1983), "An unbiased estimator for the drift of a stopped Wiener proces", *Journal of Applied Probability*, Vol. 20 No. 1, pp. 94–102.

Fisher, L.D. (1996), "Comments on Bayesian and frequentist analysis and interpretation of clinical trials", *Controlled Clinical Trials*, Vol. 17 No. 5, pp. 423–434.

Fisher, L.D. (1998), "Self-designing clinical trials.", *Statistics in Medicine*, Vol. 17 No. 14, pp. 1551–62.

Fleiss, J.L., Levin, B. and Paik, M.C. (2003), "Determining Sample Sizes Needed to Detect a Difference Between Two Proportions", *Statistical Methods for Rates and Proportions*, 3rd ed., WILEY-INTERSCIENCE: John Wiley & Sons Inc, pp. 64–83.

Fleming, T., Harrington, D. and O'Brien, P. (1984), "Designs for group sequential tests", *Controlled Clinical Trials*, Vol. 5, pp. 348–361.

Fleming, T.R. (1982), "One-sample multiple testing procedure for phase II clinical trials.", *Biometrics*, Vol. 38 No. 1, pp. 143–51.

Fleming, T.R. and DeMets, D.L. (1993), "Monitoring of clinical trials: Issues and recommendations", *Controlled Clinical Trials*, Vol. 14 No. 3, pp. 183–197.

Fleming, T.R., Sharples, K., McCall, J., Moore, A., Rodgers, A. and Stewart, R. (2008), "Maintaining confidentiality of interim data to enhance trial integrity and credibility.", *Clinical Trials (London, England)*, Vol. 5 No. 2, pp. 157–67.

Follmann, D., Proschan, M. and Geller, N. (1994), "Monitoring pairwise comparisons in multi-armed clinical trials", *Biometrics*, Vol. 50 No. 2, pp. 325–336.

Fowler, V.G., Allen, K.B., Moreira, E.D., Moustafa, M., Isgro, F., Boucher, H.W., Corey, G.R., et al. (2013), "Effect of an investigational vaccine for preventing Staphylococcus aureus infections after cardiothoracic

surgery: a randomized trial.", *JAMA*, American Medical Association, Vol. 309 No. 13, pp. 1368–78.

Freidlin, B. and Korn, E.L. (2009), "Stopping clinical trials early for benefit: impact on estimation.", *Clinical Trials (London, England)*, Vol. 6 No. 2, pp. 119–25.

Freidlin, B., Korn, E.L., Gray, R. and Martin, A. (2008), "Multi-arm clinical trials of new agents: some design considerations.", *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, Vol. 14 No. 14, pp. 4368–71.

Friede, T. and Kieser, M. (2002), "On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation.", *Statistics in Medicine*, Vol. 21 No. 2, pp. 165–76.

Friede, T. and Kieser, M. (2004), "Sample size recalculation for binary data in internal pilot study designs", *Pharmaceutical Statistics*, Vol. 3 No. 4, pp. 269–279.

Friede, T. and Kieser, M. (2006), "Sample size recalculation in Internal pilot study designs: A review", *Biometrical Journal*, Vol. 48 No. 4, pp. 537–555.

Friede, T. and Miller, F. (2012), "Blinded continuous monitoring of nuisance parameters in clinical trials", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 61 No. 4, pp. 601–618.

Friedman, L.M., Furberg, C.D. and DeMets, D.L. (2010), "Monitoring Response Variables", *Fundamentals of Clinical Trials*, 4th ed., Springer, New York, pp. 293–343.

Funk, S.G., Champagne, M.T., Wiese, R.A. and Tornquist, E.M. (1991), "BARRIERS: the barriers to research utilization scale.", *Applied Nursing Research : ANR*, Vol. 4 No. 1, pp. 39–45.

Gale, N.K., Heath, G., Cameron, E., Rashid, S. and Redwood, S. (2013), "Using the framework method for the analysis of qualitative data in multi-disciplinary health research.", *BMC Medical Research Methodology*, Vol. 13 No. 1, p. 117.

Gallo, P. (2006), "Confidentiality and trial integrity issues for adaptive designs", *Drug Information Journal*, Vol. 40, pp. 445–450.

Gallo, P. (2015), "Cautions in interpretation of conditional power-based interim action thresholds", *Joint Statistical Meetings*.

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, Vol. 16 No. 3, pp. 275–83; discussion 285–91, 293–8, 311–2.

Gallo, P., Mao, L. and Shih, V.H. (2014), "Alternative Views On Setting Clinical Trial Futility Criteria", *Journal of Biopharmaceutical Statistics*, Vol. 24 No. 5, pp. 976–993.

Gao, P., Liu, L. and Mehta, C. (2013), "Exact inference for adaptive group sequential designs.", *Statistics in Medicine*, No. April, available at:http://doi.org/10.1002/sim.5847.

Gao, P., Ware, J.H. and Mehta, C. (2008), "Sample Size Re-Estimation for Adaptive Sequential Design in Clinical Trials", *Journal of Biopharmaceutical Statistics*, Vol. 18 No. 6, pp. 1184–1196.

Gates, S., Perkins, G.D., Lamb, S.E., Kelly, C., Thickett, D.R., Young, J.D., McAuley, D.F., et al. (2013), "Beta-Agonist Lung injury TrIal-2 (BALTI-2): a multicentre, randomised, double-blind, placebo-controlled trial and economic evaluation of intravenous infusion of salbutamol versus placebo in patients with acute respiratory distress syndrome.", *Health Technology Assessment (Winchester, England)*, Vol. 17 No. 38, pp. v–vi, 1–87.

Gaydos, B., Anderson, K.M., Berry, D., Burnham, N., Chuang-Stein, C., Dudinak, J., Fardipour, P., et al. (2009), "Good Practices for Adaptive Clinical Trials in Pharmaceutical Product Development", *Drug Information Journal* , Vol. 43 No. 5 , pp. 539–556.

Gaydos, B., Koch, A., Miller, F., Posch, M., Vandemeulebroecke, M. and Wang, S.-J. (2012), "Perspective on adaptive designs: 4 years European Medicines Agency reflection paper, 1 year draft US FDA guidance – where are we now?", *Clinical Investigation*, Future Science Ltd London, UK, Vol. 2 No. 3, pp. 235–240.

Gifford, W.A., Graham, I.D. and Davies, B.L. (2013), "Multi-level barriers analysis to promote guideline based nursing care: a leadership strategy from home health care.", *Journal of Nursing Management*, Vol. 21 No. 5, pp. 762–70.

Gillen, D.L. and Emerson, S.S. (2011), *Designing , Monitoring , and Analyzing Group Sequential Clinical Trials Using the "RCTdesign" Package for R*.

Glasziou, P., Altman, D.G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., Michie, S., et al. (2014), "Reducing waste from incomplete or unusable reports of biomedical research.", *Lancet*, Vol. 383 No. 9913, pp. 267–76.

Gluud, C., Dejgaard, A., Krams, M., Wallenbeck, I., Tougas, G., Wetterslev, J., Burman, C.-F., et al. (2008), "International Symposium on Adaptive Clinical Trial Designs", *Drug Information Journal* , Vol. 42 No. 1

, pp. 93–97.

Goodacre, S., Cohen, J., Bradburn, M., Gray, A., Benger, J. and Coats, T. (2013), "Intravenous or nebulised magnesium sulphate versus standard therapy for severe acute asthma (3Mg trial): a double-blind, randomised controlled trial.", *The Lancet. Respiratory Medicine*, Elsevier, Vol. 1 No. 4, pp. 293–300.

Goodacre, S., Cohen, J., Bradburn, M., Stevens, J., Gray, A., Benger, J. and Coats, T. (2014), "The 3Mg trial: a randomised controlled trial of intravenous or nebulised magnesium sulphate versus placebo in adults with acute severe asthma.", *Health Technology Assessment (Winchester, England)*, Vol. 18 No. 22, pp. 1–168.

Goodacre, S.W., Bradburn, M., Cross, E., Collinson, P., Gray, A. and Hall, A.S. (2011), "The Randomised Assessment of Treatment using Panel Assay of Cardiac Markers (RATPAC) trial: a randomised controlled trial of point-of-care cardiac markers in the emergency department.", *Heart (British Cardiac Society)*, Vol. 97 No. 3, pp. 190–6.

Gould, A.L. (1992), "Interim analyses for monitoring clinical trials that do not materially affect the type I error rate", *Statistics in Medicine*, Vol. 11 No. 1, pp. 55–66.

Gould, A.L. (1995), "Planning and revising the sample size for a trial.", *Statistics in Medicine*, Vol. 14 No. 9-10, pp. 1039–51; discussion 1053–5.

Gould, A.L. (2006), "How Practical are Adaptive Designs Likely to be for Confirmatory Trials?", *Biometrical Journal*, Vol. 48 No. 4, pp. 644–649.

Gould, A.L. and Shih, W.J. (1992), "Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance", *Communications in Statistics - Theory and Methods*, Marcel Dekker, Inc., Vol. 21 No. 10, pp. 2833–2853.

Gould, A.L. and Shih, W.J. (2005), "On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation by T. Friede and M. Kieser, Statistics in Medicine 2002; 21:165-176.", *Statistics in Medicine*, Vol. 24 No. 1, pp. 147–54; author reply154–6.

Goyder, E., Hind, D., Breckon, J., Dimairo, M., Minton, J., Everson-Hock, E., Read, S., et al. (2014), "A randomised controlled trial and cost-effectiveness evaluation of 'booster' interventions to sustain increases in physical activity in middle-aged adults in deprived urban neighbourhoods.", *Health Technology Assessment (Winchester, England)*, NIHR Journals Library, Vol. 18 No. 13, pp. 1–210.

Gray, A., Goodacre, S., Newby, D.E., Masson, M., Sampson, F. and Nicholl, J. (2008), "Noninvasive ventilation

in acute cardiogenic pulmonary edema.", *The New England Journal of Medicine*, Vol. 359 No. 2, pp. 142–51.

Gray, A.J., Goodacre, S., Newby, D.E., Masson, M.A., Sampson, F., Dixon, S., Crane, S., et al. (2009), "A multicentre randomised controlled trial of the use of continuous positive airway pressure and non-invasive positive pressure ventilation in the early treatment of patients presenting to the emergency department with severe acute cardiogenic pulmonary oe", *Health Technology Assessment (Winchester, England)*, Vol. 13 No. 33, pp. 1–106.

Hackshaw, A., Farrant, H., Bulley, S., Seckl, M.J. and Ledermann, J.A. (2008), "Setting up non-commercial clinical trials takes too long in the UK: findings from a prospective study.", *Journal of the Royal Society of Medicine*, Vol. 101 No. 6, pp. 299–304.

Hall, P. (1981), "Asymptotic Theory of Triple Sampling for Sequential Estimation of a Mean", *The Annals of Statistics*, Institute of Mathematical Statistics, Vol. 9 No. 6, pp. 1229–1238.

Halperin, M., Gordon Lan, K.K., Ware, J.H., Johnson, N.J. and DeMets, D.L. (1982), "An aid to data monitoring in long-term clinical trials", *Controlled Clinical Trials*, Vol. 3 No. 4, pp. 311–323.

Halperin, M. and Ware, J. (1974), "Early Decision in a Censored Wilcoxon Two-Sample Test for Accumulating Survival Data", *Journal of the American Statistical Association*, Taylor & Francis, Ltd. on behalf of the American Statistical Association, Vol. 69 No. 346, pp. 414–422.

Hampson, L. V. and Jennison, C. (2013), "Group sequential tests for delayed responses (with discussion)", *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, Vol. 75 No. 1, pp. 3–54.

Hasson, F., Keeney, S. and McKenna, H. (2000), "Research guidelines for the Delphi survey technique.", *Journal of Advanced Nursing*, Vol. 32 No. 4, pp. 1008–15.

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*, BioMed Central, Vol. 17 No. 1, p. 150.

Haybittle, J. (1971), "Repeated assessment of results in clinical trials of cancer treatment", *The British Journal of Radiology*, Vol. 44, pp. 793–797.

Haynes, R., Bowman, L., Rahimi, K. and Armitage, J. (2010), "How the NHS research governance procedures could be modified to greatly strengthen clinical research", *Clinical Medicine*, Royal College of Physicians, Vol. 10 No. 2, pp. 127–129.

Heritier, S., Lô, S.N. and Morgan, C.C. (2011), "An adaptive confirmatory trial with interim treatment selection: Practical experiences and unbalanced randomization", *Statistics in Medicine*, Vol. 30 No. 13, p. n/a–n/a.

Herson, J., Buyse, M. and Wittes, J.T. (2012), "On Stopping a Randomized Clinical Trial for Futility", in Kowalski, J. and Piantadosi, S. (Eds.), *Designs for Clinical Trials: Perspectives on Current Issues*, 1st ed., Springer, pp. 109–137.

Herson, J. and Wittes, J. (1993), "The Use of Interim Analysis for Sample Size Adjustment", *Drug Information Journal*, Vol. 27 No. 3, pp. 753–760.

Hertzog, M.A. (2008), "Considerations in determining sample size for pilot studies.", *Research in Nursing & Health*, Vol. 31 No. 2, pp. 180–91.

HOCHBERG, Y. (1988), "A sharper Bonferroni procedure for multiple tests of significance", *Biometrika*, Vol. 75 No. 4, pp. 800–802.

Holcomb, J.B., Tilley, B.C., Baraniuk, S., Fox, E.E., Wade, C.E., Podbielski, J.M., del Junco, D.J., et al. (2015), "Transfusion of plasma, platelets, and red blood cells in a 1:1:1 vs a 1:1:2 ratio and mortality in patients with severe trauma: the PROPPR randomized clinical trial.", *JAMA*, American Medical Association, Vol. 313 No. 5, pp. 471–82.

Holm, S. (1979), "A simple sequentially rejective multiple test procedure", *Scandinavian Journal of Statistics*, Vol. 6 No. 2, pp. 65–70.

Holmes, D.R., Kar, S., Price, M.J., Whisenant, B., Sievert, H., Doshi, S.K., Huber, K., et al. (2014), "Prospective randomized evaluation of the Watchman Left Atrial Appendage Closure device in patients with atrial fibrillation versus long-term warfarin therapy: the PREVAIL trial.", *Journal of the American College of Cardiology*, Vol. 64 No. 1, pp. 1–12.

Hommel, G. (2001), "Adaptive Modifications of Hypotheses After an Interim Analysis", *Biometrical Journal*, Vol. 43 No. 5, pp. 581–589.

Hopewell, S., Clarke, M., Stewart, L. and Tierney, J. (2007), "Time to publication for results of clinical trials.", *The Cochrane Database of Systematic Reviews*, No. 2, p. MR000011.

Hopewell, S., Loudon, K., Clarke, M.J., Oxman, A.D. and Dickersin, K. (2009), "Publication bias in clinical trials due to statistical significance or direction of trial results.", *The Cochrane Database of Systematic Reviews*, No. 1, p. MR000006.

van der Horst, C., Chasela, C., Ahmed, Y., Hoffman, I., Hosseinipour, M., Knight, R., Fiscus, S., et al. (2009), "Modifications of a large HIV prevention clinical trial to fit changing realities: a case study of the Breastfeeding, Antiretroviral, and Nutrition (BAN) protocol in Lilongwe, Malawi.", *Contemporary Clinical Trials*, Vol. 30 No. 1, pp. 24–33.

Hughes, M.D., Freedman, L.S. and Pocock, S.J. (1992), "The impact of stopping rules on heterogeneity of results in overviews of clinical trials.", *Biometrics*, Vol. 48 No. 1, pp. 41–53.

Hughes, M.D. and Pocock, S.J. (1988), "Stopping rules and estimation problems in clinical trials.", *Statistics in Medicine*, Vol. 7 No. 12, pp. 1231–42.

Hughes, S., Cuffe, R.L., Lieftucht, A. and Garrett Nichols, W."Informing the selection of futility stopping thresholds: case study from a late-phase clinical trial.", *Pharmaceutical Statistics*, Vol. 8 No. 1, pp. 25–37.

Hurst, D. (2011), "Quality of reporting randomised controlled trials in major dental journals suboptimal.", *Evidence-Based Dentistry*, Nature Publishing Group, Vol. 12 No. 2, pp. 52–3.

Hwang, I., Shih, W. and De Cani, J. (1990), "Group sequential designs using a family of type I error probability spending functions", *Statistics in Medicine*, Vol. 9 No. 12, pp. 1439–1445.

ICH. (1998), *ICH E9: Guidance on Statistical Principles for Clinical Trials*.

ICMJE. (2004), "International Committee of Medical Journal Editors (ICMJE): Uniform Requirements for Manuscripts Submitted to Biomedical Journals: writing and editing for biomedical publication.", *Haematologica*, Vol. 89 No. 3, p. 264.

ICON. (2015), "ADDPLAN", ICON.

Jaki, T. (2013), "Uptake of novel statistical methods for early-phase clinical studies in the UK public sector.", *Clinical Trials (London, England)*, Vol. 10 No. 2, pp. 344–6.

Jaki, T. and Magirr, D. (2014), "Designing Multi-Arm Multi-Stage Studies: Package 'MAMS'".

James, N.D., Sydes, M.R., Clarke, N.W., Mason, M.D., Dearnaley, D.P., Spears, M.R., Ritchie, A.W.S., et al. (2015), "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.", *Lancet (London, England)*, Elsevier, Vol. 387 No. 10024, pp. 1163–1177.

James, N.D., Sydes, M.R., Mason, M.D., Clarke, N.W., Anderson, J., Dearnaley, D.P., Dwyer, J., et al. (2012),

"Celecoxib plus hormone therapy versus hormone therapy alone for hormone-sensitive prostate cancer: first results from the STAMPEDE multiarm, multistage, randomised controlled trial.", *The Lancet. Oncology*, Vol. 13 No. 5, pp. 549–58.

Jennison, C. and Turnbull, B. (1990), "Statistical approaches to interim monitoring of medical trials: a review and commentary", *Statistical Science*, Vol. 5 No. 3, pp. 299–317.

Jennison, C. and Turnbull, B.W. (1989), "Interim Analyses: The Repeated Confidence Interval Approach", *Journal of the Royal Statistical Society. Series B (Methodological)*, Wiley, Vol. 51 No. 3, pp. 305–361.

Jennison, C. and Turnbull, B.W. (2000a), *Group Sequential Methods with Applications to Clinical Trials*, CHAPMAN & HALL/CRC.

Jennison, C. and Turnbull, B.W. (2000b), "Internal Pilot studies: Sample Size Re-estimation", *Group Sequential Methods with Applications to Clinical Trials*, CHAPMAN & HALL/CRC, pp. 279–297.

Jennison, C. and Turnbull, B.W. (2000c), "Stochastic Curtailment", *Group Sequential Methods with Applications to Clinical Trials*, CHAPMAN & HALL/CRC, pp. 205–220.

Jennison, C. and Turnbull, B.W. (2000d), "Analysis Following a Sequential Test", *Group Sequential Methods with Applications to Clinical Trials*, Second Ed., CHAPMAN & HALL/CRC, Florida, USA, pp. 171–187.

Jennison, C. and Turnbull, B.W. (2000e), "Repeated Confidence Inervals", *Group Sequential Methods with Applications to Clinical Trials*, Second Ed., CHAPMAN & HALL/CRC, Florida, USA, pp. 189–204.

Jennison, C. and Turnbull, B.W. (2000f), "Multiple Endpoints", *Group Sequential Methods with Applications to Clinical Trials*, 2nd Ed., CHAPMAN & HALL/CRC, Florida, USA, pp. 299–314.

Jennison, C. and Turnbull, B.W. (2000g), "Flexible Monitoring: The Error Spending Approach", *Group Sequential Methods with Applications to Clinical Trials*, First., CHAPMAN & HALL/CRC, Florida, USA, pp. 148–149.

Jennison, C. and Turnbull, B.W. (2011), "From Group Sequential to Adaptive Designs", *Group Sequential and Adaptive Methods for Clinical Trials*, CHAPMAN & HALL/CRC.

Jennison, C. and Turnbull, B.W. (2015), "Adaptive sample size modification in clinical trials: start small then ask for more?", *Statistics in Medicine*, available at:http://doi.org/10.1002/sim.6575.

Jensen, K. and Kieser, M. (2010), "Blinded sample size recalculation in multicentre trials with normally

distributed outcome.", *Biometrical Journal*, Vol. 52 No. 3, pp. 377–99.

Jiang, Z., Wang, L., Li, C., Xia, J. and Wang, W. (2014), "CP function: an alpha spending function based on conditional power.", *Statistics in Medicine*, Vol. 33 No. 26, pp. 4501–14.

Jitlal, M., Khan, I., Lee, S.M. and Hackshaw, A. (2012), "Stopping clinical trials early for futility: retrospective analysis of several randomised clinical studies.", *British Journal of Cancer*, Vol. 107 No. 6, pp. 910–7.

Julious, S.A. (2010), *Sample Sizes for Clinical Trials*, CPC Press.

Julious, S.A., Dimairo, M., Stevely, A. and Todd, S. (2015), "Shortcomings of the CONSORT 2010 Statement in the Reporting of Adaptive Trials", *JSM 2015*, Seattle, Washington, p. 166.

Julious, S.A. and Owen, R.J. (2006), "Sample size calculations for clinical studies allowing for uncertainty about the variance", *Pharmaceutical Statistics*, John Wiley & Sons, Ltd., Vol. 5 No. 1, pp. 29–37.

Kairalla, J. a, Coffey, C.S., Thomann, M. a and Muller, K.E. (2012), "Adaptive trial designs: a review of barriers and opportunities.", *Trials*, Trials, Vol. 13 No. 1, p. 145.

Kaplan, R., Maughan, T., Crook, A., Fisher, D., Wilson, R., Brown, L. and Parmar, M. (2013), "Evaluating many treatments and biomarkers in oncology: a new design.", *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, Vol. 31 No. 36, pp. 4562–8.

Kaplan, R.M. and Irvin, V.L. (2015), "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time.", *PloS One*, Public Library of Science, Vol. 10 No. 8, p. e0132382.

Kieser, M. and Friede, T. (2000), "Re-calculating the sample size in internal pilot study designs with control of the type I error rate.", *Statistics in Medicine*, Vol. 19 No. 7, pp. 901–911.

Kieser, M. and Friede, T. (2003), "Simple procedures for blinded sample size adjustment that do not affect the type I error rate.", *Statistics in Medicine*, Vol. 22 No. 23, pp. 3571–81.

Kieser, M. and Wassmer, G. (1996), "On the Use of the Upper Confidence Limit for the Variance from a Pilot Sample for Sample Size Determination", *Biometrical Journal*, WILEY-VCH Verlag, Vol. 38 No. 8, pp. 941–949.

Kim, K. (1989), "Point estimation following group sequential tests", *Biometrics*, Vol. 45 No. 2, pp. 613–617.

Kim, K. and Demets, D. (1987), "Design and analysis of group sequential tests based on the type I error spending rate function", *Biometrika*, Vol. 74 No. 1, pp. 149–154.

Kim, K. and DeMets, D. (1987), “Confidence intervals following group sequential tests in clinical trials”, *Biometrics*, Vol. 43 No. 4, pp. 857–864.

Kimani, P.K., Todd, S. and Stallard, N. (2013), “Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility.”, *Statistics in Medicine*, Vol. 32 No. 17, pp. 2893–910.

Kittelson, J. and Emerson, S. (1999), “A unifying family of group sequential test designs”, *Biometrics*, Vol. 55, pp. 874–882.

Klonoff, D.C., Bergenstal, R.M., Garg, S.K., Bode, B.W., Meredith, M., Slover, R.H., Ahmann, A., et al. (2013), “ASPIRE In-Home: rationale, design, and methods of a study to evaluate the safety and efficacy of automatic insulin suspension for nocturnal hypoglycemia.”, *Journal of Diabetes Science and Technology*, SAGE Publications, Vol. 7 No. 4, pp. 1005–10.

Koch, A. (2006), “Confirmatory Clinical Trials with an Adaptive Design”, *Biometrical Journal*, Vol. 48 No. 4, pp. 574–585.

Kola, I. and Landis, J. (2004), “Can the pharmaceutical industry reduce attrition rates?”, *Nature Reviews. Drug Discovery*, Nature Publishing Group, Vol. 3 No. August, pp. 711–715.

Krams, M., Burman, C.-F., Dragalin, V., Gaydos, B., Grieve, A.P., Pinheiro, J., Maurer, W., et al. (2007), “Adaptive designs in clinical drug development: opportunities, challenges, and scope reflections following PhRMA’s November 2006 workshop.”, *Journal of Biopharmaceutical Statistics*, Vol. 17 No. 6, pp. 957–64.

Lachin, J.M. (1981), “Introduction to sample size determination and power analysis for clinical trials”, *Controlled Clinical Trials*.

Lachin, J.M. (2005), “A review of methods for futility stopping based on conditional power.”, *Statistics in Medicine*, Vol. 24 No. 18, pp. 2747–64.

Lachin, J.M. (2009), “Futility interim monitoring with control of type I and II error probabilities using the interim Z-value or confidence limit.”, *Clinical Trials (London, England)*, Vol. 6 No. 6, pp. 565–73.

Lai, D., Davis, B. and Hardy, R. (2000), “Fractional Brownian motion and clinical trials”, *Journal of Applied Statistics*, Vol. 27 No. 1, pp. 103–109.

Lai, T.L., Lavori, P.W. and Shih, M.-C. (2012), “Adaptive trial designs.”, *Annual Review of Pharmacology and*

Toxicology, Vol. 52, pp. 101–10.

Lai, T.L., Lavori, P.W. and Tsang, K.W. (2015), "Adaptive design of confirmatory trials: Advances and challenges", *Contemporary Clinical Trials*, Vol. 45, pp. 93–102.

Lamb, S. (2014), "Encouraging adaptive designs in NIHR funded clinical trials", available at: http://goo.gl/OIpGq7 (accessed 4 November 2014).

Lan, K. and DeMets, D. (1983), "Discrete sequential boundaries for clinical trials", *Biometrika*, Vol. 70 No. 3, pp. 659–663.

Lan, K. and DeMets, D. (1989), "Changing frequency of interim analysis in sequential monitoring", *Biometrics*, Vol. 45 No. 3, pp. 1017–1020.

Lan, K. and Wittes, J. (1988), "The B-value: a tool for monitoring data", *Biometrics*, Vol. 44 No. 2, pp. 579–585.

Lan, K. and Zucker, D. (1993a), "Sequential monitoring of clinical trials: the role of information and Brownian motion", *Statistics in Medicine*.

Lan, K.G., Detlets, D.L. and Halperin, M. (1984), "More flexible sequential and non-sequential designs in long-term clinical trial", *Communications in Statistics - Theory and Methods*, Vol. 13 No. 19, pp. 2339–2353.

Lan, K.G., Simon, R. and Halperin, M. (1982), "Stochastically curtailed tests in long–term clinical trials", *Sequential Analysis*, No. May 2013, pp. 37–41.

Lan, K.K. and Zucker, D.M. (1993b), "Sequential monitoring of clinical trials: the role of information and Brownian motion.", *Statistics in Medicine*, Vol. 12 No. 8, pp. 753–65.

Lan, K.K.G. and DeMets, D. (2009), "Further Comments on the Alpha Spending Function", *Statistics in Biosciences*, No. April, available at:http://doi.org/10.1007/s12561-009-9005-2.

Lan, K.K.G., Lachin, J.M. and Bautista, O. (2003), "Over-ruling a group sequential boundary--a stopping rule versus a guideline.", *Statistics in Medicine*, Vol. 22 No. 21, pp. 3347–55.

Lancaster, G.A. (2015), "Pilot and feasibility studies come of age!", *Pilot and Feasibility Studies*, BioMed Central Ltd, Vol. 1 No. 1, p. 1.

Lancaster, G.A., Dodd, S. and Williamson, P.R. (2004), "Design and analysis of pilot studies: recommendations for good practice.", *Journal of Evaluation in Clinical Practice*, Vol. 10 No. 2, pp. 307–12.

Lanini, S., Zumla, A., Ioannidis, J.P.A., Caro, A. Di, Krishna, S., Gostin, L., Girardi, E., et al. (2015), "Are

adaptive randomised trials or non-randomised studies the best way to address the Ebola outbreak in west Africa?", *The Lancet. Infectious Diseases*, Vol. 15 No. 6, pp. 738–745.

Law, L.M. and Wason, J.M.S. (2014), "Design of telehealth trials--introducing adaptive approaches.", *International Journal of Medical Informatics*, Vol. 83 No. 12, pp. 870–80.

Léauté-Labrèze, C., Hoeger, P., Mazereeuw-Hautier, J., Guibaud, L., Baselga, E., Posiunas, G., Phillips, R.J., et al. (2015), "A Randomized, Controlled Trial of Oral Propranolol in Infantile Hemangioma", *New England Journal of Medicine*, Vol. 372 No. 8, pp. 735–746.

Legard, R., Keegan, J. and Ward, K. (2003), "In-depth Interviews", in Ritchie, J. Lewis, J. (Ed.), *Qualitative Research Practice: A Guide for Social Science Students and Researchers*, First Edit., SAGE, London,UK, pp. 138–169.

Legocki, L.J., Meurer, W.J., Frederiksen, S., Lewis, R.J., Durkalski, V.L., Berry, D.A., Barsan, W.G., et al. (2015), "Clinical trialist perspectives on the ethics of adaptive clinical trials: a mixed-methods analysis.", *BMC Medical Ethics*, Vol. 16 No. 1, p. 27.

Lehmacher, W. and Wassmer, G. (1999), "Adaptive sample size calculations in group sequential trials", *Biometrics*, Vol. 55 No. December, pp. 1286–1290.

Leung, D., Wang, Y. and Amar, D. (2003), "Early stopping by using stochastic curtailment in a three-arm sequential trial", *Journal of the Royal Statistical …*, Vol. 52 No. 2, pp. 139–152.

Li, Z. and DeMets, D. (1999), "On the bias of estimation of a Brownian motion drift following group sequential tests", *Statistica Sinica*, Vol. 9, pp. 923–937.

Lin, M., Lee, S., Zhen, B., Scott, J., Horne, A., Solomon, G. and Russek-Cohen, E. (2015), "CBER's Experience With Adaptive Design Clinical Trials", *Therapeutic Innovation & Regulatory Science*, p. 2168479015604181–.

Liu, A. and Hall, W.J. (1999), "Unbiased Estimation Following a Group Sequential Test", *Biometrika*, Vol. 86 No. 1, pp. 71–78.

Liu, A. and Hall, W.J. (2001), "Unbiased estimation of secondary parameters following a sequential test", *Biometrika*, Vol. 88 No. 3, pp. 895–900.

Lohr, S.L. (1990), "Accurate Multivariate Estimation Using Triple Sampling", *The Annals of Statistics*, Institute

of Mathematical Statistics, Vol. 18 No. 4, pp. 1615–1633.

Maca, J., Dragalin, V. and Gallo, P. (2014), "Adaptive Clinical Trials: Overview of Phase III Designs and Challenges", *Therapeutic Innovation & Regulatory Science*, Vol. 48 No. 1, pp. 31–40.

Magirr, D., Jaki, T. and Whitehead, J. (2012), "A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection", *Biometrika*, Vol. 99 No. 2, pp. 494–501.

Magirr, D., Stallard, N. and Jaki, T. (2014), "Flexible sequential designs for multi-arm clinical trials.", *Statistics in Medicine*, Vol. 33 No. 19, pp. 3269–79.

Mallick, A.A. and O'Callaghan, F.J.K. (2009), "Research governance delays for a multicentre non-interventional study.", *Journal of the Royal Society of Medicine*, SAGE Publications, Vol. 102 No. 5, pp. 195–8.

Markman, M., Liu, P.Y., Wilczynski, S., Monk, B., Copeland, L.J., Alvarez, R.D., Jiang, C., et al. (2003), "Phase III randomized trial of 12 versus 3 months of maintenance paclitaxel in patients with advanced ovarian cancer after complete response to platinum and paclitaxel-based chemotherapy: a Southwest Oncology Group and Gynecologic Oncology Group trial.", *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, Vol. 21 No. 13, pp. 2460–2465.

Mascia, L., Pasero, D., Slutsky, A.S., Arguis, M.J., Berardino, M., Grasso, S., Munari, M., et al. (2010), "Effect of a lung protective strategy for organ donors on eligibility and availability of lungs for transplantation: a randomized controlled trial.", *JAMA*, American Medical Association, Vol. 304 No. 23, pp. 2620–7.

Mason, M. (2010), "Sample Size and Saturation in PhD Studies Using Qualitative Interviews", *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 24 August.

Mazumdar, M. and Bang, H. (2011), "Sequential and Group Sequential Designs in Clinical Trials: Guidelines for Practitioners", in Rao, C.R., Miller, J.P. and Rao, D.C. (Eds.), *Statistical Methods for Medical Statistics*, First., Elsevier, Amsterdam, pp. 270–287.

McDonald, A.M., Knight, R.C., Campbell, M.K., Entwistle, V.A., Grant, A.M., Cook, J.A., Elbourne, D.R., et al. (2006), "What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies.", *Trials*, Vol. 7 No. 1, p. 9.

McGauran, N., Wieseler, B., Kreis, J., Schüler, Y.-B., Kölsch, H. and Kaiser, T. (2010), "Reporting bias in medical research - a narrative review.", *Trials*, Vol. 11, p. 37.

Mehta, C., Gao, P., Bhatt, D.L., Harrington, R.A., Skerjanec, S. and Ware, J.H. (2009), "Optimizing trial design: sequential, adaptive, and enrichment strategies.", *Circulation*, Vol. 119 No. 4, pp. 597–605.

Mehta, C. and Liu, L. (2016), "An objective re-evaluation of adaptive sample size re-estimation: commentary on 'Twenty-five years of confirmatory adaptive designs'", *Statistics in Medicine*, Vol. 35 No. 3, pp. 350–358.

Mehta, C.R., Bauer, P., Posch, M. and Brannath, W. (2007), "Repeated confidence intervals for adaptive group sequential trials.", *Statistics in Medicine*, Vol. 26 No. 30, pp. 5422–33.

Mehta, C.R. and Pocock, S.J. (2011), "Adaptive increase in sample size when interim results are promising: a practical guide with examples.", *Statistics in Medicine*, Vol. 30 No. 28, pp. 3267–84.

Mehta, C.R. and Tsiatis, a. a. (2001), "Flexible Sample Size Considerations Using Information-Based Interim Monitoring", *Drug Information Journal*, Vol. 35 No. 4, pp. 1095–1112.

Mehta, R.S., Barlow, W.E., Albain, K.S., Vandenberg, T. a, Dakhil, S.R., Tirumali, N.R., Lew, D.L., et al. (2012), "Combination Anastrozole and Fulvestrant in Metastatic Breast Cancer", *The New England Journal of Medicine*, Vol. 367, pp. 435–44.

Mhaskar, R., Djulbegovic, B., Magazin, A., Soares, H.P. and Kumar, A. (2012), "Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols.", *Journal of Clinical Epidemiology*, Vol. 65 No. 6, pp. 602–9.

Middleton, G., Silcocks, P., Cox, T., Valle, J., Wadsley, J., Propper, D., Coxon, F., et al. (2014), "Gemcitabine and capecitabine with or without telomerase peptide vaccine GV1001 in patients with locally advanced or metastatic pancreatic cancer (TeloVac): an open-label, randomised, phase 3 trial.", *The Lancet. Oncology*, Vol. 15 No. 8, pp. 829–40.

Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M.G., Tsiatis, A.A., Davidian, M. and Verbeke, G. (2014), "Estimation After a Group Sequential Trial", *Statistics in Biosciences*, available at:http://doi.org/10.1007/s12561-014-9112-6.

Millard, W.B. (2012), "The gold standard's flexible alloy: adaptive designs on the advance.", *Annals of Emergency Medicine*, Vol. 60 No. 2, p. 22A–27A.

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P.J., Elbourne, D., et al. (2010), "CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials.", *BMJ (Clinical Research Ed.)*, Vol. 340 No. mar23_1, p. c869.

Moher, D., Jones, A. and Lepage, L. (2001), "Use of the CONSORT Statement and Quality of Reports of Randomized Trials", *JAMA*, American Medical Association, Vol. 285 No. 15, p. 1992.

Moher, D., Liberati, A., Tetzlaff, J. and Altman, D.G. (2009), "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.", *PLoS Medicine*, Public Library of Science, Vol. 6 No. 7, p. e1000097.

Moher, D., Schulz, K.F. and Altman, D.G. (2001), "The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials.", *Lancet*, Vol. 357 No. 9263, pp. 1191–4.

Montori, V.M., Devereaux, P.J., Adhikari, N.K.J., Burns, K.E.A., Eggert, C.H., Briel, M., Lacchetti, C., et al. (2005), "Randomized trials stopped early for benefit: a systematic review.", *JAMA*, American Medical Association, Vol. 294 No. 17, pp. 2203–9.

Moore, M.J., Hamm, J., Dancey, J., Eisenberg, P.D., Dagenais, M., Fields, a, Hagan, K., et al. (2003), "Comparison of gemcitabine versus the matrix metalloproteinase inhibitor BAY 12-9566 in patients with advanced or metastatic adenocarcinoma of the pancreas: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group.", *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, Vol. 21 No. 17, pp. 3296–3302.

Morgan, C.C., Huyck, S., Jenkins, M., Chen, L., Bedding, a., Coffey, C.S., Gaydos, B., et al. (2014), "Adaptive Design: Results of 2012 Survey on Perception and Use", *Therapeutic Innovation & Regulatory Science*, Vol. 48 No. 4, pp. 473–481.

Morse, J.M. (2000), "Determining Sample Size", *Qualitative Health Research*, Vol. 10 No. 1, pp. 3–5.

Moshman, J. (1958), "A Method for Selecting the Size of the Initial Sample in Stein's Two Sample Procedure", *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, Vol. 29 No. 4, pp. 1271–1275.

MRC. (2014), "MRC NHTMR", available at: http://www.methodologyhubs.mrc.ac.uk/ (accessed 14 November 2014).

MRC NHTMR. (2014), "MRC funded PhD projects for 2015 entry", available at: http://www.methodologyhubs.mrc.ac.uk/about_us/phd2015/phd2015-projects.aspx (accessed 4 November 2014).

MRC NHTMR. (2016), "Adaptive Designs: Scope and Future Objectives", available at:

http://www.methodologyhubs.mrc.ac.uk/research/working-groups/adaptive-designs/ (accessed 11 December 2015).

Müller, H.H. and Schäfer, H. (2001), "Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches.", *Biometrics*, Vol. 57 No. 3, pp. 886–91.

Newcombe, R.G. (1998), "Two-sided confidence intervals for the single proportion: Comparison of seven methods", *Statistics in Medicine*, Vol. 17 No. May 1995, pp. 857–872.

Newton, D. and Tanner, J. (1956), "N-acetyl-para-aminophenol as an analgesic. A controlled clinical trial using the method of sequential analysis.", *BMJ*, Vol. 2, pp. 1096–9.

NIHR. (2014a), "NIHR Clinical Trials Unit (CTU) Support Funding", John Williams, available at: http://www.nets.nihr.ac.uk/programmes/ctu (accessed 4 November 2014).

NIHR. (2014b), "NIHR Evaluation, Trials and Studies Project Portfolio", available at: http://www.nets.nihr.ac.uk/projects?collection=netscc&meta_P_sand=Project (accessed 1 May 2014).

NIHR. (2015), *NIHR Annual Report 2013/14*, NIHR.

NIHR. (n.d.). "Feasibility and pilot studies", available at: http://www.nihr.ac.uk/CCF/RfPB/FAQs/Feasibility_and_pilot_studies.pdf (accessed 4 November 2014).

NIHR CRN. (2015), "The Clinical Research Network: delivering research to make patients, and the NHS, better", available at: http://www.crn.nihr.ac.uk/ (accessed 4 November 2014).

NIHR HTA. (2014a), "Health Technology Assessment (HTA) Programme", available at: http://www.nets.nihr.ac.uk/programmes/hta (accessed 20 August 2014).

NIHR HTA. (2014b), "NIHR HTA Programme: Our people", available at: http://www.nets.nihr.ac.uk/programmes/hta/our-people (accessed 20 August 2014).

NIHR HTA. (2014c), "Funding for Primary Research Using Efficient Study Designs to Evaluate Clinical and Public Health Interventions for the NHS", available at: http://www.wales.nhs.uk/sites3/documents/970/NIHR HTA Researcher Led (efficient study) Specification.pdf (accessed 4 November 2014).

O'Brien, P. and Fleming, T. (1979a), "A multiple testing procedure for clinical trials", *Biometrics*, Vol. 35 No. 3,

pp. 549–556.

O'Brien, P.C. and Fleming, T.R. (1979b), "A multiple testing procedure for clinical trials.", *Biometrics*, Vol. 35 No. 3, pp. 549–56.

O'Neill, R.T. (2006), "FDA's critical path initiative: a perspective on contributions of biostatistics.", *Biometrical Journal. Biometrische Zeitschrift*, Vol. 48 No. 4, pp. 559–64.

O'Reilly, M. and Parker, N. (2013), "'Unsatisfactory Saturation': a critical exploration of the notion of saturated sample sizes in qualitative research", *Qualitative Research* , Vol. 13 No. 2 , pp. 190–197.

Pampallona, S. and Tsiatis, A. a. (1994), "Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis", *Journal of Statistical Planning and Inference*, Vol. 42 No. 1-2, pp. 19–35.

Pampallona, S., Tsiatis, A.A. and Kim, K. (2001), "Interim Monitoring of Group Sequential Trials Using Spending Functions for the Type I and Type II Error Probabilities", *Drug Information Journal*, Vol. 35 No. 4, pp. 1113–1121.

Parmar, M.K.B., Barthel, F.M.-S., Sydes, M., Langley, R., Kaplan, R., Eisenhauer, E., Brady, M., et al. (2008), "Speeding up the evaluation of new agents in cancer.", *Journal of the National Cancer Institute*, Vol. 100 No. 17, pp. 1204–14.

Parmar, M.K.B., Carpenter, J. and Sydes, M.R. (2014), "More multiarm randomised trials of superiority are needed.", *Lancet*, Vol. 384 No. 9940, pp. 283–4.

Parsons, N., Friede, T., Todd, S. and Stallard, N. (2011), "Software tools for implementing simulation studies in adaptive seamless designs: introducing R package ASD", *Trials*, Vol. 12 No. Suppl 1, p. A8.

Patra, K., Cree, B.A.C., Katz, E., Pulkstenis, E., Dmitrienko, A. and Cutter, G. (2015), "Statistical Considerations for an Adaptive Design for a Serious Rare Disease", *Drug Information Journal*, p. 2168479015619203–.

Pearce, W., Raman, S. and Turner, A. (2015), "Randomised trials in context: practical problems and social aspects of evidence-based medicine and policy.", *Trials*, Vol. 16 No. 1, p. 394.

Pepe, M. and Anderson, G. (1992), "Two-stage experimental designs: early stopping with a negative result", *Applied Statistics*, Vol. 41 No. 1, pp. 181–190.

Peto, R., Pike, M. and Armitage, P. (1977), "Design and analysis of randomized clinical trials requiring prolonged

observation of each patient. II. analysis and examples.", *British Journal of …*.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S. V, Mantel, N., et al. (1976), "Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design.", *British Journal of Cancer*, Vol. 34 No. 6, pp. 585–612.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S. V, Mantel, N., et al. (1977), "Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples.", *British Journal of Cancer*, Vol. 35 No. 1, pp. 1–39.

Piaggio, G., Elbourne, D.R., Pocock, S.J., Evans, S.J.W. and Altman, D.G. (2012), "Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement.", *JAMA*, Vol. 308 No. 24, pp. 2594–604.

Pinheiro, J. (1997), "Estimating and reducing bias in group sequential designs with Gaussian independent increment structure", *Biometrika*, Vol. 84 No. 4, pp. 831–845.

Pocock, S.J. (1977), "Group Sequential Methods in the Design and Analysis of Clinical Trials", *Biometrika*, Vol. 64 No. 2, p. 191.

Pocock, S.J. (1983), "Monitoring Trial Progress", *Clinical Trials: A Practical Approach*, John Wiley & Sons Ltd, Chichester, England, pp. 142–159.

Pocock, S.J. (2006), "Current controversies in data monitoring for clinical trials", *Clinical Trials*, Vol. 3 No. 6, pp. 513–521.

Pocock, S.J. and Hughes, M.D. (1989), "Practical problems in interim analyses, with particular regard to estimation.", *Controlled Clinical Trials*, Vol. 10 No. 4 Suppl, p. 209S–221S.

Pope, C., Ziebland, S. and Mays, N. (2000), "Qualitative research in health care. Analysing qualitative data.", *BMJ (Clinical Research Ed.)*, Vol. 320 No. 7227, pp. 114–6.

Pope, C., Ziebland, S. and Mays, N. (2006), "Analysing Qualitative Data", in Pope, C. Mays, N. (Ed.), *Qualitative Research in Health Care*, Third Edit., Blackwell Publishing Ltd, Oxford, UK, pp. 63–81.

Posch, M. and Bauer, P. (1999), "Adaptive Two Stage Designs and the Conditional Error Function", *Biometrical Journal*, Vol. 41 No. 6, pp. 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*, Vol. 24 No. 24, pp. 3697–714.

Pritchett, Y., Jemiai, Y., Chang, Y., Bhan, I., Agarwal, R., Zoccali, C., Wanner, C., et al. (2011), "The use of group sequential, information-based sample size re-estimation in the design of the PRIMO study of chronic kidney disease.", *Clinical Trials (London, England)*, Vol. 8 No. 2, pp. 165–74.

Pritchett, Y.L., Menon, S., Marchenko, O., Antonijevic, Z., Miller, E., Sanchez-Kam, M., Morgan-Bouniol, C.C., et al. (2015), "Sample Size Re-estimation Designs In Confirmatory Clinical Trials—Current State, Statistical Considerations, and Practical Guidance", *Statistics in Biopharmaceutical Research*, Taylor & Francis, Vol. 7 No. 4, pp. 309–321.

Proschan, M. a and Dodd, L.E. (2014), "A modest proposal for dropping poor arms in clinical trials.", *Statistics in Medicine*, Vol. 33 No. 19, pp. 3241–52.

Proschan, M. a. (2009), "Sample size re-estimation in clinical trials", *Biometrical Journal*, Vol. 51 No. 2, pp. 348–357.

Proschan, M. a., Follmann, D. a. and Geller, N.L. (1994), "Monitoring multi-armed trials", *Statistics in Medicine*, Vol. 13 No. 13-14, pp. 1441–1452.

Proschan, M., Lan, K. and Wittes, J. (2006a), "Adaptive Sample Size Methods", *Statistical Monitoring of Clinical Trials - A Unified Approach*, 2nd ed., Springer, pp. 185–211.

Proschan, M., Lan, K. and Wittes, J. (2006b), "Historical Monitoring Boundaries", *Statistical Monitoring of Clinical Trials - A Unified Approach*, Springer, New York, USA, pp. 67–79.

Proschan, M., Lan, K.K.G. and Wittes, J.T. (2006c), "Power: Conditional, Unconditional, and Predictive", *Statistical Monitoring of Clinical Trials - A Unified Approach*, 1st ed., Springer, New York, pp. 43–66.

Proschan, M.A. (2005), "Two-stage sample size re-estimation based on a nuisance parameter: a review.", *Journal of Biopharmaceutical Statistics*, Vol. 15 No. 4, pp. 559–74.

Proschan, M.A. and Hunsberger, S.A. (1995), "Designed extension of studies based on conditional power.", *Biometrics*, Vol. 51 No. 4, pp. 1315–24.

Proschan, M.A., Lan, K.K.G. and Wittes, J.T. (2006d), "Inference Following a Group-Sequantial Trial", *Statistical Monitoring of Clinical Trials - A Unified Approach*, Springer, New York, USA, pp. 113–135.

Proschan, M.A. and Wittes, J. (2000), "An improved double sampling procedure based on the variance.", *Biometrics*, Vol. 56 No. 4, pp. 1183–1187.

QRS International. (2014), "NVivo: Qualitative Data Analysis Software", available at: http://www.qsrinternational.com/products_nvivo.aspx (accessed 8 November 2014).

Quinlan, J., Gaydos, B., Maca, J. and Krams, M. (2010), "Barriers and opportunities for implementation of adaptive designs in pharmaceutical product development", *Clinical Trials*, Vol. 7 No. 2, pp. 167–173.

Quinlan, J. and Krams, M. (2006), "Implementing adaptive designs: logistical and operational considerations", *Drug Information Journal*, Vol. 40, pp. 437–444.

Renfro, L.A., An, M.-W. and Mandrekar, S.J. (2016), "Precision oncology: A new era of cancer clinical trials", *Cancer Letters*, available at:http://doi.org/10.1016/j.canlet.2016.03.015.

Ritchie, J. and Lewis, J. (Eds.). (2003), *QUALITATIVE RESEARCH PRACTICE: A Guide for Social Sience Students and Researcher*, First Edit., SAGE Publications, Inc.

Robertson, J. and Armitage, P. (1959), "Report of a clinical trial to compare two hypotensive agents", *Anaesthesia*, Vol. 14, pp. 53–64.

Rogers, C. a, Pike, K., Campbell, H., Reeves, B.C., Angelini, G.D., Gray, A., Altman, D.G., et al. (2014), "Coronary artery bypass grafting in high-RISk patients randomised to off- or on-Pump surgery: a randomised controlled trial (the CRISP trial).", *Health Technology Assessment (Winchester, England)*, Vol. 18 No. 44, pp. v–xx, 1–157.

Rosner, G.L. and Tsiatis, A. a. (1988), "Exact Confidence Intervals Following a Group Sequential Trial: A Comparison of Methods", *Biometrika*, Vol. 75 No. 4, p. 723.

Royston, P., Barthel, F.M.-S., 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*, BioMed Central, Vol. 12 No. 1, p. 81.

Rudser, K.D. and Emerson, S.S. (2008), "Implementing type I & type II error spending for two-sided group sequential designs.", *Contemporary Clinical Trials*, Vol. 29 No. 3, pp. 351–8.

RUMM Laboratory Pty Ltd. (2014), "RUMM2030".

SAACTD Workshop Committee. (2009), "Connecting Non-Profits to Adaptive Clinical Trial Designs : Themes

and Recommendations from the Scientific Advances in Adaptive Clinical Trial Designs Workshop", available at: https://custom.cvent.com/536726184EFD40129EF286585E55929F/files/2627e73646ce4733a2c03692fab26fff.pdf (accessed 3 October 2014).

Sandvik, L., Erikssen, J., Mowinckel, P. and Rødland, E.A. (1996), "A method for determining the size of internal pilot studies.", *Statistics in Medicine*, Vol. 15 No. 14, pp. 1587–90.

Schäfer, H. (2006), "Adaptive designs from the viewpoint of an academic biostatistician.", *Biometrical Journal. Biometrische Zeitschrift*, Vol. 48 No. 4, pp. 586–90; discussion 613–22.

Schäfer, H., Timmesfeld, N. and Müller, H.-H. (2006), "An Overview of Statistical Approaches for Adaptive Designs and Design Modifications", *Biometrical Journal*, WILEY-VCH Verlag, Vol. 48 No. 4, pp. 507–520.

Scharfstein, D.O. and Tsiatis, A.A. (1998), "The use of simulation and bootstrap in information-based group sequential studies.", *Statistics in Medicine*, Wiley Subscription Services, Inc., A Wiley Company, Vol. 17 No. 1, pp. 75–87.

Schlosser, R.W., Wendt, O., Bhavnani, S. and Nail-Chiwetalu, B. (2006), "Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive Pearl Growing. A review.", *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, Vol. 41 No. 5, pp. 567–82.

Schulz, K.F., Altman, D.G. and Moher, D. (2010), "CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials.", *Annals of Internal Medicine*, Vol. 152 No. 11, pp. 726–32.

Schulz, K.F. and Grimes, D.A. (2002), "Blinding in randomised trials: hiding who got what.", *Lancet (London, England)*, Vol. 359 No. 9307, pp. 696–700.

Schulz, K.F. and Grimes, D.A. (2005), "Sample size calculations in randomised trials: mandatory and mystical.", *Lancet*, Vol. 365 No. 9467, pp. 1348–53.

Sebille, V. and Bellissant, E. (2000), "Comparison of four sequential methods allowing for early stopping of comparative clinical trials", *Clinical Science*, Vol. 578, pp. 569–578.

Sebille, V. and Bellissant, E. (2003), "Sequential methods and group sequential designs for comparative clinical trials", *Fundamental and Clinical Pharmacology*, Vol. 17 No. 5, pp. 505–516.

Seelbinder, B.M. (1953), "On Stein's Two-stage Sampling Scheme", *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, Vol. 24 No. 4, pp. 640–649.

Shanyinde, M., Pickering, R.M. and Weatherall, M. (2011), "Questions asked and answered in pilot and feasibility randomized controlled trials.", *BMC Medical Research Methodology*, Vol. 11, p. 117.

Shih, W.J. and Zhao, P.L. (1997), "Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes.", *Statistics in Medicine*, John Wiley & Sons, Ltd., Vol. 16 No. 17, pp. 1913–23.

Siegmund, D. (1978), "Estimation following sequential tests", *Biometrika*, Vol. 65 No. 2, pp. 341–349.

Sim, J. and Lewis, M. (2012), "The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency.", *Journal of Clinical Epidemiology*, Vol. 65 No. 3, pp. 301–8.

Singer, J. (1999), "A method for determining the size of internal pilot studies by L. Sandvik, J. Erikssen, P. Mowinckel and E. A. Rødland, Statistics in Medicine, 15, 1587–1590 (1996)", *Statistics in Medicine*, John Wiley & Sons, Ltd., Vol. 18 No. 9, pp. 1151–1153.

Sjögren, P. and Halling, A. (2002), "Quality of reporting randomised clinical trials in dental and medical research.", *British Dental Journal*, Vol. 192 No. 2, pp. 100–3.

Slud, E. and Wei, L.J. (1982), "Two-Sample Repeated Significance Tests Based on the Modified Wilcoxon Statistic", *Journal of the American Statistical Association*, Taylor & Francis, Ltd. on behalf of the American Statistical Association, Vol. 77 No. 380, pp. 862–868.

Smajdor, A., Sydes, M.R., Gelling, L. and Wilkinson, M. (2009), "Applying for ethical approval for research in the United Kingdom.", *BMJ (Clinical Research Ed.)*, BMJ Publishing Group, Vol. 339, p. b4013.

Smith, J. and Firth, J. (2011), "Qualitative data analysis: the framework approach.", *Nurse Researcher*, Vol. 18 No. 2, pp. 52–62.

Snell, E. and Armitage, P. (1957), "Clinical comparison of diamorphine and pholcodine as cough suppressants by a new method of sequential analysis", *Lancet*, Vol. 1, pp. 860–862.

Song, F., Parekh, S., Hooper, L., Loke, Y.K., Ryder, J., Sutton, A.J., Hing, C., et al. (2010), "Dissemination and publication of research findings: an updated review of related biases.", *Health Technology Assessment (Winchester, England)*, Vol. 14 No. 8, pp. iii, ix–xi, 1–193.

Spiegelhalter, D.J., Freedman, L.S. and Blackburn, P.R. (1986), "Monitoring clinical trials: Conditional or predictive power?", *Controlled Clinical Trials*, Vol. 7 No. 1, pp. 8–17.

Stallard, N., Hamborg, T., Parsons, N. and Friede, T. (2014), "Adaptive designs for confirmatory clinical trials with subgroup selection.", *Journal of Biopharmaceutical Statistics*, Vol. 24 No. 1, pp. 168–87.

Stallard, N. and Todd, S. (2003), "Sequential designs for phase III clinical trials incorporating treatment selection.", *Statistics in Medicine*, Vol. 22 No. 5, pp. 689–703.

Stallard, N. and Todd, S. (2011), "Seamless phase II/III designs.", *Statistical Methods in Medical Research*, Vol. 20 No. 6, pp. 623–34.

StataCorp. (2014), "Stata", StataCorp, College Station, Texas 77845 USA.

Stein, C. (1945), "A Two-Sample Test for a Linear Hypothesis Whose Power is Independent of the Variance", *The Annals of Mathematical Statistics*, The Institute of Mathematical Statistics, Vol. 16 No. 3, pp. 243–258.

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", edited by Shamji, M.*PLOS ONE*, Public Library of Science, Vol. 10 No. 11, p. e0141104.

Stovold, E., Beecher, D., Foxlee, R. and Noel-Storr, A. (2014), "Study flow diagrams in Cochrane systematic review updates: an adapted PRISMA flow diagram.", *Systematic Reviews*, BioMed Central Ltd, Vol. 3 No. 1, p. 54.

Sugano, K., Choi, M.-G., Lin, J.-T., Goto, S., Okada, Y., Kinoshita, Y., Miwa, H., et al. (2014), "Multinational, double-blind, randomised, placebo-controlled, prospective study of esomeprazole in the prevention of recurrent peptic ulcer in low-dose acetylsalicylic acid users: the LAVENDER study.", *Gut*, BMJ Publishing Group Ltd and British Society of Gastroenterology, Vol. 63 No. 7, pp. 1061–8.

Sully, B.G.O., Julious, S.A. and Nicholl, J. (2013), "A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies.", *Trials*, Vol. 14 No. 1, p. 166.

Sully, B.G.O., Julious, S.A. and Nicholl, J. (2014), "An investigation of the impact of futility analysis in publicly funded trials.", *Trials*, Vol. 15, p. 61.

SurveyMonkey. (2014), "SurveyMonkey webpage", available at: https://www.surveymonkey.com/ (accessed 2

April 2014).

Sutton, L., Julious, S.A. and Goodacre, S.W. (2012), "Influence of adaptive analysis on unnecessary patient recruitment: reanalysis of the RATPAC trial.", *Annals of Emergency Medicine*, Vol. 60 No. 4, pp. 442–8.e1.

Sydes, M.R., Parmar, M.K.B., James, N.D., Clarke, N.W., Dearnaley, D.P., Mason, M.D., Morgan, R.C., et al. (2009a), "Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial.", *Trials*, Vol. 10, p. 39.

Sydes, M.R., Parmar, M.K.B., James, N.D., Clarke, N.W., Dearnaley, D.P., Mason, M.D., Morgan, R.C., et al. (2009b), "Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial.", *Trials*, Vol. 10, p. 39.

Sydes, M.R., Parmar, M.K.B., Mason, M.D., Clarke, N.W., Amos, C., Anderson, J., de Bono, J., et al. (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*, Trials, Vol. 13 No. 1, p. 1.

Teare, M.D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A. and Walters, S.J. (2014), "Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study.", *Trials*, Vol. 15 No. 1, p. 264.

Temple, R. (2000), "Current definitions of phases of investigation and the role of the FDA in the conduct of clinical trials.", *American Heart Journal*, Vol. 139 No. 4, pp. S133–5.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L.P., Robson, R., et al. (2010), "A tutorial on pilot studies: the what, why and how.", *BMC Medical Research Methodology*, Vol. 10 No. 1, p. 1.

Thadhani, R., Appelbaum, E., Pritchett, Y., Chang, Y., Wenger, J., Tamez, H., Bhan, I., et al. (2012), "Vitamin D therapy and cardiac structure and function in patients with chronic kidney disease: the PRIMO randomized controlled trial.", *JAMA*, American Medical Association, Vol. 307 No. 7, pp. 674–84.

Thadhani, R., Wenger, J., Tamez, H., Cannata, J., Thompson, B.T., Andress, D., Manning, W.J., et al. (2012), "Vitamin D Therapy and Cardiac Structure and Function in Patients With Chronic Kidney Disease", Vol. 02114.

Thall, P.F., Simon, R. and Ellenberg, S.S. (1988), "Two-Stage Selection and Testing Designs for Comparative Clinical Trials", *Biometrika*, Biometrika Trust, Vol. 75 No. 2, pp. 303–310.

The EQUATOR Network. (2006), “Enhancing the QUAlity and Transparency Of health Research: Reporting guidelines under development”, available at: http://www.equator-network.org/library/reporting-guidelines-under-development/ (accessed 23 September 2014).

Thompson, W.R. (1933), “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples”, *Biometrika*, Biometrika Trust, Vol. 25 No. 3/4, pp. 285–294.

Todd, S. (2003), “An adaptive approach to implementing bivariate group sequential clinical trial designs.”, *Journal of Biopharmaceutical Statistics*, Vol. 13 No. 4, pp. 605–19.

Todd, S. (2007), “A 25-year review of sequential methodology in clinical studies.”, *Statistics in Medicine*, England, Vol. 26 No. 2, pp. 237–252.

Todd, S., Whitehead, A., Stallard, N. and Whitehead, J. (2001), “Interim analyses and sequential designs in phase III studies.”, *British Journal of Clinical Pharmacology*, Vol. 51 No. 5, pp. 394–9.

Todd, S., Whitehead, J. and Facey, K. (1996), “Point and interval estimation following a sequential clinical trial”, *Biometrika*, Vol. 83 No. 2, pp. 453–461.

Tong, A., Sainsbury, P. and Craig, J. (2007), “Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups”, *International Journal for Quality in Health Care*, Vol. 19 No. 6, pp. 349–357.

Troendle, J.F. and Yu, K.F. (2010), “Conditional estimation following a group sequential clinical trial”, *Communications in Statistics - Theory and Methods*, Marcel Dekker, Inc., Vol. 28 No. 7, pp. 1617–1634.

Tröger, W., Galun, D., Reif, M., Schumann, A., Stanković, N. and Milićević, M. (2013), “Viscum album [L.] extract therapy in patients with locally advanced or metastatic pancreatic cancer: a randomised clinical trial on overall survival.”, *European Journal of Cancer (Oxford, England : 1990)*, Vol. 49 No. 18, pp. 3788–97.

Tsiatis, A., Rosner, G. and Mehta, C. (1984), “Exact confidence intervals following a group sequential test”, *Biometrics*, Vol. 40 No. 3, pp. 797–803.

Tudur Smith, C., Hickey, H., Clarke, M., Blazeby, J. and Williamson, P. (2014), “The trials methodological research agenda: results from a priority setting exercise.”, *Trials*, Vol. 15, p. 32.

Tudur Smith, C., Williamson, P.R. and Beresford, M.W. (2014), “Methodology of clinical trials for rare diseases.”, *Best Practice & Research. Clinical Rheumatology*, Vol. 28 No. 2, pp. 247–62.

Turner, L., Shamseer, L., Altman, D.G., Schulz, K.F. and Moher, D. (2012), "Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review.", *Systematic Reviews*, Vol. 1, p. 60.

U.S. NIH. (1997), "ClinicalTrials.gov A service of the U.S. National Institutes of Health", available at: https://clinicaltrials.gov/ (accessed 1 June 2014).

UK CRC. (2014), "UKCRC Registered Clinical Trials Units Network", available at: http://www.ukcrc-ctu.org.uk/ (accessed 14 November 2014).

Vagias, W.M. (2006), "Likert-type scale response anchors", *Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management, Clemson University*, Vol. 257 No. 6, pp. 50–58.

Vandemeulebroecke, M. (2008), "Group sequential and adaptive designs - a review of basic concepts and points of discussion.", *Biometrical Journal. Biometrische Zeitschrift*, Vol. 50 No. 4, pp. 541–57.

Vandemeulebroecke, M. (2014), "Package ' adaptTest ': Adaptive two-stage tests Version".

Vickers, A.J. (2003), "Underpowering in randomized trials reporting a sample size calculation", *Journal of Clinical Epidemiology*, Vol. 56 No. 8, pp. 717–720.

Waksman, J.A. (2007), "Assessment of the Gould-Shih procedure for sample size re-estimation.", *Pharmaceutical Statistics*, Vol. 6 No. 1, pp. 53–65.

Wang, S. and Tsiatis, A. (1987), "Approximately optimal one-parameter boundaries for group sequential trials", *Biometrics*, Vol. 43 No. 1, pp. 193–199.

Wang, S.-J. (2010), "Editorial: Adaptive designs: appealing in development of therapeutics, and where do controversies lie?", *Journal of Biopharmaceutical Statistics*, Vol. 20 No. 6, pp. 1083–7.

Wang, Y. and Leung, D. (1997), "Bias reduction via resampling for estimation following sequential tests", *Sequential Analysis*, Vol. 16 No. 3, pp. 249–267.

Ware, J.H., Muller, J.E. and Braunwald, E. (1985), "The futility index. An approach to the cost-effective termination of randomized clinical trials", *The American Journal of Medicine*, Vol. 78 No. 4, pp. 635–643.

Warschkow, R., Tarantino, I., Jensen, K., Beutner, U., Clerici, T., Schmied, B.M. and Steffen, T. (2012), "Bilateral superficial cervical plexus block in combination with general anesthesia has a low efficacy in thyroid

surgery: a meta-analysis of randomized controlled trials.", *Thyroid : Official Journal of the American Thyroid Association*, Vol. 22 No. 1, pp. 44–52.

Wason, J., Magirr, D., Law, M. and Jaki, T. (2013), "Some recommendations for multi-arm multi-stage trials.", *Statistical Methods in Medical Research*, p. 0962280212465498–.

Wason, J.M.S. (2015), "OptGS : An R Package for Finding Near-Optimal Group-Sequential Designs", *Journal of Statistical Software*, Vol. 66 No. 2, pp. 1–13.

Wason, J.M.S. and Jaki, T. (2012), "Optimal design of multi-arm multi-stage trials.", *Statistics in Medicine*, Vol. 31 No. 30, pp. 4269–79.

Wassmer, G. (1999), "Multistage Adaptive Test Procedures Based on Fisher's Product Criterion", *Biometrical Journal*, Vol. 41 No. 3, pp. 279–293.

Watkinson, G. (1958), "Treatment of ulcerative colitis with topical hydrocortisone hemisuccinate sodium. A controlled trial employing restricted sequential analysis", *BMJ*, Vol. 2, pp. 1077–82.

Wears, R.L. (2015), "Are We There Yet? Early Stopping in Clinical Trials", *Annals of Emergency Medicine*, Vol. 65 No. 2, pp. 214–215.

White, W.B., Grady, D., Giudice, L.C., Berry, S.M., Zborowski, J. and Snabes, M.C. (2012), "A cardiovascular safety study of LibiGel (testosterone gel) in postmenopausal women with elevated cardiovascular risk and hypoactive sexual desire disorder.", *American Heart Journal*, Vol. 163 No. 1, pp. 27–32.

Whitehead, A.L., Julious, S.A., Cooper, C.L. and Campbell, M.J. (2015), "Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable.", *Statistical Methods in Medical Research*, p. 0962280215588241–.

Whitehead, A.L., Sully, B.G.O. and Campbell, M.J. (2014), "Pilot and feasibility studies: is there a difference from each other and from a randomised controlled trial?", *Contemporary Clinical Trials*, Vol. 38 No. 1, pp. 130–3.

Whitehead, J. (1986), "On the bias of maximum likelihood estimation following a sequential test", *Biometrika*, Vol. 73 No. 3, pp. 573–581.

Whitehead, J. (1997), *The Design and Analysis of Sequential Clinical Trials*, Revised 2n., John Wiley & Sons Ltd.

Whitehead, J. (1999), “A unified theory for sequential clinical trials.”, *Statistics in Medicine*, Vol. 18 No. 17-18, pp. 2271–86.

Whitehead, J. (2000), *The Design and Analysis of Sequential Clinical Trials*, Revised 2n., John Wiley & Sons Ltd, Chichester, England.

Whitehead, J. and Jones, D. (1979), “The analysis of sequential clinical trials”, *Biometrika*, Vol. 66 No. 3, pp. 443–452.

Whitehead, J. and Matsushita, T. (2003), “Stopping clinical trials because of treatment ineffectiveness: a comparison of a futility design with a method of stochastic curtailment.”, *Statistics in Medicine*, Vol. 22 No. 5, pp. 677–87.

Whitehead, J. and Stratton, I. m. (1983), “Group Sequential Clinical Trials with Triangular Continuation Regions”, *Biometrics*, Vol. 39 No. 1, pp. 227–236.

WHO ICTRP. (2004), “WHO International Clinical Trials Registry Platform”, available at: http://www.who.int/ictrp/en/ (accessed 14 May 2014).

Wittes, J. (2002), “On changing a long-term clinical trial midstream”, *Statistics in Medicine*, Vol. 21 No. 19, pp. 2789–2795.

Wittes, J. (2012), “Stopping a trial early - and then what?”, *Clinical Trials (London, England)*, Vol. 9 No. 6, pp. 714–20.

Wittes, J. and Brittain, E. (1990), “The role of internal pilot studies in increasing the efficiency of clinical trials.”, *Statistics in Medicine*, Vol. 9 No. 1-2, pp. 65–71; discussion 71–2.

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E. and Proschan, M. (1999), “Internal pilot studies I: type I error rate of the naive t-test.”, *Statistics in Medicine*, Vol. 18 No. 24, pp. 3481–91.

Wolff, A.C., Lazar, A.A., Bondarenko, I., Garin, A.M., Brincat, S., Chow, L., Sun, Y., et al. (2013), “Randomized phase III placebo-controlled trial of letrozole plus oral temsirolimus as first-line endocrine therapy in postmenopausal women with locally advanced or metastatic breast cancer.”, *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, Vol. 31 No. 2, pp. 195–202.

Woodroofe, M. (1992), “Estimation After Sequential Testing: A Simple Approach for a Truncated Sequential Probability Ratio Test”, *Biometrika*, Biometrika Trust, Vol. 79 No. 2, pp. 347–353.

Yao, A.C., Khajuria, A., Camm, C.F., Edison, E. and Agha, R. (2014), "The reporting quality of parallel randomised controlled trials in ophthalmic surgery in 2011: a systematic review.", *Eye (London, England)*, Vol. 28 No. 11, pp. 1341–9.

Yardley, D.A., Brufsky, A., Coleman, R.E., Conte, P.F., Cortes, J., Glück, S., Nabholtz, J.-M.A., et al. (2015), "Phase II/III weekly nab-paclitaxel plus gemcitabine or carboplatin versus gemcitabine/carboplatin as first-line treatment of patients with metastatic triple-negative breast cancer (the tnAcity study): study protocol for a randomized controlled trial", *Trials*, BioMed Central, Vol. 16 No. 1, p. 575.

Yuan, Y. (2009), *Group Sequential Analysis Using the New SEQDESIGN and SEQTEST Procedures, Paper 311-2009*, SAS Institute Inc.

Zang, Y. and Lee, J.J. (2014), "Adaptive clinical trial designs in oncology.", *Chinese Clinical Oncology*, Vol. 3 No. 4, p. 49.

Zannad, F., Stough, W.G., McMurray, J.J. V, Remme, W.J., Pitt, B., Borer, J.S., Geller, N.L., et al. (2012), "When to stop a clinical trial early for benefit: Lessons learned and future approaches", *Circulation: Heart Failure*, Vol. 5, pp. 294–302.

Zellner, D., Zellner, G.E. and Keller, F. (2001), "A SAS macro for sample size re-estimation", *Computer Methods and Programs in Biomedicine*, Vol. 65 No. 3, pp. 183–190.

Zhang, J.J., Blumenthal, G.M., He, K., Tang, S., Cortazar, P. and Sridhara, R. (2012), "Overestimation of the effect size in group sequential trials.", *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, Vol. 18 No. 18, pp. 4872–6.

Zhang, Q., Lai, D. and Davis, B.R. (2015), "Stochastically Curtailed Tests Under Fractional Brownian Motion", *Communications in Statistics - Theory and Methods*, Vol. 44 No. 5, pp. 1053–1064.

Zucker, D.M., Wittes, J.T., Schabenberger, O. and Brittain, E. (1999), "Internal pilot studies II: comparison of various procedures.", *Statistics in Medicine*, Vol. 18 No. 24, pp. 3493–509.

Zwarenstein, M., Treweek, S., Gagnier, J.J., Altman, D.G., Tunis, S., Haynes, B., Oxman, A.D., et al. (2008), "Improving the reporting of pragmatic trials: an extension of the CONSORT statement.", *BMJ (Clinical Research Ed.)*, Vol. 337 No. nov11_2, p. a2390.

Appendices

Appendix 2.1: Other extensions to Stein's internal pilot concept.

Appendix 2.2: Supplementary stopping boundaries.

Appendix 2.3: Specification of spending functions.

Appendix 2.4: A review of methods to compute point estimates following a group sequential test.

Appendix 2.5: Statistical concepts underpinning more complex and flexible ADs.

Appendix 3.1: Interview invitation letter.

Appendix 3.2: Interview information sheet.

Appendix 3.3: Research governance and ethics approval.

Appendix 3.4: Signed ethics application declaration form.

Appendix 3.5: Interview consent form.

Appendix 3.6: Case Study A – An adaptive design evaluating rapid evolving pandemics.

Appendix 3.7: Case Study B – MAMS treatment selection in Tuberculosis.

Appendix 3.8: Case Study C – Bayesian adaptive response randomisation.

Appendix 4.1: Survey instrument for UK CTUs.

Appendix 4.2: Survey instrument for the private sector.

Appendix 4.3: Survey instrument for Public Funders.

Appendix 4.4: Level of personal and UK CTU research group awareness of and experiences in the design and conduct of confirmatory adaptive designs.

Appendix 4.5: Level of personal and the private sector research group awareness of and experiences in the design and conduct of confirmatory adaptive designs.

Appendix 4.6: Awareness of confirmatory adaptive designs, reviewing and commissioning experience of AD-related grant proposals.

Appendix 4.7: Supplementary summary data on UK CTUs' perceptions of important barriers to adaptive designs use in confirmatory trials.

Appendix 4.8: Supplementary summary data on private sector's perceptions of important barriers to adaptive designs use in confirmatory trials.

Appendix 4.9: Supplementary summary data on Public Funders' perceptions of important barriers to adaptive designs use in confirmatory trials.

Appendix 4.10: Supplementary summary data on cross-sector perceptions of concerns towards adaptive designs use in confirmatory trials.

Appendix 5.1: Adaptive design-related search terms and strategy.

Appendix 5.2: List of case studies of confirmatory adaptive designs found in the literature.

Appendix 6.1: Summary data of compliance in the reporting of general CONSORT 2010 checklist items.

Appendix 7.1: Unrestricted SSR results for 3CPO trial.

Appendix 2.1: Other extensions to Stein's internal pilot concept

Stein's 2-stage design has been extended in a clinical trial setting in the context of an 'unrestricted' design assuming an independent two sample t-test (Proschan, 2005; Zucker et al., 1999). The revised sample size (per group) is expressed as:

$$n^* = max\left\{\frac{s_1^2\left(t_{(\frac{\alpha}{2},m)} + t_{(\beta,m)}\right)^2}{\theta_\delta^2} + 1, n_1\right\},$$

where s_1^2 is the internal pilot estimate of the pooled or within-group variance, and $t_{(\frac{\alpha}{2},m)}$ and $t_{(\beta,m)}$ are the Student t-distribution percentiles with m degrees of freedom. Computation of the final test statistic based on the final estimated intervention effect and interim pooled variance has been suggested (Proschan et al., 2006a; Proschan, 2005). This is then compared to the critical value with degrees of freedom associated with s_1^2 and CIs subsequently constructed. Although the statistical properties of the latter design have been studied, it has limitations in clinical trials since it does not use all the data collected for inference (Kieser and Friede, 2000). For the 'restricted' design, Proschan and Wittes (2000) describe a more accurate and unbiased weighted pooled variance estimator alternative to Stein's variance (s_1^2). The authors compared its performance with test statistics computed using Stein's and naïve variance estimators through simulations.

Appendix 2.2: Supplementary stopping boundaries

Wang and Tsiatis

WT (1987) propose a unified theory for the computation of stopping boundaries that are expected to produce the least sample size for a fixed effect size sought, type I and II errors. WT boundaries are used to stop the trial early to reject H_0 for efficacy and/or safety applicable for a one or two-sided test, but not for futility. However, theoretically, their boundaries also require equally spaced information fractions and the number of interim analyses to be fixed and specified in advance. The general form of WT boundaries is expressed by:

$$c_j = C_{j_{WT}}(k, \alpha, \Delta)\, t_j^{\Delta - \frac{1}{2}} \; ; \forall\, j \leq k$$

The shape parameter Δ is chosen to minimise the expected sample size at the scheduled end of the trial. When Δ is 0.5 or 0 the WT boundaries produce the Pocock and OBF boundaries, respectively. The smaller the value of Δ the more difficult it is to stop the trial at earlier interims. Again, the constant $C_{j_{WT}}(k, \alpha, \Delta)$ is chosen through recursive numerical integration by solving equation (2:23). Emerson and Fleming (1989) further describe the WT boundary family for symmetric one and two-sided tests applicable when the number and timing of interim analyses are not fixed in advance, but not allowing for stopping early for futility under H_0.

Whitehead Triangular Shape

Whitehead and Stratton (1983; 2000) propose an approach which does not require pre-specification of the timing of the interim analyses, allows unlimited continuous monitoring, and also applies to a group sequential setting. The authors use a correction resembling a 'Christmas Tree' to boundaries when interim analyses are conducted in a group sequential manner rather than continuous monitoring. The approach depends on the cumulative intervention effect measure (θ) referred to as the efficient score and the Fishers' information (I) which indicates the amount of information about θ contained in the efficient score. The authors point out that these quantities can be computed at any point during the trial and the stopping boundaries sets $\{l_j, u_j\}$ are not fixed in advance, but are calculated according to a pre-defined rule.

Whitehead (2000) defines the upper and lower continuous stopping boundaries functions on the efficient score scale as $a + cI$ and $-a + 3cI$ respectively. The constants a and c are computed through numerical integration such that $P_1(\theta_\delta) = 1 - \beta$. Whitehead states that in practice, monitoring is discrete (at interims). As a

result, the sets $\{l_j, u_j\}$ are calculated based on the efficient score estimates at t_j, $I(t_j)$and $I(t_j) - I(t_{j-1})$ by imposing an overshoot 'Christmas Tree' correction such that:

$$\{l_j, u_j\} = \{-a + 3cI_j + 0.583\sqrt{I(t_j) - I(t_{j-1})}, a + cI_j - 0.583\sqrt{I(t_j) - I(t_{j-1})}\}$$

This is a special case for a design where demonstration of superiority of the intervention is paramount rather than futility. Whitehead later presents a reverse triangular stopping boundaries when stopping early for futility is important than superiority such that $P_1(-\theta_\delta) = 1 - \beta$. In addition, the Whitehead shows the formulation of a 'double triangular' test when both superiority and futility are equally important by combining the 'triangular' and 'reverse triangular' designs. Kittelson and Emerson (1999) extend the concept of power boundaries and describe a unified family of several stopping boundaries defined by the shape and location parameters, including the triangular test under different hypothesis tests of interests.

Conditional Power

Section 2.6 describes stochastic curtailment methods based on CP for a given futility-stopping threshold for decision-making. Literature shows that for a chosen CP futility stopping threshold, equation (2:13) can be reorganised to formulate futility stopping boundaries (Davis and Hardy, 1994; Lachin, 2009; Zhang et al., 2015). Whitehead and Matsushita (2003) compare the performance of stochastic curtailment boundaries and triangular test in futility decision-making. Jiang et al (2014) formulate α spending functions constructed from CP futility boundaries by reorganising equation (2:13). In summary, for a chosen futility stopping threshold, a GSD can be constructed with futility boundaries based on CP to control the desired statistical properties.

Appendix 2.3: Specification of spending functions

OBF and Pocock equivalent

$\alpha_1^*(t_j)$ give the approximate corresponding OBF α spending boundaries that do not necessarily require pre-specification of the number and timing of interim analyses. DeMets et al (1999) suggest that this approach is the most widely adopted in practice.

$$\alpha_1^*(t_j) = \begin{cases} 0\,; & t_j = 0, \\ 4 - 4\boldsymbol{\Phi}\left(\frac{Z_{1-\frac{\alpha}{4}}}{\sqrt{t_j}}\right);\ 0 < t_j \leq 1\ (for\ two-sided\ test), \\ 2 - 2\boldsymbol{\Phi}\left(\frac{Z_{1-\frac{\alpha}{2}}}{\sqrt{t_j}}\right);\ 0 < t_j \leq 1\ (for\ one-sided\ test). \end{cases}.$$

The $\alpha_1^*(t_j)$ is a monotone increasing function of the information fraction defined everywhere within the interval $t_j \in [0,1]$ such that $\alpha_1^*(0) = 0$ and $\alpha_1^*(1) = \alpha$. At the design stage, neither the number nor the timing of the interim analyses needs to be necessarily specified in advance, but only the spending function to be used (Ellenberg et al., 2003; Lan and DeMets, 1983, 1989). However, it assumes that the choice of the future interim analyses is not driven by previous trends in the interim results (Whitehead, 2000).

$\alpha_2^*(t_j)$ gives the approximate corresponding Pocock α spending boundaries for one-sided test are given by (Lan and DeMets, 1983):

$$\alpha_2^*(t_j) = \begin{cases} 0 & ; \quad t_j = 0, \\ \alpha\, ln\left[1 + (e-1)t_j\right]; & 0 < t_j \leq 1. \end{cases}$$

For a two-sided test, α is replaced with 0.5α (Lan and DeMets, 2009). Lan and DeMets (1983) did not provide an 'optimum' criteria for the selection of the α spending function under different design scenarios, which is imperative in clinical trials practice.

Power or Rho family

Kim and DeMets (1987) formalise this idea to generate two-sided symmetric or asymmetric boundaries and propose additional α spending functions in the form of $\alpha {t_j}^{3/2}$, αt_j and $\alpha {t_j}^2$. The authors also studied their properties as a function of the timing of interim analyses (early or late; frequent or equally spaced) for a fixed number of interims in terms of the shape and expected stopping times. Jennison and Turnbull (1989, 1990, 2000a) later adopt a unified expression for these additional α spending functions referred to as ρ family expressed by.

$$\alpha_3^*(t_j) = \begin{cases} 0 & ; \quad t_j = 0, \\ \alpha {t_j}^{\rho} & ; \ 0 < t_j \leq 1, \rho > 0. \end{cases}$$

The $\alpha_3^*(t_j)$ function yields wider early stopping boundaries for higher value of ρ. In addition, the Pocock and OBF boundaries are approximated when ρ corresponds to 0.75 or 1 and 2.5 or 3, respectively.

Gamma Family

Hwang et al (1990) extend the idea by Lan and DeMets (1983) and describe a generalised one parameter family of α spending functions in the form of truncated exponential distributions with shape parameter γ. The function describes the rate at which the overall planned type I error is spent during the entire course of the trial regardless of the statistical test to be used. The general form is expressed by:

$$\alpha(t_j, \gamma) = \begin{cases} \alpha\left(\frac{1 - e^{-\gamma t_j}}{1 - e^{-\gamma}}\right) & ; \ \gamma \neq 0, \\ \alpha t & ; \gamma = 0, \end{cases}$$

where $t_j \in [0,1]$. The function $\alpha(t_j, \gamma)$ is a monotone increasing function defined everywhere in the interval $t_j \in [0,1]$, such that $\alpha(0, \gamma) = 0$ and $\alpha(1, \gamma) = \alpha$. The authors illustrate that α spending functions similar to OBF and Pocock boundaries are approximated by $\alpha(t_j, -4)$ or $\alpha(t_j, -5)$ and $\alpha(t_j, 1)$, respectively. The authors studied the 'optimality' of the subset of these boundaries with respect to minimisation of the expected sample size compared to other α spending functions, Pocock and WT stopping boundaries. The authors conclude that absolute values $|\gamma| \leq 4$ yield boundaries that possess properties that minimises the expected sample size, and are marginally greater than those produced by WT boundaries. However, their approach was not exhaustive in searching for γ values. The authors recommend values of γ between -5 and -1 as appropriate for large long term trials where recruitment is slow and staggered.

Appendix 2.4: A review of methods to compute point estimates following a group sequential test

For continuous outcomes, Whitehead (1986) numerically evaluates the bias of MLE ($\hat{\theta}$) at early stopping for a triangular test as a function of boundaries evaluated using Newton-Raphson iteration. Whitehead tabulates the bias with its standard errors as a function of the effect size.

Whitehead (1986) proposes a bias correction factor to obtain a bias-adjusted estimator by subtracting the bias from the MLE of θ ($\tilde{\theta}_{UWT}$) evaluated at the adjusted estimate. Jennison and Turnbull (2000a) point out that Whitehead's bias-adjusted estimator has a noticeably lower mean square error. More so, the estimator is independent of any particular ordering of the sample space (Emerson and Fleming, 1990). However, as Proschan et al (2006b) highlight, the question still remains concerning whether this bias-adjusted MLE $\tilde{\theta}_{UWT}$ is the optimal estimator.

Pinheiro and DeMets (1997) provide an analytical expression of bias as a function of the information fraction (t_j), variance of the test statistic at the scheduled end, and the exit probabilities of sequential test at t_j. For a two-sided test, the authors investigate the influence of the frequency of interim analyses and magnitude of the effect size on bias through simulation for a class of α spending functions. The authors conclude that bias is a function of the stopping boundaries used, influenced by the timing of the interim analyses, and is greatest in regions when the trial has high probability of stopping early. The authors also suggest a bias-adjusted estimator of θ similar to Whitehead (1986) and found substantial bias reduction for effect sizes commonly used in practice. Citing difficulties in the derivation of the bias based on the above approaches, Wang and Leung (1997) propose bias reduction using a parametric bootstrap method. The authors used simulations to evaluate its efficiency and claim that the method performs competitively compared to other existing methods such as the one proposed by Whitehead (1986), but is slightly inferior.

Ferebee (1983) proposes an algebraic expression to compute an unbiased estimator of θ and proves its existence assuming a Brownian motion process for continuous monitoring. Liu and Hall (1999) state that the sufficient statistic at early stopping $\big(\tau, Z(\tau)\big)$; $\tau = t_j \notin \{l_j, u_j\}$ is not complete for θ. In addition, following a group sequential test, there does not exist a uniformly minimum variance unbiased estimator (UMVUE) of θ, although there exits infinitely many unbiased estimators of θ. Proschan et al (2006b) underline that the completeness property ensures that there cannot be more than one sufficient function of the test statistic that is unbiased for θ. Emerson (1993) adds that the advantage of a UMVUE is that it is unbiased and only depends on the stopping boundaries at interim analyses prior to early stopping. Liu and Hall (1999) went on to show that the

Ferebee estimator is unique and UMVUE. However, this claim was later refuted (Proschan et al., 2006d). Liu and Hall (1999) also claim that their estimator is superior to the MLE, but slightly inferior compared to the Whitehead (1986) biased-adjusted estimator.

Kim (1989) suggests a median unbiased estimators ($\tilde{\theta}^{*}_{MUE}$) as presented in Section 2.7.11.4. Kim also presents a related estimators ($\tilde{\theta}^{*}_{MUE2}$) which is just a midpoint of the lower (θ_L) and upper (θ_U) exact confidence limits computed based on the stagewise ordering as described by Kim and DeMets (1987). Kim evaluates their bias properties compared to the MLE and found a marked decrease in bias for both estimators. However, bias reduction was at a cost of increased mean square errors, which were substantial in some cases depending on the stopping boundaries used and the interim analyses patterns. Kim concludes that both estimates are appropriate as they yield a remarkable reduction in bias and the mean square effort cost is minimal for boundaries often encountered and recommended in practice, such as the OBF.

Emerson (1993) presents a method to estimate a UMVUE of θ by numerical computation of the expectation of $\hat{\theta}(t_1)$ using the first interim analysis conditional on its sufficient statistic at early stopping ($\tau \in t_j \notin \{l_j, u_j\}$) given by $\mathrm{E}\{\hat{\theta}(t_1) \,|(\tau, Z(\tau))\}$. Emerson and Kittelson (1997) further describes a simpler computational approach for its evaluation. Emerson and Fleming (1990) investigate the performance of a number of estimators with respect to bias and mean square error for Pocock and OBF stopping boundaries; $\hat{\theta}$, $\tilde{\theta}_{MUE}$ based on the stagewise, MLE or sample mean ordering, $\tilde{\theta}_{UWT}$ and $\tilde{\theta}_{UMVUE}$. The authors found that $\tilde{\theta}_{UMVUE}$ was associated with no bias, but, its mean square error was relatively larger compared to other estimators. Proschan *et al* (2006b) highlight that Emerson and Fleming mistakenly claimed $\tilde{\theta}_{UMVUE}$ being a UMVUE because $(\tau, Z(\tau))$ is incomplete as shown by Liu and Hall (1999), although it is appealing since the first interim analysis is guaranteed to be observed. Emerson and Fleming (1990) also claim that the estimator based on the MLE ordering performs much better than the one based on the Tsiatis et al (1984) ordering when results from earlier stopping times are treated as being more extreme than the later ones. However, neither of these estimates performed as well as the biased-adjusted estimator ($\tilde{\theta}_{UWT}$) studied by Whitehead (1986).

In search of a UMVUE, Proschan et al (2006b) present an estimator similar to the one suggested by Emerson and Fleming (1990) by computing the conditional expectation of $\hat{\theta}$ given its sufficient statistic $(\tau, Z(\tau))$. The authors show that the variance of the estimators is not greater than that of $\hat{\theta}$. However, the authors did not evaluate the performance of the estimator with respect to bias and mean square error using numerical integration or simulation.

Li and DeMets (1999) investigate bias of the MLE ($\hat{\theta}$) for a subset of α spending functions under a Brownian motion process. The authors found that the magnitude of the bias depends on the stopping boundary used, timing of the interim analysis, and the magnitude of the treatment effect. In addition, boundaries that are more conservative are associated with less bias compared to their counterparts. The authors further studied the Whitehead biased-adjusted estimator ($\tilde{\theta}_{UWT}$), presented an analytical expression for its computation through numerical integration, and evaluate its bias reduction performance compared to the MLE. The authors observed marked bias reduction, although residual bias remains.

Pocock and Hughes (1989) suggests a Bayesian approach to adjust the intervention effect through shrinkage. Here, θ is estimated by the mean of its posterior distribution averaged over the prior distribution about θ which can be elicited from previous clinical trials of the similar intervention. However, problems still exist in cases where there is little information about the intervention under investigation.

Appendix 2.5: Statistical Concepts Underpinning Flexible Complex ADs

Brief background

So far, focus has been paid to ADs with standalone class of trial adaptation such as early stopping or SSR. As displayed in Figure 2.2, some researchers may wish to include multiple adaptations in an ongoing trial for some reasons depending on the rationale provided. For example, in addition to SSR, one may be interested in dropping worst performing intervention(s) ('drop-the-loser') or selecting best performing treatment(s) ('pick-the-winner') or the selection of patient subgroups who are most likely to benefit from the intervention (subgroup selection or enrichment) (Mehta et al., 2009; Stallard et al., 2014). Furthermore, trial adaptation has been assumed to be conducted as pre-planned in order to control the statistical properties of the design. However, this may not be the case in some trials. This section therefore introduces the statistical concepts underpinning more complex and flexible ADs. The approach is flexible in the sense that changes to the design outside the pre-planned adaptations are permitted while controlling for the desired statistical properties such as the type I error. This approach can also be adopted for ADs such as inferential seamless introduced in Section 2.9, flexible variant of MAMS (Magirr et al., 2014), and SSR based on promising intervention effect (Chen et al., 2004; Chen, Li, et al., 2015a; Jennison and Turnbull, 2015; Mehta and Pocock, 2011).

Combination Test Methodology

The general principle of combination tests is based on the analyses of subsequent interim analyses data independently followed by the combination of the independently computed interim test statistics using a pre-planned statistical rule (Schäfer et al., 2006). In other words, data from different interim analyses are theoretically treated as individual studies. The pre-specified combination rule defines how the interim independent test statistics are combined and weighted to produce an overall test statistic for statistical inference to reserve the type I error. A number of approaches have been proposed to achieve this objective (Bauer and Kieser, 1999; Bauer and Kohne, 1994; Chi et al., 1999; Denne, 2001; Fisher, 1998; Lehmacher and Wassmer, 1999; Müller and Schäfer, 2001; Posch and Bauer, 1999; Proschan and Hunsberger, 1995). The general approach is summarised as follows:

1) Specify the combination rule to be used,
2) Specify the decision-making rules guiding the trial adaptation process at interim analysis stages such as stopping criteria,

3) At the 1st interim analysis, determine the associated test statistic and p-value then make a decision regarding the proposed adaptation as specified on item 2).
4) If the trial is continued, based on independent data between subsequent interims, calculate the interim test statistics of independent data increments up to the current interim,
5) Combine the interim independent incremental test statistics and their distributions using the pre-specified combination rule on item 1) to give an overall test statistic with its distribution and associated p-value,
6) Perform statistical inference based on the overall test statistic and p-value computed and adapt the trial accordingly.

The next subsections introduce some of the combination test procedures.

Fisher's Product Combination Function

For a trial with two interim analysis, Bauer and Kohne (1994) introduce the concept of combining data from different interim analysis based on a product of p-values from independent interim analysis. At the 1st interim analysis, the trial is stopped if the associated p-value (p_1) does not exceed some pre-specified values (α_0 and α_1); reject H_0 if $p_1 \leq \alpha_1$ or 'accept' H_0 if $p_1 \leq \alpha_0$. Otherwise the trial proceeds to the 2nd interim analysis if $\alpha_1 < p_1 < \alpha_0$. An independent p-value (p_2) associated with the 2nd interim analysis is computed. The independent p-values from the 1st and 2nd interim analyses are then combined such that H_0 is rejected if their product is less than some critical value ($p_1 p_2 \leq c_2$). Since independent p-values are known to be uniformly distributed, the joint distribution of their product can be easily derived. It can be shown that $c_2 = e^{-0.5\chi^2_{4,1-\alpha}}$; where $\chi^2_{4,1-\alpha}$ is the $(1-\alpha)$ of the Chi-Square distribution with 4 degrees of freedom. For pre-specified values of α_0 and α, in order to preserve the type I error across interim analyses, α_1 is obtained by solving $\alpha_1 + c_2(\ln \alpha_0 - \ln \alpha_1) = \alpha$. The authors also laid a foundation for a trial with three interim analyses. Wassmer (1999) generalises the method to k interim analyses such that the trial is stopped to reject H_0 at the k^{th}interim analysis if $\prod_{i=1}^{k} p_i \leq c_k$; where $c_k = e^{-0.5\chi^2_{2k,1-\alpha}}$. Bauer and Kieser (1999) illustrate the Fisher's p-value combination method for a trial with two interim analysis investigating two interventions compared to a shared control. Vandemeulebroecke (2014) developed an R package '*AdaptTest*' to implement the method for ADs with two interim analyses.

Inverse Normal Combination Function

For a trial with k, interim analyses, Lehmacher and Wassmer (1999) describe a weighted inverse normal method where data from independent interim analyses with associated p-values are combined using a function given by:

$$1 - \Phi\left(\sum_{i=1}^{k} w_i \Phi^{-1}(1 - p_i)\right),$$

where $w_i \in (0,1)$ are the arbitrary chosen weights, pre-specified to indicate the contribution of interim analyses data to the overall test statistic, and $\sum_{i=1}^{k} w_i = 1$. The method is equivalent to a group sequential test when optimum weights are chosen to represent the corresponding interim information fractions (Bretz et al., 2006; Stallard and Todd, 2011). Cui et al (1999) provide a variant of this approach in the context of unblinded SSR.

Conditional Error Function

A number of authors describe methods for combining data from independent interim analyses based on a conditional error function (Denne, 2001; Posch and Bauer, 1999; Proschan and Hunsberger, 1995). The function is a variant of the CP defined by equation 2:12 in Section 2.6.4. This approach has been adopted for ADs with unblinded SSR using the promising zone concept (Bowden and Mander, 2014; Chen et al., 2004; Gao et al., 2008; Mehta and Liu, 2016; Mehta and Pocock, 2011).

Appendix 3.1: Interview Invitation Letter.

Subject Heading: "Invitation to participate in an in-depth interview about your views, attitudes towards and experiences of adaptive trial designs"

Dear Colleague,

When planning a trial, it is common to have sub-optimal information to inform its design which could later undermine its validity. Furthermore, the assumptions made at the planning stage are often overoptimistic. Adaptive designs, in which accumulating trial data may be used to modify key aspects of the trial or make decisions about that ongoing trial, may be beneficial. Nonetheless, such designs have their drawbacks and are considered controversial in some quarters and may not be amenable to the constraints of public funding bodies. At present, adaptive designs are perceived to be rarely applied in publicly funded trials. This study which is funded by the National Institute for Health Research (NIHR) investigates and addresses the issues raised by adaptive designs, specifically in publicly funded trials, and to provide guidance on their implementation in confirmatory trials from a statistical and practical perspective (details of the project are found on this link http://goo.gl/1kWC5E).

As part of this study, we are therefore seeking 20 to 30 participants to participate in nested in-depth qualitative interviews (face-to-face or telephone) to generate themes in order to inform the design of a national quantitative survey among key stakeholders exploring issues on the barriers and how some of these could be alleviated where possible, potential opportunities, perceptions, awareness, experiences (if any) and attitudes towards the use of adaptive designs in confirmatory trials. You are eligible to take part regardless of your positive or negatives views or experiences if you are one of the following key stakeholders directly involved in clinical trials;

- trial statistician in either a publicly or commercially funded sector
- lead clinical trials investigator in publicly funded setting
- chair or vice chair of a public funding panel such as NIHR and MRC (current and previous)
- independent data monitoring committee member
- UK Clinical Trials Unit director/deputy director
- academic interested in adaptive designs
- health economist in clinical trials
- clinical trials regulator such as from MHRA

Appendix 3.1: Interview Invitation Letter.

The interview should take approximately 30-45 minutes to complete. Copies of the information sheet and the consent form are also found on these links for further information (http://goo.gl/ERQD6x and http://goo.gl/po4SXe). If you are eligible and willing to take part in our qualitative interviews please contact Munya Dimairo the Investigator of this project by email (m.dimairo@sheffield.ac.uk) or telephone (+44 (0)l142225204) for further information and on what to do next.

The research is funded by NIHR as part of a Doctoral Research Fellowship (NIHR DRF-2012-05-182) and reviewed favourable ethical opinion has been obtained from ScHARR Research Ethics Committee at the University of Sheffield. The research is supervised by Prof Steven Julious and Prof Jon Nicholl (University of Sheffield), and Prof Sue Todd (University of Reading).

Best regards,

Munya

Appendix 3.1: Interview Invitation Letter.

Project Title: Utility of adaptive designs in publicly funded clinical trials; *a qualitative interviews sub-study*

1 Invitation to take part

We are cordially inviting you to take part in our research. Before making a decision to take part, you will need to understand why the research is being done and what is involved for you. In this regard, please take your time to read this information sheet carefully and feel free to contact the Investigator of this research if you wish using the contact details provided at the bottom.

This information sheet tells you the purpose of this study, what will happen to you if you decide to take part and details about the contact of the research.

2 Purpose of the study and details of what is involved

2.1 What is the purpose of the study?

This qualitative interview sub-study is nested within a large statistical methodology research investigating the application of adaptive designs (where accumulating data from an ongoing trial is used to modify the design aspects of that trial) in publicly funded clinical trials. The overarching aim of the main study is to describe forms of adaptive designs in confirmatory trials with potential to be implemented in publicly funded setting in order to provide statistical efficiency, patient benefit and economic value, and to provide practical recommendations or guidelines to their implementation.

In order to achieve our main study goal, this nested qualitative sub-study aims to generate themes to feed into the quantitative survey which is another component of this larger study. Its objectives are to explore; a) opportunities to the application of adaptive designs, b) forms of adaptations with potential to be implemented in confirmatory phase of clinical trials and when and where they are applicable, c) barriers to their application in public funded setting and how these could be alleviated in order to improve the uptake of adaptive designs in this setting were applicable. Exploring experiences, perceptions and attitudes towards adaptive designs among experts is key to their adoption in routine practice. The research will add to the knowledge base and is also educational, and not for commercial purposes.

2.2 Why have I been selected?

We intend to interview experts directly involved in clinical trials research working in commercial, publicly funded and academic settings. These experts include (but not limited to); trial statisticians, trials methodologists, trial investigators and funding panel members. We will make sure we have a broad representation of views among experts thereby robustly informing later stages of this research. Sampling of participants will continue until we reach a point of saturation with respect to themes which we expect to be around less than 50 participants (expected to be between 20 and 30).

2.3 Do I have to take part?

The decision to take part in this research is entirely upon you. This Information Sheet gives you more information on our study and on what is involved. We will further ask you to sign the consent form to show that you have agreed to take part which could be done electronically or hand written depending on your choice, practicalities and whether interview is done by telephone or face to face.

2.4 What will happen to me if I take part?

When you decide to take part you will be involved in a one off face to face or telephone interview which is expected to last around 30-45 minutes. The interviewer will ask you about your experiences, perceptions, attitudes and views on the application of adaptive designs in clinical trials with respects to barriers, opportunities and prospects. The interview will be audio taped and transcribed for further analysis.

2.5 Expenses and payments

Taking part in this research work is entirely voluntary and there will be no payments made for your participation. In addition, no payments will be made to cover any form of expenses.

2.6 What are the potential benefits of taking part?

There are no personal direct benefits for you to take part in this research. However, your taking part will contribute to the knowledge base on adaptive research designs and how this may improve clinical trials research.

2.7 What are the potential risks or disadvantages of taking part?

We do not foresee any potential risks or disadvantages for you taking part in this one off interview except that you will be expected to spare about 30-45 minutes of your time.

2.8 Will my taking part in this study be kept confidential and what will happen to the data following my interview?

Following the interview, your outputs will be kept in a secure encrypted hard drive. An unencrypted voice recorder will be used during interviews. However, all audio recorded files will be immediately transferred onto an encrypted hard drive as soon as possible after the recording and permanently deleted from an unencrypted voice recorder. The information you provide will be anonymised, kept confidential and transcribed for further analysis. However, anonymous audio transcripts could be used in presentations or conference proceedings.

2.9 What happens if I don't feel like carrying on with the study?

You can withdraw from the interview at any time if you wish not to continue with the interview and you do not have to give a reason. We would like to keep the data you have given us up to that point. However, we would like to keep the data to this point unless you requested us to remove it completely.

2.10 What happens to the results of the research study?

The results of this research will be made public in form of peer reviewed publications and presentations. They will also form part of guidelines or recommendations in form of a monogram on the applications of adaptive designs tailored for publicly funded setting. The summary of the findings will also be made available to you on request as soon as possible following the completion of this research. If you need the summary of the findings, you will need to provide the Principal Investigator of this project with your contact details in form of email address (which will be kept confidential) in order to send you the summary of the results of the study following the completion of the study.

2.11 Who has reviewed the study?

The ethics of this project has been reviewed and approved by ScHARR Research Ethics Committee at the University of Sheffield.

2.12 Who is funding the study?

The research is funded by National Institute of Health and Research (NIHR) as part of a Doctoral Research Fellowship (NIHR DRF-2012-05-182).

3 Further Information and contact details

For further information, please feel free to contact the Principal Investigator of this project.

Munya Dimairo
NIHR Research Fellow in Medical Statistics
University of Sheffield
School of Health and Related Research (ScHARR)
Clinical Trials Research Unit
Regent Court, 30 Regent Street

Appendix 3.1: Interview Invitation Letter.

Sheffield
S1 4DA

Physical address:
Room 3.10, Innovation Centre
Email : m.dimairo@sheffield.ac.uk
Tel : +44 (0) 114 22 25204

Appendix 3.1: Interview Invitation Letter.

The University Of Sheffield.

Munya Dimairo
ScHARR
University of Sheffield

Research Ethics Administrator
Miss Kirsty Woodhead

School of Health and Related Research (ScHARR)
Regent Court
30 Regent Street
Sheffield
S1 4DA

Telephone: +44 (0) 114 222 5453
Fax: +44 (0) 114 272 4095
Email: k.woodhead@sheffield.ac.uk

17 March 2014

Project title: *Utility of adaptive designs in public health funded trials*
6 digit URMS number: *132660*

Dear *Munya*

LETTER TO CONFIRM THAT THE UNIVERSITY OF SHEFFIELD IS THE PROJECT'S RESEARCH GOVERNANCE SPONSOR

The University has reviewed the following documents:

1. A University approved URMS costing record;
2. Confirmation of independent scientific approval (or the award letter if externally funded);
3. Confirmation of independent ethics approval.

All the above documents are in place. Therefore, the University can now **confirm** that it is the project's research governance sponsor and, as research governance sponsor, **authorises** the project to commence any non-NHS research activities. Please note that NHS R&D approval will be required before the commencement of any activities which involve the NHS.

You are expected to deliver the research project in accordance with the University's policies and procedures, which includes the University's Good Research & Innovation Practices Policy: www.shef.ac.uk/ris/other/gov-ethics/grippolicy, Ethics Policy: www.sheffield.ac.uk/ris/other/gov-ethics/ethicspolicy and Data Protection Policies: www.shef.ac.uk/cics/records
As the Principal Investigator you are responsible for monitoring the project on an ongoing basis. Your Head of Department is responsible for independently monitoring the project as appropriate. The project may be audited during or after its lifetime by the University. The monitoring responsibilities are listed in **Annex 1**.

Yours sincerely

Miss Kirsty Woodhead

cc. ScHARR Research Administrator: Mari Bullock

Annex 1

To access the University's research governance website go to:

http://www.sheffield.ac.uk/ris/other/gov-ethics/governance

Monitoring responsibilities of the Principal Investigator ('PI'):

The primary responsibility for project monitoring lies with the PI. You agree to:

1. Establish a **site file** before the start of the project and ensure it remains up to date over the project's entire lifetime:
   http://www.sheffield.ac.uk/ris/other/gov-ethics/governance/rg-forms
2. Provide **progress reports/written updates** to the Head of Department at reasonable points over the project's lifetime, for example at:
   a. three months after the project has started; and
   b. on an annual basis (only if the project lasts for over 18 months); and
   c. at the end of the project.
   See: http://www.sheffield.ac.uk/ris/other/gov-ethics/governance/rg-forms
3. Report **adverse events**, should they occur, to the Head of Department:
   http://www.sheffield.ac.uk/ris/other/gov-ethics/governance/rg-forms
4. Provide progress reports to the research funder (if externally-funded).
5. Establish appropriate arrangements for recording, reporting and reviewing significant developments as the research proceeds – i.e. developments that have a significant impact in relation to one or more of the following:
   - the safety or physical or mental integrity of the participants in the project;
   - the project's scientific direction;
   - the conduct or management of the project.

   The Head of Department should be alerted to significant developments in advance wherever possible.

****

Monitoring responsibilities of the Head of Department

You agree to:

1. Review the **standard monitoring progress reports**, submitted by the PI, and follow up any issues or concerns that the reports raise with the PI.
2. Verify that **adverse events**, should they occur, have been reported properly and that actions have been taken to address the impact of the adverse event(s) and/or to limit the risk of similar adverse event(s) reoccurring.
3. Verify that a project is complying with any **ethics conditions** (e.g. that the information sheet and consent form approved by ethics reviewers is being used; e.g. that informed consent has been obtained from participants).
4. Introduce a form of **correspondence** (e.g. regular email, annual meeting) with a project's PI, that is **proportionate to the project's potential level of risk**, in order to verify that a project is complying with the approved protocol and/or with any research funder conditions. Whatever correspondence is chosen the Head of Department should, as a minimum, ensure that s/he is informed sufficiently in advance about significant developments wherever possible.

Appendix 3.1: Interview Invitation Letter.

The University Of Sheffield.

Kirsty Woodhead
Ethics Committee Administrator

Regent Court
30 Regent Street
Sheffield S1 4DA
Telephone: +44 (0) 114 2225453
Fax: +44 (0) 114 272 4095 (non confidential)
Email: k.woodhead@sheffield.ac.uk

Our ref: 0676/KW

22 August 2013

Munya Dimairo
ScHARR

Dear Munya

Utility of Adaptive Designs in Publically Funded Clinical Trials.

Thank you for submitting the above research project for approval by the ScHARR Research Ethics Committee. On behalf of the University Chair of Ethics who reviewed your project, I am pleased to inform you that on 22 August 2013 the project was approved on ethics grounds, on the basis that you will adhere to the documents that you submitted for ethics review.

The research must be conducted within the requirements of the hosting/employing organisation or the organisation where the research is being undertaken. You are also required to ensure that you meet any research ethics and governance requirements in the country in which you are researching. It is your responsibility to find out what these are.

If during the course of the project you need to deviate significantly from the documents you submitted for review, please inform me since written approval will be required. Please also inform me should you decide to terminate the project prematurely.

Yours sincerely

Kirsty Woodhead
Ethics Committee Administrator

The University Of Sheffield.

Munya Dimairo <m.dimairo@sheffield.ac.uk>

Ehics question (0676)

3 messages

Munya Dimairo <m.dimairo@sheffield.ac.uk> 13 November 2014 15:20
To: Jane Spooner <J.Spooner@sheffield.ac.uk>

Hi Jane

I have a quick ethics question regarding my ethics approval 0676. It's basically for qualitative interviews followed by two quantitative surveys (private and public sector). Based on the findings from the inteviewes, I thought it's prudent to separate the public sector into 2 surveys (one focusing on Clinical Trials Research Units and the other on members of public funding panels such as NIHR, MRC etc). This separation is a techiical issue since members of the funding panels were also part of my sampling frame for the public sector survey. However, the survey was becoming too long which may impact on the response rate. So I was advised to separate the two. The information sheet is still the same but would want the ethics committee to be aware of this. Does this require an amendment? If so, which form do I need to use?

best regards

Munya

--

"Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning" ~ Albert Einstein

****

Munyaradzi Dimairo
NIHR Research Fellow in Medical Statistics
University of Sheffield
ScHARR/DTS
Clinical Trials Research Unit
Regent Court, 30 Regent Street
Sheffield
S1 4DA

Physical address:
Room 3.10, Innovation Centre

email: m.dimairo@sheffield.ac.uk
tel: +44 (0) 114 22 25204
Link to my publications

****

Jane Spooner <j.spooner@sheffield.ac.uk> 18 November 2014 10:09
To: Munya Dimairo <m.dimairo@sheffield.ac.uk>

Dear Munya

Sorry to be so long in responding, but I've been off sick so am catching up on emails now. As I've only just taken over ethics I'm not sure, but have forwarded your request to Jennifer Burr so she can give advice from the Chair. One of use will get back to you as soon as possible.

Best wishes

Jane .1 of 3

****
Jane M Spooner
Information Manager
School of Health & Related Research
University of Sheffield
Regent Court, 30 Regent Street
Sheffield
S1 4DA
Telephone 0114 222 2965
http://www.shef.ac.uk/scharr/

[Quoted text hidden]

Jennifer A Burr <j.a.burr@sheffield.ac.uk> 18 November 2014 10:27
To: Jane Spooner <j.spooner@sheffield.ac.uk>, Munya Dimairo <m.dimairo@sheffield.ac.uk>

Dear Munya and Jane,
Thanks for your enquiry Munya.
I don't think this requires further action. Jane, we'll just keep a copy of the email on record please.
Kind regards
Jennifer

On 18 November 2014 10:07, Jane Spooner <j.spooner@sheffield.ac.uk> wrote:

> Dear Jennifer
>
> Could you advise on this, please? I'm not sure whether the Ethics Committee would see this as a significant amendment.
>
> Thanks
>
> *Jane*
>
> **********
> ***Jane M Spooner***
> *Information Manager*
> *School of Health & Related Research*
> *University of Sheffield*
> *Regent Court, 30 Regent Street*
> *Sheffield*
> *S1 4DA*
> *Telephone 0114 222 2965*
> **http://www.shef.ac.uk/scharr/**
>
> [Quoted text hidden]

--
Dr Jennifer Burr
Senior Lecturer in Medical Sociology
School of Health and Related Research (ScHARR)
University of Sheffield
Regent Court
30 Regent Street
Sheffield
S1 4DA
Tel: 0114 2220792
J.a.burr@sheffield.ac.uk

The University Of Sheffield.

ScHARR Research Ethics Application Form for Staff and PGRs

This form has been approved by the University Research Ethics Committee (UREC)

Date:	08 July 2013
Name of applicant:	Munyaradzi Dimairo
Research project title:	Utility of Adaptive Designs in Publicly Funded Clinical Trials

Complete this form if you are a **member of staff or a postgraduate research student** who plans to undertake a research project which requires ethics approval via the University Ethics Review Procedure. [x]

or

Complete this form if you plan to submit a **'generic' research ethics application (i.e. an application** that will cover several sufficiently similar research projects). Information on the 'generic' route is at: www.sheffield.ac.uk/ris/other/gov-ethics/ethicspolicy/approval-procedure/review-procedure/generic-research-projects [ ]

If you are an undergraduate or a postgraduate-taught student, this is the wrong form.

This form should be accompanied, where appropriate, by all Information Sheets/Covering Letters/Written Scripts which you propose to use to inform the prospective participants about the proposed research, and/or by a Consent Form where you need to use one.

Further guidance on how to apply is at: http://www.shef.ac.uk/scharr/research/ethicsgovernance

Guidance on the possible routes for obtaining ethics approval (i.e. on the University Ethics Review Procedure, the NHS procedure and the Social Care Research Ethics Committee, and the Alternative procedure) is at: www.shef.ac.uk/ris/other/gov-ethics/ethicspolicy/approval-procedure/ethics-approval

Once you have completed this research ethics application form in full, and other documents where appropriate, check that your name, the title of your research project and the date is contained in the footer of each page and email, as a word document, to the Ethics Administrator k.woodhead@sheffield.ac.uk. Please note that the original signed and dated version of 'Part B' of the application form should be provided to the Ethics Administrator in hard copy.

I confirm that I have read the current version of the University of Sheffield 'Ethics Policy Governing Research Involving Human Participants, Personal Data and Human Tissue', as shown on the University's research ethics website at: www.shef.ac.uk/ris/other/gov-ethics/ethicspolicy [x]

Part A

A1. Title of Research Project: Utility of Adaptive Designs in Publicly Funded Clinical Trials

A2. Contact person (normally the Principal Investigator, in the case of staff-led research projects, or the student in the case of supervised-postgraduate researcher projects):

Title: Mr
Post: Research Fellow
Email: m.dimairo@sheffield.ac.uk
Name: Munyaradzi Dimairo
Department: ScHARR
Telephone: +44 (0) 1142225204

A2.1. Is this a postgraduate researcher project? If yes, please provide the Supervisor's contact details:

Title: Professor
Post: Professor in Medical Statistics
Email: s.a.julious@sheffield.ac.uk
Name: Steven A Julious
Department: ScHARR
Telephone: + 44 (0) 1142220709

A2.2. Other key investigators/co-applicants (within/outside University), where applicable. Please list all (add more if necessary):

Title: Professor
Post: Dean of ScHARR
Email: j.nicholl@sheffield.ac.uk
Name: Jonathan P Nicholl
Department: ScHARR
Telephone: +44 (0) 114 222 5453

Title: Professor
Post: Professor of Medical Statistics
Email: s.c.todd@reading.ac.uk
Name: Sue C Todd
Department: Mathematics and Statistics
Telephone: +44 (0) 118 378 8917

A3. Proposed Project Duration:

Start date: 1st January 2013 End date: 31st December 2015

A4. Mark 'X' in one or more of the following boxes if your research:

	involves adults with mental incapacity or mental illness
	involves prisoners or others in custodial care (e.g. young offenders)
	involves children or young people aged under 18 years
	involves using samples of human biological material collected before for another purpose
	involves taking new samples of human biological material (e.g. blood, tissue) *
	involves testing a medicinal product *
	involves taking new samples of human biological material (e.g. blood, tissue) *
	involves additional radiation above that required for clinical care *
	involves investigating a medical device *
	is social care research
	is ESRC funded
x	**Is taking place in the health service but does not require NHS ethical approval****
x	**URMS number if required (please see below)**

* If you have marked boxes marked ***** then you also need to obtain confirmation that appropriate University insurance is in place. The procedure for doing so is entirely by email. Please send an email addressed to insurance@shef.ac.uk and request a copy of the 'Clinical Trial Insurance Application Form'.

- If you have marked the box** your supervisor, needs to obtain an URMS number (details on the ScHARR web site http://www.shef.ac.uk/scharr/research/ethicsgovernance/ugpgt)

It is recommended that you familiarise yourself with the University's Ethics Policy Governing Research Involving Human Participants, Personal Data and Human Tissue before completing the following questions. Please note that if you provide sufficient information about the research (what you intend to do, how it will be carried out and how you intend to minimise any risks), this will help the ethics reviewers to make an informed judgement quickly without having to ask for further details.

A5. Briefly summarise:

i. The project's aims and objectives:
(this must be in language comprehensible to a lay person)

We aim to explore innovative, effective and acceptable ways for testing new treatments in publicly funded clinical trials where accumulating data from the trial could be used to modify the design and make decisions about an ongoing trial (adaptive designs) which could offer patient and economic benefits. The overarching long term goal is to come up with some recommendations or guidelines for trialists and funders on the implementation of "acceptable" adaptive designs.

Another aim is to explore barriers, opportunities and attitudes, and use of adaptive designs in public funded setting among key stakeholders (UK Clinical Trial Units (CTUs), trialists, academia, funding panels and pharmaceutical industry)

Lastly, noting the gap between statistical theory and practical implementations of these designs, we intend to illustrate practically how these designs are implemented statistically and decisions made (such as stopping a trial as soon as there is evidence that the treatment under investigation is not effective) during the course of the trial using case studies of secondary clinical trial data.

ii. The project's methodology:
(this must be in language comprehensible to a lay person)

In order to answer the above aims and objectives, this research is structured in 3 phases;

Phase 1 (first year): An extensive statistical literature review will be undertaken to explore forms of adaptive designs which could be implemented in publicly funded confirmatory clinical trials. A narrative pearl growing literature review strategy will be used.

Phase 2 (part of second year): This stage is in two parts;

1) Qualitative in-depth and semi-structured interviews of important stakeholders defined in section A5 (i) will be undertaken to generate themes and understand the attitudes towards, barriers and opportunities in the use of adaptive designs, and under what circumstances they may have used adaptive designs in the past, present and future. This will be either face-to-face or by telephone and designed to inform the development of a questionnaire to be used in a subsequent quantitative survey. A stratified purposive sampling technique will be employed in order to obtain broad views among key stakeholders. Participants will be interviewed until a saturation point is reached.
2) National quantitative surveys utilising the themes generated from the qualitative interviews will be undertaken to further explore opportunities, attitudes towards and

barriers to the use of adaptive designs. One is targeted at leads of CTU (which could also be completed by the lead Statistician of that CTU) and the other is generic targeting all experts in clinical trials in the UK. Target key holders for both qualitative interviews and quantitative survey will include trial methodologists within UK CTUs, academia, charity organisations, funding panel members and pharmaceutical industry involved in clinical trials. In addition, qualitative interviews (either face to face or by telephone) will be done with informed consent. It should be noted that although this will not be done within the NHS, the study uses participants within health services structures. Therefore, NHS staff (such as clinicians involved in trials) could be part of the stakeholders mentioned above. It is the quantitative interviews and qualitative survey for which ethical approval is being sought.

Phase 3 (third year): Case studies of five clinical trials will be used to illustrate how adaptive designs are implemented statistically and the decisions made during the course of the trials. Key questions will be on whether any of the trials could have stopped early, e.g. in situations where there is evidence that the treatment under investigation is not effective and how sample size re-estimations are conducted. Further simulation work will be done under different assumptions to investigate impact of changes to the design characteristics on decision making. In addition, impact of decisions on financial, patients and study time will be explored. Application for the use secondary data has been completed using a different form through a different route to this application.

A6. What is the potential for physical and/or psychological harm/distress to participants?

It should be noted that this study is not a clinical trial although participants will take part in qualitative and quantitative interviews. In addition, secondary clinical trial data will be used. Hence, we don't envisage any physical, psychological and distress to participants due to their involvement in qualitative interviews and quantitative survey.

A7. Does your research raise any issues of personal safety for you or other researchers involved in the project? (especially if taking place outside working hours or off University premises)

We don't envisage any issues of personal safety to the PI and no other researchers will be involved during the qualitative interviews even though it is likely that some interviews could be done outside working hours. In addition, some face to face interviews will be conducted off the University premises during the PI's internship within the pharmaceutical industry although we don't foresee any personal safety issues.

If yes, explain how these issues will be managed.

Not applicable

A8. How will the potential participants in the project be:

i. **Identified?** *(please ensure that all practical issues about contacting individuals are covered and that you are not requesting the personal details of individuals be given over without their consent)*

Identification of key stakeholders for the quantitative survey will be done through the NIHR Evaluation, Trials and Statistics infrastructure already in existence, UK CTU network, Clinical Research Networks, funding panels (NIHR and MRC), Statisticians in the Pharmaceutical Industry

forum, personal contacts of experts both in academia and pharmaceutical industry, MRC infrastructure, Charity Research Organisation (such as Wellcome Trust and Cancer Research) and academic training events through universities. The contact details of these experts are publicly available and we will not be getting contact details of individuals from third parties. This survey will be anonymous and completed online, and no personal sensitive information will be collected. However, additional information on gender, age (in bands), employment sector and years of experience (in bands) will be collected. Name of employment organisation will not be collected. Participants will consent to this survey by completing an online questionnaire. A consent field (yes/no) will be created on top of the first page of the questionnaire and participants can only complete the survey is they answered yes on this field. The field will explicitly explain that the data from the survey will only be used for research purposes (see Appendix A and B).

As for the qualitative interviews, purposive stratified sampling will be done to capture views across the body of key stakeholders above. Interviews will be either face-to-face or by telephone and informed consent will be sought prior to interviews (electronically or hard copy) among all interview participants. In addition, all interview outputs will be kept in a secure encrypted hard drive. An unencrypted voice recorder will be used during interviews. However, all audio files will be immediately transferred onto an encrypted hard drive as soon as possible after the recording and permanently deleted from an unencrypted voice recorder.

As for case studies of clinical trials data, the PI has received written consent (through email confirmation) from all the Principal or Chief Investigators and have also agreed to the identification of the trials as part of this work. Furthermore, the data is anonymised with unique identifier per record without post code information and participants' names. The data will be kept on an encrypted internal and external hard drives to ensure security and confidentiality of participants information.

ii. Approached?

Generic emails will be send to email list within existing network infrastructures (such as UK CTUs, NIHR, MRC Methodology Hub, medical statistics groups) inviting them to take part and complete an online questionnaire survey if they are involved in clinical trials. Funding panels mainly from NIHR and MRC will be approached through these funding bodies. In addition, participants will also be invited to take part during qualitative interviews through personal contacts of experts.

iii. Recruited?

All key stakeholders who are involved in clinical trials research and approached as explained above will be eligible to take part.

A9. Will informed consent be obtained from the participants?

Yes [x] No [ ]

If informed consent or consent is NOT to be obtained please explain why. Further guidance is at: www.shef.ac.uk/ris/other/gov-ethics/ethicspolicy/policy-notes/consent

Written informed consent will be sought for the qualitative interviews for all participants prior to interviews. For participants who will be interviewed through the telephone, an electronic consent form will be sent to them prior to interviews which they will need to sign and return. As for

participants who will be interviewed face to face, signed hard copies will be obtained prior to interviews.

As for the quantitative online survey, a consent field will be on the front page of an online questionnaire and ticking a yes on this field will be regarded as giving consent to take part in the study.

A9.1. This question is only applicable if you are planning to obtain informed consent:
How do you plan to obtain informed consent? (i.e. the proposed process?):

The informed consent form is attached as part of this ethics process and will be given to participants in plenty of time prior to the interviews. The consent form will be explained to participants and they will only be interviewed if they have agreed in writing (electronic or hard copy) to take part in the study.

Remember to attach your consent form and information sheet (where appropriate)

A10. What measures will be put in place to ensure confidentiality of personal data, where appropriate?
(As a minimum please ensure details are included of: how long data will be kept; when and how it will be destroyed; that PCs and other devices are password protected; that personal details are encrypted. This information should also be included on your information sheet).

All project data will be stored in an encrypted internal and external drives to ensure confidentiality of participants and trial data. Data will be kept for future research purposes with informed consent (see Appendix A, B and C), mainly for publications related to this study following its completion.

A11. Will financial/in kind payments (other than reasonable expenses and compensation for time) be offered to participants? (Indicate how much and on what basis this has been decided)

- No financial or any kind of payments (such as vouchers) will be offered to participants in kind or any sort of compensation.

A12. Will the research involve the production of recorded media such as audio and/or video recordings?

YES [x] NO [ ]

A12.1. This question is only applicable if you are planning to produce recorded media:
How will you ensure that there is a clear agreement with participants as to how these recorded media may be stored, used and (if appropriate) destroyed?

The participant information sheet and consent form (Appendix A) for the interviews will clearly explain how the interviews will be recorded with participants, transcribed for further analysis and stored in a secure place for future research purposes following the completion of the study.

Guidance on a range of ethical issues, including safety and well-being, consent and anonymity, confidentiality and data protection are available at: www.shef.ac.uk/ris/other/gov-ethics/ethicspolicy/policy-notes

University Research Ethics Application Form - Part B - The Signed Declaration

Title of Research Project:

Utility of Adaptive Designs in Publicly Funded Clinical Trials

I confirm my responsibility to deliver the research project in accordance with the University of Sheffield's policies and procedures, which include the University's *'Financial Regulations'*, *'Good Research Practice Standards'* and the *'Ethics Policy Governing Research Involving Human Participants, Personal Data and Human Tissue'* (Ethics Policy) and, where externally funded, with the terms and conditions of the research funder.

In signing this research ethics application form I am also confirming that:

- The form is accurate to the best of my knowledge and belief.
- The project will abide by the University's Ethics Policy.
- There is no potential material interest that may, or may appear to, impair the independence and objectivity of researchers conducting this project.
- Subject to the research being approved, I undertake to adhere to the project protocol without unagreed deviation and to comply with any conditions set out in the letter from the University ethics reviewers notifying me of this.
- I undertake to inform the ethics reviewers of significant changes to the protocol (by contacting my academic department's Ethics Administrator in the first instance).
- I am aware of my responsibility to be up to date and comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data, including the need to register when necessary with the appropriate Data Protection Officer (within the University the Data Protection Officer is based in CiCS).
- I understand that the project, including research records and data, may be subject to inspection for audit purposes, if required in future.
- I understand that personal data about me as a researcher in this form will be held by those involved in the ethics review procedure (e.g. the Ethics Administrator and/or ethics reviewers) and that this will be managed according to Data Protection Act principles.
- If this is an application for a 'generic' project, all the individual projects that fit under the generic project are compatible with this application.
- **I understand that this project cannot be submitted for ethics approval in more than one department, and that if I wish to appeal against the decision made, this must be done through the original department.**

Name of the Supervisor:

Steven A Julious

Name of the student:

Munyaradzi Dimairo

Signature of the Supervisor:

Steven A Julious	**Date:** 22 Aug 2013.

Email the completed application form and provide a signed, hard copy of 'Part B' to the Ethics Administrator (also enclose, if relevant, other documents).

Title: Utility of adaptive designs in publicly funded clinical trials

Consent Form for Interviews: a Qualitative Sub-study

Thank you for reading the information sheet about the interview sub-study. If you are happy to participate then please complete and sign the form below. Please initial the boxes below to confirm that you agree with each statement:

	Please Initial box:
I confirm that I have read and understood the information sheet dated **[22/08/2013]** and have had the opportunity to ask questions.	☐
I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline.	☐
I understand that my responses will be kept strictly confidential. I understand that my name will not be linked with the research materials, and will not be identified or identifiable in the report or reports that result from the research.	☐
I agree for this interview to be tape-recorded. I understand that the audio recording made of this interview will be used only for analysis and that extracts from the interview, from which I would not be personally identified, may be used in any conference presentation, report or journal article developed as a result of the research. I understand that no other use will be made of the recording without my written permission, and that no one outside the research team will be allowed access to the original recording.	☐
I agree that my anonymised data will be kept for future research purposes such as publications related to this study after the completion of the study.	☐
I agree to take part in this interview.	☐

_____	_____	_____
Name of participant	Date	Signature
_____	_____	_____
Principal Investigator	Date	Signature

To be counter-signed and dated electronically for telephone interviews or in the presence of the participant for face to face interviews

Copies: *Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form, and the information sheet. A copy of the signed and dated consent form should be placed in the main project file which must be kept in a secure location.*

Appendix 3.6: Case Study A – An adaptive design evaluating rapid evolving pandemics

"Yes, it was a mild pandemic so I guess that gives you some context for what a bad pandemic would be like but in developed countries there was probably in the order of 200,000 to 300,000 patients admitted to intensive care with severe influenza and as best as we can work out about 30 – 40 were randomised in a controlled trial. The reason for that is that the lag time for setting up any sort of interventional trial is at least 6 – 12 months requiring Funders to agree, ethics committees to approve, governance to be organised, study drug and study procedures, case report forms, randomisation project management and none of those things can be put together in time lines measured in weeks to a month or so. So part of the concept behind this Toronto meeting which has also been a now merged effort between the people who went to that Toronto meeting and a group called ISARIC (International Severe and Respiratory Illness Consortium) is the broad proposal that we should have "mothballed" randomised controlled trials that can be activated in a couple of weeks in the event of an arrival of a pandemic. So that means we would know the interventions that we are going to test, we would know the sites that are going to participate, they would ethics approval, case report form would be developed, the study drug and all other study procedures would be warehoused and able to be activated at short notice. But having gone to the point of feeling that that was a good idea, there was then a lot of really useful discussion about whether a classic frequentist design is optimal or whether some sort adaptive Bayesian design would be preferable. In a pandemic there is a rapidly rising number of patients with life threatening illness so a pandemic wave will typically last 6 – 10 weeks in any location but move around the world in different dynamics, so where an emergency is going to appear there is a time imperative to provide results to Clinicians and policymakers. If you do a classic frequentist design you have got to make guesses about adequate sample size and your fixed to that design and then you are not able to utilise the information that is accrued to adapt the design as you go and so you could find that you have randomised 5000 patients and was being futile and proven to be futile up to 3000 or you could have randomised 3000 and found that you have got an effective intervention after 2000 and many patients who would have benefited from that knowledge and have been through the health care system without that knowledge being applied to their treatment. So there is this time imperative that doesn't so much exist where there is a stable incidence of disease, adaptation really comes to the fore as a design feature to generate the maximum amount of useful knowledge in a shortest possible period of time." (QL21 Chief Investigator, design experience)

Appendix 3.7: Case Study B – MAMS treatment selection in Tuberculosis

"So particularly the one particular trial that we're involved in a trial in tuberculosis using the multi –arm and multi - stage design. And it's a trial that's being conducted in South and East Africa. So there are lots of challenges … I suppose first of all the methodology itself was developed in Oncology. It was developed for a different disease area for a different set of end points (time-to-event). And so there's a little bit of work in adapt in the methodology to fit the particular disease area so in terms of what the end points were and some issues around how the analysis is done. So most of the methodologies around cancer where the end point is death - in tuberculosis the phase 2 end point is culture conversion. So there are a few minor tweaks like that. So the first challenge I guess was adapting methodology.

We didn't have any problem in communicating (the design) to the other stakeholders (investigators) and the other investigators based in Africa. This was the right design and everyone seemed to appreciate that it seemed to be a sensible approach to use an adaptive design where we might be dropping arms that are not performing well. The five arm study was just four intervention arms with one driven analysis. I get challenges with data entry ensuring that the data was entered in a way that it could be collected on a central database in a rapid fashion and put it into analysis. So it could be cleaned and entered on an ongoing basis but at any one time you have real time access to the data. And so we've used tablet computers for data entry with a central database. It's clear that adaptive design is somewhat complicated to implement. And because of interim analysis there are obviously issues around controlling type one error but also controlling the power as well. You don't want to be stopping arms or indeed declaring arms; it's a bit efficacious too early. So there are concerns there. I think in our context the methodology was fairly well worked out so that helped. So I work in [organisation] and so I have colleagues who are working on the design and methodology so it was use …….so the power of work on the methodology alongside running the trial. And so that helped to provide the necessary underpinning of methodology and to have confidence that it was working and that you know time and error was adequately controlled and that it made sense. I guess that's always a concern with adaptive design; it's a little bit more complicated than a traditional design. So in our study we had interim analysis so we looked at the data. We had an independent data monitoring committee who view the results. And based on the sort of pre-specified threshold they recommended stopping 2 arms. And so there were challenges in how to communicate that. We worked very carefully about how to communicate it. Explaining that it would that it was about lack of benefit rather than necessarily toxicity. So these arms are being dropped because there was insufficient evidence for benefit, not necessarily that they were bad arms or the patients were being put at risk or that they were necessarily inferior drugs. But just that in the context of the trail we're

optimising resources and focusing on the other 2 arms rather than which arms are being dropped. There was a lot of discussion about how to present that. So that's definitely a challenge around you know how you present the results of or how you present the modifications during adaptive design" (QL26 Statistician, design and conduct experience)

Appendix 3.8: Case Study C – Bayesian adaptive response randomisation design

"Yes, so I have been involved in the design that's currently the [organisation] on a large multi-centre combined phase 2 /3 trial that is on the verge of having regulatory approval looking at a novel vasopressor agent. The patient relation remains static over time and there's a Bayesian prior … but it involves a response adaptive randomisation in the beginning phase to try to narrow down the population … so we start the trial with a placebo arm and three dose arms and then there is a possibility of adding a fourth dose arm. The model that the design incorporates is essentially a model of dose response to try to gleam information from the overall behaviour of the drug across the different doses and that's used with a set of pre-trial assumptions about likely dose response so that in the absence of a dose response we will go with the lowest dose as long as there are no safety concern that is dose related and there appears to be efficacy it will tend to favour higher dose and in fact the fourth dose that could be added in is an even higher dose, it's like we haven't hit the ceiling. Then for the switch from phase 2 to phase 3 is variable depending on the behaviour in the opening part but the final half will still have a minimum number of patients into the final dose and the final placebo so that the total sample size is potentially variable because there could be a delay before we choose the dose if there isn't much of a difference in the dose so the dose selection study is variable in size but the second half will have minimum number of patients. It's going through European regulatory approval right now but hasn't been received, so it's in the planning phase. I've also been involved in generally thinking through whether an adaptive design is more flexible for studying interventions in the critically ill during epidemics. A trial could be up and running, targeting a broad population but could then be more flexible to adjust potentially by bringing in new study arms in the middle of an epidemic as opposed to trying to launch the trial at the time the epidemic starts so seeing whether adaptive design is more suitable for just in time research if that makes sense.

Probably the biggest regulatory hurdle has been a discussion around – there's a couple – one is around controlling alpha (type 1) error and the concern that it's not that easy to generate empirically but rather you have to run simulations and there's a discussion about whether the simulations truly model across the entire space of the trial. So whether it is giving a robust understanding of the alpha error, that has been a concern. Another concern is that if you are recruiting placebo patients all the way through but you disproportionately recruit intervention patients into the final arm more heavily towards the end of the trial then there is the risk that if there is some sort of time bias across the trial then, if for example people get better at caring for these patients later or if there's changes in case mix then the comparison may not be taking the full advantage of trying to use true randomisation to hope that you would have a balanced distribution of both known and unknown covariance at baseline. Then a

third hurdle is the overall combined trial most definitely speaks to whether drug on average is superior to placebo but it's less clear that it really gives a clear statement about whether the particular dose is better. You know, you could argue that the statement about dose can only really be made about the later patients enrolled in that arm and so there is a question about what would be the label that one could right from the results of the trial. I guess another issue is using a Bayesian framework, it does not necessarily return a classic p-value, which does not necessarily concern the regulatory authorities but it might be hard for Clinicians to understand what that means. So those are some of the hurdles which I think are relatively generic but those are some of the issues we have been discussing at length with the regulatory authorities" (QL22 Chief Investigator, design experience)

Adaptive Designs Survey

7%

This nested survey aims to assess the uptake of adaptive designs (ADs) in human confirmatory trials and perceived associated barriers with potential facilitators to their use in the publicly funded setting. We consider ADs designed, implemented and analysed using Frequentist methods (excludes Bayesian). The details of the main project are found on this link: http://goo.gl/hD7czi. The findings will help to identify priority areas to improve uptake and facilitate successful implementation of ADs when appropriate in the UK.

Definition: ***By AD, we mean prospectively planned changes to the design or decisions to stop an ongoing trial based on interim primary outcome(s) related data from that trial without undermining its scientific integrity, validity and credibility. This excludes decisions based solely on external information or operational feasibility such as poor recruitment as part of internal pilot risk management assessment criteria.*** **Note that an internal pilot trial can still be classified as an AD so long it has statistical related objectives based on the primary outcome data. For instance, estimating event rate in the control group or variability for a binary or continuous endpoint respectively for sample size review.**

The survey is aimed to be completed by the CTU Director/Deputy Director or designated Senior Statistician. We will keep your responses and any identifiable information completely confidential. This study has been approved (0676) by ScHARR Ethics Committee at the University of Sheffield. Your participation is voluntary and you may wish to discontinue at any point in time. Most of the questions are closed, there are a few open-ended questions to allow you to give further detail on your responses or suggestions where necessary.

The survey should take no more than 15 minutes to complete. Thank you for taking part in this survey. Your feedback is very important to us. Please complete the consent statement below.

I consent for my anonymised data to be used for research purposes

- ○ Yes
- ○ No

Next

Adaptive Designs Survey

14%

Q1. Unique UK CTU registration number

Q2. What is your experience in trials research (years)?

- Below 5
- 5 to <10
- 10 to <15
- 15 to <20
- At least 20

Q3. What is your age group (years)?

- Below 35
- 35 to <40
- 40 to <45
- 45 to <50
- 50 to <55
- At least 55

Q4. What is your main role or responsibility in trials research?

- CTU Director/Deputy Director
- Senior Statistician

Prev Next

Adaptive Designs Survey

21%

Q5. How would you describe your level of familiarity with adaptive designs (ADs)?

Not at all familiar	Slightly familiar	Moderately familiar	Very familiar	Extremely familiar
○	○	○	○	○

Q6. How would you rate the *level of awareness of types of ADs in confirmatory trials* among your research team directly involved in trial design and implementation?

Not at all aware	Slightly aware	Moderately aware	Very aware	Extremely aware
○	○	○	○	○

Q7. How would you rate the level of the following **within your CTU**?

	None	Little experience	Some experience	Substantial experience
Experience in the design of ADs in confirmatory trials?	○	○	○	○
Experience in the conduct of ADs in confirmatory trials?	○	○	○	○

Q7.1. How would you rate **your** level of the following?

	None	Little experience	Some experience	Substantial experience
Experience in the design of ADs in confirmatory trials?	○	○	○	○
Experience in the conduct of ADs in confirmatory trials?	○	○	○	○

Prev Next

Adaptive Designs Survey

29%

Q8. To what extent do you view the following as *main barriers* to the use of ADs when appropriate in confirmatory trials *within your CTU*?

	Not an important barrier	Somewhat an important barrier	Moderately important barrier	Extremely important barrier
Lack of awareness of benefits of ADs	○	○	○	○
Lack of awareness of when ADs are appropriate	○	○	○	○
Lack of awareness of which scope of ADs are acceptable in confirmatory trials	○	○	○	○
Difficulties in marketing ADs to key stakeholders in trials research (such as collaborators, funders and regulators)	○	○	○	○
Research team being more comfortable with the conventional mainstream designs compared to ADs	○	○	○	○
Lack of practical implementation knowledge	○	○	○	○
Lack of applied training to facilitate practical implementation	○	○	○	○
Lack of practical hands-on experience	○	○	○	○

	Not an important barrier	Somewhat an important barrier	Moderately important barrier	Extremely important barrier
Insufficient access to case studies to facilitate practical learning	○	○	○	○
Inadequate data management support infrastructure for timely capturing, cleaning and transfer for decision making as part of the adaptation	○	○	○	○
Lack of statistical expertise	○	○	○	○
Unfamiliarity with key implementation resources such as validated statistical software	○	○	○	○
Lack of knowledge to use existing validated statistical software	○	○	○	○
Lack of capacity of proposal developers with basic knowhow	○	○	○	○
Lack of time to support planning in relation to other competing conventional mainstream design priorities	○	○	○	○
Amount of work and effort required at the design or planning stage	○	○	○	○
Practical complexities during trial conduct for successful implementation	○	○	○	○
Statistical complexities during planning (such as simulation work)	○	○	○	○
Difficulties in setting up acceptable upfront decision making criteria to guide the adaptation	○	○	○	○
Statistical complexities during implementation (such as analysis and reporting)	○	○	○	○
Lack of bridge funding required to support design work of time consuming and complex ADs	○	○	○	○
Worry about the impact of stopping early on full-time research staff employment contracts	○	○	○	○
Costing complexities on grant application	○	○	○	○
Fear of regulatory reluctance and jeopardising chances of obtaining regulatory approval due to the use of an AD	○	○	○	○
Tension during early stopping decision making among key decision makers (such as data monitoring committees and funders)	○	○	○	○

	Not an important barrier	Somewhat an important barrier	Moderately important barrier	Extremely important barrier
Previous negative experiences with ADs based on funders/reviewers comments	○	○	○	○
Previous negative experiences during implementation	○	○	○	○

Other (Please provide any further comments regarding any other perceived important barriers within your CTU not mentioned above)

Prev Next

Adaptive Designs Survey

36%

Q9. To what extent would you rate your *level of concern* relating to the use of ADs in confirmatory trials?

	Not at all concerned	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned
Early stopping of trials for efficacy	○	○	○	○	○
Early stopping of trials for non-inferiority	○	○	○	○	○
Early stopping of trials for futility	○	○	○	○	○
Robustness of AD methodology to influence policy decision making when trials are stopped early	○	○	○	○	○
Acceptability of the findings from ADs by the research community or regulators in order to change practice	○	○	○	○	○
Fear of introducing operational bias by leaking of information related to the adaptation thereby compromising the scientific integrity, validity and credibility of the trial results	○	○	○	○	○
Potential change in the population during the course of an adaptive trial and its impact on interpretation of the findings	○	○	○	○	○
Impact of ADs on secondary trial objectives (such as health economics) when trials are stopped early	○	○	○	○	○

Q10. To what extent do you agree or disagree with the following statements?

	Strongly Disagree	Disagree	Somewhat disagree	Neither Disagree Nor Agree	Somewhat agree	Agree	Strongly Agree
General attitude towards ADs by public funders has changed positively in the past 10 years	○	○	○	○	○	○	○

Appendix 4.1: Survey Instrument for UK CTUs.

	Strongly Disagree	Disagree	Somewhat disagree	Neither Disagree Nor Agree	Somewhat agree	Agree	Strongly Agree
Clinical trial investigators are generally positive towards ADs depending on how they are marketed to them	○	○	○	○	○	○	○
There are negative attitudes towards ADs among some influential statistical communities	○	○	○	○	○	○	○
Regulatory awareness and experiences of ADs is improving	○	○	○	○	○	○	○
Independent Data Monitoring Committee (IDMC) members are often reluctant to stop trials early unless for safety reasons	○	○	○	○	○	○	○
IDMC members are generally unfamiliar with ADs	○	○	○	○	○	○	○
Lack of financial incentives beyond self-esteem among public sector IDMC members may negatively influence their reluctance to take key trial advisory decisions	○	○	○	○	○	○	○
There is a general conceived perception among peer reviewers or journal editors that stopping a trial early is failure	○	○	○	○	○	○	○

Appendix 4.1: Survey Instrument for UK CTUs.

	Strongly Disagree	Disagree	Somewhat disagree	Neither Disagree Nor Agree	Somewhat agree	Agree	Strongly Agree
There are general negative attitudes among peer reviewers/journal editors towards ADs	○	○	○	○	○	○	○
Ethics boards are generally unfamiliar with AD methodology	○	○	○	○	○	○	○
Scientific boards are generally unfamiliar with AD methodology	○	○	○	○	○	○	○
Public funders are generally risk averse to fund complex ADs associated with high financial uncertainty	○	○	○	○	○	○	○

Prev Next

Adaptive Designs Survey

43%

Q11. How useful do you think the following would be in facilitating the use of ADs when appropriate in confirmatory trials?

	Not at all useful	Not very useful	Somewhat useful	Very useful
A consensus guidance document on the acceptable scope of ADs tailored for publicly funded confirmatory trials	○	○	○	○
A troubleshooting toolkit of specific questions grant applicants need to ask themselves before considering various types of ADs	○	○	○	○
Accessible published case studies of ADs such as focusing on the design, implementation, challenges, lessons learnt, statistical issues and facilitators to challenges	○	○	○	○
An AD tailored CONSORT guidance document as a way to enhance transparency and completeness in the conduct and reporting	○	○	○	○

Other (please provide any further comments or suggestions you may have regarding potential facilitators to the use of ADs in confirmatory trials)

Q12. In your CTU, how would you rank the theme of ADs (of use or research of ADs related methods) in confirmatory in the next 5 to 10 years?

Not a priority	Low priority	Medium priority	High priority	Essential
○	○	○	○	○

Please specify ADs related themes/areas of interest which you would need help if applicable

Q13. Do you have an AD working group (focusing on the<u>use or research of ADs related methods</u> in confirmatory trials) within your CTU?

- ○ Yes
- ○ No
- ○ Prefer not to say

If anwered No, could you please provide further comment on why?

Q14. Would you consider using ADs in some of your future confirmatory trials (when appropriate) to answer research questions?

Would not consider	Might or might not consider	Definitely consider
○	○	○

Please provide any further comments which you may have

Prev Next

Adaptive Designs Survey

50%

Q15. Have you ever submitted an AD confirmatory trial grant application to any funding source?

- ○ Yes
- ○ No
- ○ Prefer not to say

Q16. If answered Yes above, approximately how many of these grant applications have been successfully funded by the ... ?

	Approximate number
Public sector (such as UK government bodies and charity orgainisations)	
Private sector (such as pharmaceutical or biotech companies)	
Both private and public sector	

Q17. What best describes each type of confirmatory AD(s) which has been successfully funded from any source (public and/or private)? This includes trials which have been completed or are ongoing or awaiting commencement.

Q17a. Sample size review

- ○ Yes
- ○ No

If answered Yes above, what best describes the type of sample size review and approximate number of trials?

	Type of sample size review	Approx. number of trials
Blinded review only allowing for an increase in sample size		
Blinded review allowing an increase or decrease in sample size		
Unblinded review only allowing for an increase in sample size		
Unblinded review allowing for an increase or decrease in sample size		
Unblinded review based on interim treatment effect or conditional power (promising zone concept)		
Other 1		
Other 2		

Other (Please specify the meaning of "Other 1" and "Other 2" where applicable)

Prev Next

Adaptive Designs Survey

57%

Q17b. Standard two arm group sequential design

- Yes
- No

If answered Yes above, what best describes the*type of planned stopping criteria* and approx. number of trials?

	Type of early stopping criteria	Approx. number of trials
Futility only		
Efficacy only		
Either futility or efficacy		
Safety/harm only		
Either futility or safety/harm		
Non-inferiority		
Other 1		
Other 2		
Other 3		
Other 4		

Other (Please specify the meaning of "Other 1" to "Other 4" where applicable)

Prev Next

Adaptive Designs Survey

64%

Q17c. Futility assessment (outside group sequential framework)

- Yes
- No

If answered Yes above, what best describes the type of*futility assessment* and approx. number of trials?

	Type of futility assessment	Approx. number of trials
Based on conditional power		
Based on predictive power		
Based on confidence interval of the treatment effect		
Other 1		
Other 2		

Other (Please specify the meaning of "Other 1" and "Other 2" from abover where applicable)

Prev Next

Adaptive Designs Survey

71%

Q17d. Operational Seamless 2/3 design

- Yes
- No

If answered YES above, what best describes the type of*operational seamless adaptation* and approx. number of trials?

	Type of adaptation	Number of trials
Only allowing dropping of futile treatments in phase 2		
Selection of only one promising treatment in phase 2		
Selection of multiple promising treatments in phase 2		
Allowing addition of new treatments or dropping of futile treatments in phase 2		
Other 1		
Other 2		
Other 3		

Other (please specify the meaning of "Other 1" to "Other 3" from above where applicable)

Prev Next

Adaptive Designs Survey

79%

Q17e. Inferential Seamless 2/3 design

- Yes
- No

If answered YES above, what best describes the type of*inferential seamless adaptation* and approx. number of trials?

	Type of adaptation	Approx. number of trials
Only allowing dropping of futile treatments in phase 2		
Selection of only one promising treatment in phase 2		
Selection of multiple promising treatments in phase 2		
Allowing addition of new treatments or dropping of futile treatments in phase 2		
Other 1		
Other 2		
Other 3		

Other (please specify the meaning of "Other1" to "Other 3" from above where applicable)

Prev Next

Adaptive Designs Survey

86%

Q17f. Strictly phase 3 multi-arm multi-stage (MAMS) design

- Yes
- No

If answered YES above, what best describes the type of *MAMS adaptation* and number of trials?

	Type of adaptation	Approx. number of trials
Only allowing dropping of futile treatments		
Option to stop the trial for futility or stopping futile treatments		
Option to stop the trial for futility or efficacy or stopping futile treatments		
Allowing addition of new treatments or dropping of futile treatments		
Other 1		
Other 2		
Other 3		

Other (please specify the meaning of "Other 1" to "Other 3" from above where applicable)

Prev Next

Adaptive Designs Survey

93%

Q17g. Other types of ADs in confirmatory trials

	Yes or No	Approx. number of trials
Information based group sequential design		
Standard group sequential design plus sample size review		
Patient enrichment (subgroup selection) design		
Response adaptive randomisation (strictly based on the primary outcome or biomarker)		
Other 1		
Other 2		
Other 3		
Other 4		
Other 5		
Other 6		

Other (please specify the meaning of "Other 1" to "Other 6" from above where applicable)

Prev Next

Adaptive Designs Survey

100%

Q18. Approximately what is the current number of confirmatory trials in your research portfolio (grants won regardless of the stage to completion)?

Q19. Approximately what is the percentage of the following confirmatory trial interventions in your CTU research portfolio? (*give a number between 0 and 100*)

	Approx. percentage
Drugs	
Biologics	
Devices	
Surgical	
Other non-pharmacological (such as complex, physiotherapy, behavioural, educational and nutritional interventions)	

Q20. Approximately what percentage of your confirmatory trials require regulatory approval (such as from MHRA, EMA and FDA)? This exclude local ethicals approval required for every trial.

Q21. What best describes the main disease areas in your confirmatory research portfolio? (*tick all that apply*)

- Cancer/oncology
- Cardiovascular
- Diabetes
- Emergency medicine
- Mental healh
- Health services
- Infectious diseases
- Primary care
- Rare/orphan diseases
- Respiratory
- Musculoskeletal
- Public Health

Other (please specify)

Q22. At what email address would you like to be contacted informing you about our findings?

Prev Done

Private Sector Adaptive Designs Survey

7%

This nested survey aims to assess the uptake of adaptive designs (ADs) in human confirmatory trials and perceived associated barriers with potential facilitators to their use in the publicly funded setting. We consider ADs designed, implemented and analysed using Frequentist methods (excludes Bayesian). The details of the main project are found on this link: http://goo.gl/hD7czi. The findings will help to identify priority areas to improve uptake and facilitate successful implementation of ADs when appropriate. Moreover, we will also compare and contrast between private and public sector perspectives predominantly in the UK.

Definition: ***By AD, we mean prospectively planned changes to the design or decisions to stop an ongoing trial based on interim primary outcome(s) related data from that trial without undermining its scientific integrity, validity and credibility. This excludes decisions based solely on external information or operational feasibility such as poor recruitment as part of internal pilot risk management assessment criteria.*** **Note that an internal pilot trial can still be classified as an AD so long it has statistical related objectives based on the primary outcome data. For instance, estimating variability for a continuous primary endpoint for sample size review.**

The survey is aimed to be completed by the Research Leader or designated Lead/Senior/Principal Statistician of a clinical trials research group within a company (pharmaceutical or biotech or CRO). We will keep your responses and any identifiable information completely confidential. This study has been approved (0676) by ScHARR Ethics Committee at the University of Sheffield. Your participation is voluntary and you may wish to discontinue at any point in time. Most of the questions are closed, there are a few open-ended questions to allow you to give further detail on your responses or suggestions where necessary.

The survey should take no more than 15 minutes to complete. Thank you for taking part in this survey. Your feedback is very important to us. Please complete the consent statement below.

I consent for my anonymised data to be used for research purposes

- ○ Yes
- ○ No

Next

Private Sector Adaptive Designs Survey

13%

Appendix 4.2: Survey Instrument for the Private Sector.

Q1. How would you describe your company?

- ○ Pharmaceutical
- ○ Biotech
- ○ CRO
- ○ Other (please specify)

Q2. In what country is your company located?

- ○ United Kingdom
- ○ Switzerland
- ○ Other (please specify)

Q3. What is your experience in trials research (years)?

- ○ Below 5
- ○ 5 to <10
- ○ 10 to <15
- ○ 15 to <20
- ○ At least 20

Q4. What is your age group (years)?

- ○ Below 35
- ○ 35 to <40
- ○ 40 to <45
- ○ 45 to <50
- ○ 50 to <55
- ○ At least 55

Q5. What is your main role or responsibility in trials research?

- ○ Research Leader
- ○ Lead/Senior/Principal Statistician

Q5.1. Which section of your company do you belong to?

- ○ Statistics
- ○ Clinical development

Prev Next

Appendix 4.2: Survey Instrument for the Private Sector.

Private Sector Adaptive Designs Survey

20%

Q6. How would you describe your level of familiarity with adaptive designs (ADs)?

Not at all familiar	Slightly familiar	Moderately familiar	Very familiar	Extremely familiar
○	○	○	○	○

Q7. How would you rate the level of the following among your research team directly involved in trial design and implementation?

	Not at all aware	Slightly aware	Moderately aware	Very aware	Extremely aware
Awareness of types of ADs in confirmatory trials	○	○	○	○	○

Q8. How would you rate the level of the following **in your company**?

	None	Little experience	Some experience	Substantial experience
Experience in the design of ADs in confirmatory trials?	○	○	○	○
Experience in the conduct of ADs in confirmatory trials?	○	○	○	○

Q8.1 How would you rate **your** level of the following?

	None	Little experience	Some experience	Substantial experience
Experience in the design of ADs in confirmatory trials?	○	○	○	○
Experience in the conduct of ADs in confirmatory trials?	○	○	○	○

Prev Next

Private Sector Adaptive Designs Survey

27%

Q9. To what extent do you view the following as *main barriers* to the use of ADs when appropriate in confirmatory trials *within your company*?

	Not an important barrier	Somewhat an important barrier	Moderately important barrier	Extremely important barrier	N/A
Lack of awareness of the benefits of ADs	○	○	○	○	○
Lack of awareness of when ADs are appropriate	○	○	○	○	○
Lack of awareness of which scope of ADs are acceptable in confirmatory trials	○	○	○	○	○
Difficulties in marketing ADs to key stakeholders in trials research (such as collaborators, R& D and regulators)	○	○	○	○	○
Research team being more comfortable with the conventional mainstream designs compared to ADs	○	○	○	○	○
Lack of practical implementation knowledge	○	○	○	○	○
Lack of applied training to facilitate practical implementation	○	○	○	○	○
Lack of practical experiences	○	○	○	○	○
Insufficient access to case studies to facilitate practical learning	○	○	○	○	○
Inadequate data management support infrastructure for timely capturing, cleaning and transfer for decision making as part of the adaptation	○	○	○	○	○
Lack of statistical expertise	○	○	○	○	○
Unfamiliarity with key implementation resources such as validated statistical software	○	○	○	○	○
Lack of knowledge to use existing validated statistical software	○	○	○	○	○
Lack of general expertise around ADs at the trial planning stage	○	○	○	○	○
Lack of time to support planning in relation to other competing conventional mainstream design priorities	○	○	○	○	○
Amount of work and effort required at the design or planning stage	○	○	○	○	○
Practical complexities during trial conduct for successful implementation	○	○	○	○	○
Statistical complexities during planning (such as simulation work)	○	○	○	○	○
Difficulties in setting up acceptable upfront decision making criteria to guide the adaptation	○	○	○	○	○

	Not an important barrier	Somewhat an important barrier	Moderately important barrier	Extremely important barrier	N/A
Statistical complexities during implementation (such as analysis and reporting)	○	○	○	○	○
Lack of bridge funding required to support design work of time consuming and complex ADs	○	○	○	○	○
Worry about the impact of stopping early on staff contracts	○	○	○	○	○
Complexity in deriving the cost of the proposed trial	○	○	○	○	○
Fear of regulatory reluctance and jeopardising chances of obtaining regulatory approval due to the use of an AD	○	○	○	○	○
Tension during early stopping decision making among key decision makers (such as data monitoring committees and sponsors/funders)	○	○	○	○	○
Previous negative regulatory experiences with ADs such as based on regulatory comments or unsuccessful implementation	○	○	○	○	○
Previous negative experiences with ADs during implementation	○	○	○	○	○
Insufficient financial support from R&D to build an infrastructure to support ADs	○	○	○	○	○
Lack of motivational support from R&D to build an infrastructure to support ADs	○	○	○	○	○
Difficulties outsourcing expertise to support ADs	○	○	○	○	○

Other (Please provide any further comments regarding any other perceived important barriers within your company not mentioned above)

Prev Next

Private Sector Adaptive Designs Survey

33%

Q10. To what extent would you rate your *level of concern* relating to the use of ADs in confirmatory trials within your company?

	Not at all concerned	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned
Early stopping of trials for efficacy	○	○	○	○	○
Early stopping of trials for non-inferiority	○	○	○	○	○
Early stopping of trials for futility	○	○	○	○	○
Robustness of AD methodology to influence policy decision making when trials are stopped early	○	○	○	○	○
Acceptability of the findings from ADs by the research community or regulators in order to change practice	○	○	○	○	○
Fear of introducing operational bias by leaking of information related to the adaptation thereby compromising the scientific integrity, validity and credibility of the trial results	○	○	○	○	○
Potential change in the population during the course of an adaptive trial and its impact on interpretation of the findings	○	○	○	○	○
Impact of ADs on secondary trial objectives (such as health economics) when trials are stopped early	○	○	○	○	○

Q11. To what extent do you agree or disagree with the following statements?

	Strongly Disagree	Disagree	Somewhat disagree	Neither Disagree Nor Agree	Somewhat agree	Agree	Strongly Agree	N/A
Clinical trial investigators are generally positive towards ADs depending on how they are marketed to them	○	○	○	○	○	○	○	○
There are negative attitudes towards ADs among some influential statistical communities	○	○	○	○	○	○	○	○
Regulatory awareness and experiences of ADs is improving	○	○	○	○	○	○	○	○
Independent Data Monitoring Committee (IDMC) members are often reluctant to stop trials early unless for safety reasons	○	○	○	○	○	○	○	○
IDMC members are generally unfamiliar with ADs	○	○	○	○	○	○	○	○
There is a general conceived perception among peer reviewers or journal editors that stopping a trial early is failure	○	○	○	○	○	○	○	○
Ethics boards are generally unfamiliar with AD methodology	○	○	○	○	○	○	○	○
Scientific boards are generally unfamiliar with AD methodology	○	○	○	○	○	○	○	○
Funders/Sponsors are generally risk averse to fund complex ADs associated with high financial uncertainty	○	○	○	○	○	○	○	○

Prev Next

Private Sector Adaptive Designs Survey

40%

Q12. How useful do you think the following would be in facilitating the use of ADs when appropriate in confirmatory trials?

	Not at all useful	Not very useful	Somewhat useful	Very useful
A troubleshooting toolkit of specific questions researchers need to ask themselves before considering various types of ADs	○	○	○	○
Accessible published case studies of ADs such as focusing on the design, implementation, challenges, lessons learnt, statistical issues and facilitators to challenges	○	○	○	○
An AD tailored CONSORT guidance document as a way to enhance transparency and completeness in the conduct and reporting	○	○	○	○

Other (please provide any further comments or suggestions you may have regarding potential facilitators to the use of ADs in confirmatory trials)

Q13. In your company, how would you rank the theme of ADs (of *use or research of ADs related methods* in confirmatory) in the next 5 to 10 years?

Not a priority	Low priority	Medium priority	High priority	Essential
○	○	○	○	○

Q14. Do you have an AD working group (focusing on the *use or research of ADs related methods* in confirmatory trials) within your company?

- ○ Yes
- ○ No
- ○ Prefer not to say

If anwered No, could you please provide further comment on why?

Prev Next

Private Sector Adaptive Designs Survey

47%

Q15. Would you consider using ADs in some of your future confirmatory trials (when appropriate) to answer research questions?

Would not consider	Might or might not consider	Definitely consider	N/A
○	○	○	○

Q16. Have you ever submitted an AD confirmatory trial grant application or won a contract for an AD confirmatory trial or has your company funded an AD confirmatory trial?

- ○ Yes
- ○ No
- ○ Prefer not to say

Q17: Have you ever worked on or your company been involved with an AD confirmatory trial?

- ○ Yes
- ○ No

Prev Next

Private Sector Adaptive Designs Survey

53%

Q18. What best describes each type of confirmatory AD(s) you or your company have been involved with? This includes trials which have been completed or are ongoing or awaiting commencement.

Q18a. Sample size review

- ○ Yes
- ○ No

If answered Yes above, what best describes the type of sample size review and approximate number of trials?

	Type of sample size review	Approx. number of trials
Blinded review only allowing for an increase in sample size		
Blinded review allowing an increase or decrease in sample size		
Unblinded review only allowing for an increase in sample size		
Unblinded review allowing for an increase or decrease in sample size		
Unblinded review based on interim treatment effect or conditional power (promising zone concept)		
Other 1		
Other 2		

Other (Please specify the meaning of "Other 1" and "Other 2" where applicable)

Prev Next

Private Sector Adaptive Designs Survey

60%

Q18b. Standard two arm group sequential design

- Yes
- No

If answered Yes above, what best describes the*type of planned stopping criteria* and approx. number of trials?

	Type of early stopping criteria	Approx. number of trials
Futility only		
Efficacy only		
Either futility or efficacy		
Safety/harm only		
Either futility or safety/harm		
Non-inferiority		
Other 1		
Other 2		
Other 3		
Other 4		

Other (Please specify the meaning of "Other 1" to "Other 4" where applicable)

Prev Next

Private Sector Adaptive Designs Survey

67%

Q18c. Futility assessment (outside group sequential framework)

- Yes
- No

If answered Yes above, what best describes the type of*futility assessment* and approx. number of trials?

	Type of futility assessment	Approx. number of trials
Based on conditional power		
Based on predictive power		
Based on confidence interval of the treatment effect		
Other 1		
Other 2		

Other (Please specify the meaning of "Other 1" and "Other 2" from abover where applicable ）

Prev Next

Private Sector Adaptive Designs Survey

73%

Q18d. Operational Seamless 2/3 design

- Yes
- No

If answered YES above, what best describes the type of*operational seamless adaptation* and approx. number of trials?

	Type of adaptation	Number of trials
Only allowing dropping of futile treatments in phase 2		
Selection of only one promising treatment in phase 2		
Selection of multiple promising treatments in phase 2		
Allowing addition of new treatments or dropping of futile treatments in phase 2		
Other 1		
Other 2		
Other 3		

Other (please specify the meaning of "Other 1" to "Other 3" from above where applicable)

Prev Next

Private Sector Adaptive Designs Survey

80%

Q18e. Inferential Seamless 2/3 design

- Yes
- No

If answered YES above, what best describes the type of*inferential seamless adaptation* and approx. number of trials?

	Type of adaptation	Approx. number of trials
Only allowing dropping of futile treatments in phase 2		
Selection of only one promising treatment in phase 2		
Selection of multiple promising treatments in phase 2		
Allowing addition of new treatments or dropping of futile treatments in phase 2		
Other 1		
Other 2		
Other 3		

Other (please specify the meaning of "Other1" to "Other 3" from above where applicable)

Prev Next

Private Sector Adaptive Designs Survey

87%

Q18f. Strictly phase 3 multi-arm multi-stage (MAMS) design

- Yes
- No

If answered YES above, what best describes the type of *MAMS adaptation* and number of trials?

	Type of adaptation	Approx. number of trials
Only allowing dropping of futile treatments		
Option to stop the trial for futility or stopping futile treatments		
Option to stop the trial for futility or efficacy or stopping futile treatments		
Allowing addition of new treatments or dropping of futile treatments		
Other 1		
Other 2		
Other 3		

Other (please specify the meaning of "Other 1" to "Other 3" from above where applicable)

Prev Next

Private Sector Adaptive Designs Survey

93%

Q18g. Other types of ADs in confirmatory trials

	Yes or No	Approx. number of trials
Information based group sequential design		
Standard group sequential design plus sample size review		
Patient enrichment (subgroup selection) design		
Response adaptive randomisation (strictly based on the primary outcome or biomarker)		
Other 1		
Other 2		
Other 3		
Other 4		
Other 5		
Other 6		

Other (please specify the meaning of "Other 1" to "Other 6" from above where applicable)

Prev Next

Private Sector Adaptive Designs Survey

100%

Q19. Approximately what is the percentage of the following confirmatory trial interventions in your company research/contract portfolio? (*give a number between 0 and 100*)

	Approx. percentage
Drugs	
Biologics	
Devices	
Surgical	
Other non-pharmacological (such as complex, physiotherapy, behavioural, educational and nutritional interventions)	

Q20. Would you consider using AD in pivotal confirmatory trials within your company?

Would not consider	Might or might not consider	Definitely consider	N/A
○	○	○	○

Q21. What best describes the main disease areas in your confirmatory research/contract portfolio? (*tick all that apply*)

- ☐ Cancer/oncology
- ☐ Cardiovascular
- ☐ Diabetes
- ☐ Emergency medicine
- ☐ Mental healh
- ☐ Health services
- ☐ Infectious diseases
- ☐ Primary care
- ☐ Rare/orphan diseases
- ☐ Respiratory
- ☐ Musculoskeletal
- ☐ Public Health

Other (please specify)

Q22. At what email address would you like to be contacted informing you about our findings (optional)?

Prev Done

Public funders Adaptive Designs Survey

17%

This nested survey aims to assess the uptake of adaptive designs (ADs) in human confirmatory trials and perceived associated barriers with potential facilitators to their use in the publicly funded setting. We consider ADs designed, implemented and analysed using Frequentist methods (excludes Bayesian). The details of the main project are found on this link: http://goo.gl/hD7czi. The findings will help to identify priority areas to improve uptake and facilitate successful implementation of ADs when appropriate in the UK.

Definition: ***By AD, we mean prospectively planned changes to the design or decisions to stop an ongoing trial based on interim primary outcome(s) related data from that trial without undermining its scientific integrity, validity and credibility. This excludes decisions based solely on external information or operational feasibility such as poor recruitment as part of internal pilot risk management assessment criteria.*** **Note that an internal pilot trial can still be classified as an AD so long it has statistical related objectives based on the primary outcome data. For instance, estimating variability for a continuous primary endpoint for sample size review.**

This survey is aimed to be completed by panel board members of public funders. We will keep your responses and any identifiable information completely confidential. This study has been approved (0676) by ScHARR Ethics Committee at the University of Sheffield. Your participation is voluntary and you may wish to discontinue at any point in time. Most of the questions are closed although there are a few open-ended questions to allow you to give further detail on your responses or suggestions where necessary.

The survey should take no more than 10 minutes to complete. Thank you for taking part in this survey. Your feedback is very important to us. Please complete the consent statement below.

I consent for my anonymised data to be used for research purposes

- ○ Yes
- ○ No

Next

Public funders Adaptive Designs Survey

33%

Appendix 4.3: Survey Instrument for Public Funders.

Q1. How would you describe the funding source of the funding board you are a member of?

- ○ Government Funded (such as NIHR and MRC)
- ○ Charity Organisations (such as Cancer Research UK and Wellcome Trust)
- ○ Both Government and Charity Funded Organisations
- ○ Other (please specify)

Q2. For how long have you previously served as a funding panel board member (years)?

○ Below 5 ○ 5 to <10 ○ 10 to <15 ○ 15 to <20 ○ At least 20

Q3. What is your age group (years)?

○ Below 35 ○ 35 to <40 ○ 40 to <45 ○ 45 to <50 ○ 50 to <55 ○ At least 55

Q4. What is your main roles or responsibilities in trials research (*tick all that apply*)?

- ☐ CTU Director/Deputy Director
- ☐ Trial Statistician
- ☐ Chief Investigator
- ☐ Clinical Expert
- ☐ Trial Methodologist
- ☐ Other (please specify)
- ☐ Trials Management Expert
- ☐ Health Economist
- ☐ Trial Steering Committee member
- ☐ Independent Data Monitoring Committee member
- ☐ Patient Representative

Q5. How would you describe your board membership on the funding panel?

- ○ Panel Chair
- ○ Vice Panel Chair
- ○ Ordinary Member
- ○ Lay Member
- ○ Other (please specify)

Prev Next

Public funders Adaptive Designs Survey

50%

Q6. How would you describe your level of familiarity with adaptive designs (ADs)?

Not at all familiar	Slightly familiar	Moderately familiar	Very familiar	Extremely familiar
○	○	○	○	○

Q7. How would you rate the level of the following ...?

	Not at all aware	Slightly aware	Moderately aware	Very aware	Extremely aware
Your awareness of types of ADs in confirmatory trials	○	○	○	○	○
Awareness of types of ADs in confirmatory trials **among your funding panel board members**	○	○	○	○	○

Other (please provide any related further comments you may have)

Q8. How would you rate the level of the following ...?

	None	Little experience	Some experience	Substantial experience
Your experience in the reviewing of ADs grant applications in confirmatory trials to recommend for funding?	○	○	○	○
Funding panel board experience in the reviewing of ADs grant applications in confirmatory trials to recommend for funding?	○	○	○	○
Your experience in the commissioning of ADs grant applications in confirmatory trials (such as contract negotiation and trial monitoring)?	○	○	○	○
Funding panel board experience in the commissioning of ADs grant applications in confirmatory trials (such as contract negotiation and trial monitoring)?	○	○	○	○

Prev Next

Public funders Adaptive Designs Survey

67%

Q9. To what extent do you view the following as *main barriers* to the recommendation for funding of ADs (when appropriate) in confirmatory trials by your funding panel board? Please leave item(s) blank if you do not know

	Not an important barrier	Somewhat an important barrier	Moderately important barrier	Extremely important barrier
Lack of awareness of benefits of ADs	○	○	○	○
Lack of awareness of when ADs are appropriate	○	○	○	○
Lack of awareness of which scope of ADs are acceptable in confirmatory trials	○	○	○	○
Rationale for ADs not well explained in the grant application	○	○	○	○
The type AD proposed and its scope not well described in the grant application	○	○	○	○
Decision making criteria to guide the adaptation not well described	○	○	○	○
Inadequate description of the costing scenarios of ADs in the grant application	○	○	○	○
Funding panel board members being more comfortable with the conventional mainstream designs compared to ADs	○	○	○	○
Difficulties in drawing up flexible contractual agreements suitable for ADs such as for Clinical Trials Units	○	○	○	○
Lack of commissioning experience of ADs	○	○	○	○
Funding board generally being risk averse to fund complex ADs associated with high financial uncertainty	○	○	○	○
Lack of expertise of reviewers of ADs to help funding panel boards during grant review process	○	○	○	○
Tension during early stopping decision making of ADs among key decision makers (such as data monitoring committees and funders)	○	○	○	○
Negative attitudes towards ADs among some funding panel board members	○	○	○	○

Other (Please provide any further comments regarding any other perceived important barriers not mentioned above from a funding panel perspective)

Prev Next

Public funders Adaptive Designs Survey

83%

Q10. To what extent would you rate your *level of concern* relating to the use of ADs in confirmatory trials?

	Not at all concerned	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned	Don't know
Early stopping of trials for efficacy	○	○	○	○	○	○
Early stopping of trials for non-inferiority	○	○	○	○	○	○
Early stopping of trials for futility	○	○	○	○	○	○
Robustness of AD methodology to influence policy decision making when trials are stopped early	○	○	○	○	○	○
Acceptability of the findings from ADs by the research community or regulators in order to change practice	○	○	○	○	○	○
Fear of introducing operational bias by leaking of information related to the adaptation thereby compromising the scientific integrity, validity and credibility of the trial results	○	○	○	○	○	○
Potential change in the population during the course of an adaptive trial and its impact on interpretation of the findings	○	○	○	○	○	○
Impact of ADs on secondary trial objectives (such as health economics) when trials are stopped early	○	○	○	○	○	○

Other (please specify any other comments you may have)

Q11. To what extent do you agree or disagree with the following statements relating to ADs use?

	Strongly Disagree	Disagree	Somewhat disagree	Neither Disagree Nor Agree	Somewhat agree	Agree	Strongly Agree
General attitude towards ADs by public funders has changed positively in the past 10 years	○	○	○	○	○	○	○
There is lack of bridge funding required to support design work of time consuming and complex ADs by researchers	○	○	○	○	○	○	○
Researchers are more worried about the impact of early trial stopping on full-time research staff employment contracts	○	○	○	○	○	○	○
There is lack of time to support adequate planning of complex ADs in relation to other competing conventional mainstream design priorities and turnaround time for grant submission	○	○	○	○	○	○	○
Independent Data Monitoring Committee (IDMC) members are often reluctant to stop trials early unless for safety reasons	○	○	○	○	○	○	○
IDMC members are generally unfamiliar with ADs	○	○	○	○	○	○	○
Lack of financial incentives beyond self-esteem among public sector IDMC members may negatively influence their reluctance to take key trial advisory decisions	○	○	○	○	○	○	○
There is a general conceived perception among peer reviewers or journal editors that stopping a trial early is failure	○	○	○	○	○	○	○
There are general negative attitudes among peer reviewers/journal editors towards ADs	○	○	○	○	○	○	○
Ethics boards are generally unfamiliar with ADs	○	○	○	○	○	○	○
Scientific boards are generally unfamiliar with ADs	○	○	○	○	○	○	○

Prev Next

Public funders Adaptive Designs Survey

100%

Q12. How useful do you think the following would be in facilitating the use of ADs when appropriate in confirmatory trials?

	Not at all useful	Not very useful	Somewhat useful	Very useful
Refresher training of funding panel board members to improve awareness of ADs prior to their panel meetings	○	○	○	○
A consensus guidance document on the acceptable scope of ADs tailored for publicly funded confirmatory trials	○	○	○	○
A troubleshooting toolkit of specific questions grant applicants need to ask themselves before considering various types of ADs	○	○	○	○
Accessible published case studies of ADs such as focusing on the design, implementation, challenges, lessons learnt, statistical issues and facilitators to challenges	○	○	○	○
An AD tailored CONSORT guidance document as a way to enhance transparency and completeness in the conduct and reporting	○	○	○	○

Other (please provide any further comments or suggestions you may have regarding potential facilitators to the use of ADs in publicly funded confirmatory trials)

Q13. As a funding board, how would you rank the theme of ADs (of use or research of ADs related methods) in confirmatory trials in the next 5 to 10 years?

Not a priority	Low priority	Medium priority	High priority	Essential
○	○	○	○	○

Q14. As a funding board, have you ever recommended funding of an AD in confirmatory trials?

○ Yes

○ No

Q15. Would you consider recommending ADs in future confirmatory trials for funding (when appropriate) to answer clinically important research question(s)?

Would not consider	Might or might not consider	Definitely consider
○	○	○

Please provide any related further comments which you may have

Q16. At what email address would you like to be contacted informing you about our findings (optional)?

Prev Done

Appendix 4.4: Level of personal and UK CTU research group awareness of and experiences in the design and conduct of confirmatory ADs

Variable description	Scoring	Total (N=30) n(%)
Experience in clinical trials research (years)	<5	1(3%)
	5 to <10	4(13%)
	10 to <15	7(23%)
	15 to <20	5(17%)
	≥20	11(37%)
	Missing	2(7%)
Age group (years)	<35	4(13%)
	35 to <40	3(10%)
	40 to <45	6(20%)
	45 to <50	5(17%)
	50 to <55	5(17%)
	≥55	5(17%)
	Missing	2(7%)
Main roles/duties in clinical trials research	CTU Director/Deputy Director	10(33%)
	Designated Senior Statistician	18(60%)
	Missing	2(7%)
Personal level of familiarity with ADs	Slightly familiar	5(17%)
	Moderately familiar	17(57%)
	Very familiar	3(10%)
	Extremely familiar	2(7%)
	Missing	3(10%)
Perceived level awareness of type of ADs among research team within the CTU	Slightly aware	8(27%)
	Moderately aware	12(40%)
	Very aware	5(17%)
	Extremely aware	2(7%)
	Missing	3(10%)
Perceived CTU experience in the design of ADs in confirmatory trials	None	3(10%)
	Little experience	10(33%)
	Some experience	13(43%)
	Substantial experience	1(3%)
	Missing	3(10%)
Perceived CTU experience in the conduct of ADs in confirmatory trials	None	6(20%)
	Little experience	10(33%)
	Some experience	10(33%)
	Substantial experience	1(3%)
	Missing	3(10%)
	None	3(10%)

Personal experience in the design of ADs in confirmatory trials	Little experience	6(20%)
	Some experience	11(37%)
	Substantial experience	1(3%)
	Missing	9(30%)
Personal experience in the conduct of ADs in confirmatory trials	None	4(13%)
	Little experience	5(17%)
	Some experience	11(37%)
	Missing	10(33%)

Appendix 4.5: Level of personal and the private sector research group awareness of and experiences in the design and conduct of confirmatory ADs

Variable description	Scoring	Total (N=17) n(%)
Classification of the organisation	Pharmaceutical	9(53%)
	Contract Research Organisation	7(41%)
	Missing	1(6%)
Geographical location of the organisation	United Kingdom	13(76%)
	Switzerland	1(6%)
	Other	2(12%)
	Missing	1(6%)
Experience in clinical trials research (years)	5 to <10	3(18%)
	10 to <15	6(35%)
	15 to <20	1(6%)
	≥20	6(35%)
	Missing	1(6%)
Age group (years)	<35	2(12%)
	35 to <40	4(24%)
	40 to <45	6(35%)
	45 to <50	4(24%)
	Missing	1(6%)
Main roles/duties in clinical trials research	Lead/Senior/Principal Statistician	13(76%)
	Research Leader	3(18%)
	Missing	1(6%)
Section of the organisation the responder belongs to	Statistics	13(76%)
	Missing	4(24%)
Personal level of familiarity with ADs	Slightly familiar	2(12%)
	Moderately familiar	5(29%)
	Very familiar	7(41%)
	Extremely familiar	1(6%)
	Missing	2(12%)
Perceived level awareness of type of ADs among research team	Slightly aware	1(6%)
	Moderately aware	8(47%)
	Very aware	5(29%)
	Extremely aware	1(6%)
	Missing	2(12%)
Organisation's experience in the design of ADs in confirmatory trials	Little experience	3(18%)
	Some experience	6(35%)
	Substantial experience	6(35%)
	Missing	2(12%)

Organisation's experience in the conduct of ADs in confirmatory trials	None	1(6%)
	Little experience	3(18%)
	Some experience	8(47%)
	Substantial experience	3(18%)
	Missing	2(12%)
Personal experience in the design of ADs in confirmatory trials	None	1(6%)
	Little experience	1(6%)
	Some experience	5(29%)
	Substantial experience	2(12%)
	Missing	8(47%)
Personal experience in the conduct of ADs in confirmatory trials	None	2(12%)
	Little experience	3(18%)
	Some experience	3(18%)
	Substantial experience	1(6%)
	Missing	8(47%)

Appendix 4.6: Awareness of confirmatory ADs, reviewing and commissioning experience of AD-related grant proposals

Variable description	Scoring	Total (N=86) n(%)
Personal level of familiarity with ADs	Not at all familiar	7(8%)
	Slightly familiar	27(31%)
	Moderately familiar	30(35%)
	Very familiar	11(13%)
	Extremely familiar	1(1%)
	Missing	10(12%)
Personal awareness of types of ADs in confirmatory trials	Not at all aware	8(9%)
	Slightly aware	25(29%)
	Moderately aware	26(30%)
	Very aware	14(16%)
	Extremely aware	3(3%)
	Missing	10(12%)
Panel members' awareness of types of ADs in confirmatory trials	Not at all aware	9(10%)
	Slightly aware	25(29%)
	Moderately aware	29(34%)
	Very aware	8(9%)
	Extremely aware	1(1%)
	Missing	14(16%)
Personal reviewing experience of ADs grant applications in confirmatory trials	None	19(22%)
	Little experience	26(30%)
	Some experience	27(31%)
	Substantial experience	3(3%)
	Missing	11(13%)
Personal commissioning experience of ADs grant applications in confirmatory trials	None	14(16%)
	Little experience	33(38%)
	Some experience	22(26%)
	Substantial experience	4(5%)
	Missing	13(15%)
Panel reviewing experience of ADs grant applications in confirmatory trials	None	40(47%)
	Little experience	23(27%)
	Some experience	11(13%)
	Substantial experience	1(1%)
	Missing	11(13%)
Panel commissioning experience of ADs grant applications in confirmatory trials	None	27(31%)
	Little experience	30(35%)
	Some experience	14(16%)
	Substantial experience	1(1%)
	Missing	14(16%)

Appendix 4.7: Supplementary summary data on UK CTUs' perceptions of important barriers to ADs use in confirmatory trials

Barrier	Perceived importance				Relative importance parameter (95% CI)	Rank
	Not important	Somewhat important	Moderately important	Extremely important		
Lack of bridge funding required to support design work of time consuming and complex ADs	3(12%)	10(40%)	4(16%)	8(32%)	-1.05(-1.59 to -0.52)	1
Lack of practical implementation knowledge	5(20%)	5(20%)	9(36%)	6(24%)	-1.02(-1.56 to -0.49)	2
Lack of practical hands-on experience	5(20%)	5(20%)	9(36%)	6(24%)	-1.02(-1.55 to -0.48)	3
Research team being more comfortable with the conventional mainstream designs compared to ADs	3(12%)	10(40%)	6(24%)	6(24%)	-0.90(-1.43 to -0.37)	4
Difficulties in marketing ADs to key stakeholders in trials research (such as Collaborators, Funders, and Regulators)	5(20%)	7(28%)	8(32%)	5(20%)	-0.72(-1.25 to -0.19)	5
Amount of work and effort required at the design or planning stage	6(24%)	7(28%)	6(24%)	6(24%)	-0.67(-1.20 to -0.14)	6
Lack of time to support planning in relation to other competing conventional mainstream design priorities	5(20%)	8(32%)	7(28%)	5(20%)	-0.67(-1.20 to -0.14)	7
Lack of applied training to facilitate practical implementation	5(20%)	7(28%)	10(40%)	3(12%)	-0.60(-1.13 to -0.07)	8
Insufficient access to case studies to facilitate practical learning	4(16%)	10(40%)	7(28%)	4(16%)	-0.60(-1.12 to -0.07)	9
Practical complexities during trial conduct for successful implementation	5(20%)	9(36%)	8(32%)	3(12%)	-0.47(-1.00 to 0.06)	10
Statistical complexities during planning (such as simulation work)	8(32%)	4(16%)	10(40%)	3(12%)	-0.41(-0.94 to 0.12)	11
Difficulties in setting up acceptable upfront decision making criteria to guide the adaptation	10(40%)	4(16%)	6(24%)	5(20%)	-0.27(-0.80 to 0.26)	12
Lack of capacity of proposal developers with basic knowhow	6(24%)	11(44%)	6(24%)	2(8%)	-0.10(-0.64 to 0.43)	13
Costing complexities on grant application	6(24%)	12(48%)	4(16%)	3(12%)	-0.10(-0.64 to 0.44)	14
Lack of awareness of which scope of ADs are acceptable in confirmatory trials	6(24%)	12(48%)	5(20%)	2(8%)	-0.03(-0.57 to 0.51)	15
Lack of awareness of when ADs are appropriate	8(32%)	10(40%)	6(24%)	1(4%)	0.18(-0.36 to 0.73)	16
Worry about the impact of stopping early on full-time research staff employment contracts	10(40%)	8(32%)	4(16%)	3(12%)	0.21(-0.34 to 0.76)	17
Unfamiliarity with key implementation resources such as validated statistical software	8(32%)	11(44%)	6(24%)	-	0.34(-0.22 to 0.90)	18
Fear of regulatory reluctance and jeopardising chances of obtaining regulatory approval due to the use of an AD	11(44%)	6(24%)	7(28%)	1(4%)	0.37(-0.19 to 0.93)	19
Inadequate data management support infrastructure for timely capturing, cleaning and transfer for decision making as part of the adaptation	12(48%)	6(24%)	4(16%)	3(12%)	0.39(-0.17 to 0.96)	20
Statistical complexities during implementation (such as analysis and reporting)	9(36%)	10(40%)	6(24%)	-	0.42(-0.15 to 0.98)	21
Lack of knowledge to use existing validated statistical software	9(36%)	11(44%)	4(16%)	1(4%)	0.42(-0.14 to 0.99)	22
Lack of statistical expertise	9(36%)	10(40%)	6(24%)	-	0.43(-0.13 to 0.99)	23
Lack of awareness of benefits of ADs	12(48%)	6(24%)	7(28%)	-	0.60(0.02 to 1.18)	24
Tension during early stopping decision making among key decision makers (such as IDMC and Funders)	15(60%)	7(28%)	3(12%)	-	1.32(0.66 to 1.98)	25

Previous negative experiences during implementation	19(76%)	3(12%)	2(8%)	1(4%)	1.79(1.04 to 2.53)	26
Previous negative experiences with ADs based on Funders/Reviewers comments	19(76%)	5(20%)	1(4%)	-	2.15(1.32 to 2.99)	27

Appendix 4.8: Supplementary summary data on private sector's perceptions of important barriers to ADs use in confirmatory trials

Barrier	Perceived importance				Relative importance parameter (95% CI)	Rank
	Not important	Somewhat important	Moderately important	Extremely important		
Lack of practical implementation knowledge	1(8%)	3(23%)	5(38%)	4(31%)	-1.44(-2.42 to -0.46)	1
Lack of time to support planning in relation to other competing conventional mainstream design priorities	1(8%)	6(46%)	2(15%)	4(31%)	-1.24(-2.20 to -0.27)	2
Practical complexities during trial conduct for successful implementation	1(8%)	3(23%)	6(46%)	3(23%)	-1.19(-2.16 to -0.23)	3
Inadequate data management support infrastructure for timely capturing, cleaning and transfer for decision making as part of the adaptation	3(23%)	5(38%)	2(15%)	3(23%)	-1.07(-2.03 to -0.12)	4
Lack of applied training to facilitate practical implementation	2(15%)	2(15%)	7(54%)	2(15%)	-0.98(-1.93 to -0.03)	5
Lack of practical experiences	1(8%)	4(31%)	6(46%)	2(15%)	-0.92(-1.87 to 0.02)	6
Insufficient access to case studies to facilitate practical learning	2(15%)	5(38%)	4(31%)	2(15%)	-0.78(-1.72 to 0.16)	7
Research team being more comfortable with the conventional mainstream designs compared to ADs	3(23%)	2(15%)	5(38%)	3(23%)	-0.74(-1.68 to 0.20)	8
Lack of awareness of which scope of ADs are acceptable in confirmatory trials	1(8%)	4(31%)	6(46%)	2(15%)	-0.64(-1.57 to 0.30)	9
Amount of work and effort required at the design or planning stage	4(31%)	4(31%)	2(15%)	3(23%)	-0.54(-1.47 to 0.40)	10
Fear of regulatory reluctance and jeopardising chances of obtaining regulatory approval due to the use of an AD †	5(38%)	1(8%)	3(23%)	3(23%)	-0.53(-1.46 to 0.40)	11
Difficulties in marketing ADs to key stakeholders in trials research (such as Collaborators, R& D and Regulators)	2(15%)	6(46%)	4(31%)	1(8%)	-0.51(-1.45 to 0.42)	13
Lack of awareness of when ADs are appropriate	2(15%)	5(38%)	4(31%)	2(15%)	-0.46(-1.39 to 0.47)	13
Difficulties in setting up acceptable upfront decision making criteria to guide the adaptation	2(15%)	5(38%)	2(15%)	4(31%)	-0.31(-1.24 to 0.62)	14
Statistical complexities during planning (such as simulation work)	3(23%)	6(46%)	2(15%)	2(15%)	-0.09(-1.03 to 0.85)	15
Lack of bridge funding required to support design work of time consuming and complex ADs †	6(46%)	3(23%)	2(15%)	1(8%)	-0.04(-0.98 to 0.90)	16
Difficulties outsourcing expertise to support ADs †	5(38%)	3(23%)	4(31%)	-	0.20(-0.76 to 1.16)	17
Tension during early stopping decision making among key decision makers (such as data monitoring committees and Sponsors/Funders)	4(31%)	6(46%)	1(8%)	2(15%)	0.34(-0.64 to 1.31)	18
Statistical complexities during implementation (such as analysis and reporting)	5(38%)	6(46%)	-	2(15%)	0.38(-0.60 to 1.36)	19
Lack of awareness of benefits of ADs	7(54%)	2(15%)	3(23%)	1(8%)	0.38(-0.60 to 1.370	20
Lack of knowledge to use existing validated statistical software	6(46%)	4(31%)	2(15%)	1(8%)	0.50(-0.50 to 1.50)	21

Unfamiliarity with key implementation resources such as validated statistical software †	6(46%)	3(23%)	3(23%)	1(8%)	0.50(-0.50 to 1.50)	22
Lack of motivational support from R&D to build an infrastructure to support ADs	6(46%)	5(38%)	2(15%)	-	0.65(-0.37 to 1.67)	23
Previous negative regulatory experiences with ADs such as based on regulatory comments or unsuccessful implementation †	5(38%)	3(23%)	2(15%)	-	0.70(-0.33 to 1.73)	24
Lack of general expertise around ADs at the trial planning stage	5(38%)	4(31%)	4(31%)	-	0.71(-0.32 to 1.74)	25
Complexity in deriving the cost of the proposed trial †	6(46%)	3(23%)	2(15%)	1(8%)	0.91(-0.16 to 1.99)	26
Previous negative experiences with ADs during implementation †	7(54%)	2(15%)	2(15%)	-	0.97(-0.12 to 2.05)	27
Insufficient financial support from R&D to build an infrastructure to support ADs	8(62%)	2(15%)	2(15%)	1(8%)	1.00(-0.09 to 2.10)	28
Lack of statistical expertise	8(62%)	2(15%)	3(23%)	-	1.09(-0.03 to 2.20)	29
Worry about the impact of stopping early on staff contracts †	10(77%)	2(15%)	-	-	3.16(1.01 to 5.31)	30

† Some respondents selected not applicable to their organisation

Appendix 4.9: Supplementary summary data on Public Funders' perceptions of important barriers to ADs use in confirmatory trials

Barrier	Perceive importance				Relative importance parameter (95% CI)	Rank
	Not important	Somewhat important	Moderately important	Extremely important		
Funding panel board members being more comfortable with traditional mainstream designs compared to ADs	7(11%)	17(27%)	21(33%)	19(30%)	-0.58 (-0.93 to -0.24)	1
Funding board generally being risk averse to fund complex ADs associated with high financial uncertainty	10(17%)	16(27%)	15(25%)	19(32%)	-0.45 (-0.81 to -0.10)	2
Decision making criteria to guide the adaptation not well described	7(11%)	18(29%)	26(41%)	12(19%)	-0.33 (-0.67 to 0.02)	3
Rationale for ADs not well explained in the grant application	5(8%)	25(40%)	18(29%)	15(24%)	-0.32 (-0.66 to 0.02)	4
Lack of expertise of Reviewers of ADs to help funding panel boards during grant review process	8(14%)	19(32%)	21(36%)	11(19%)	-0.22 (-0.57 to 0.14)	5
Lack of commissioning experience of ADs	8(14%)	17(30%)	23(40%)	9(16%)	-0.19 (-0.55 to 0.17)	6
The type AD proposed and its scope not well described in the grant application	8(13%)	20(32%)	23(37%)	11(18%)	-0.14 (-0.48 to 0.21)	7
Lack of awareness of which scope of ADs are acceptable in confirmatory trials	8(13%)	21(34%)	25(40%)	8(13%)	-0.03 (-0.38 to 0.31)	8
Lack of awareness of when ADs are appropriate	11(18%)	21(34%)	18(29%)	12(19%)	0.02 (-0.32 to 0.37)	9
Inadequate description of the costing scenarios of ADs in the grant application	7(11%)	25(41%)	22(36%)	7(11%)	0.04 (-0.31 to 0.39)	10
Lack of awareness of benefits of ADs	16(25%)	17(27%)	17(27%)	14(22%)	0.18 (-0.16 to 0.52)	11
Difficulties in drawing up flexible contractual agreements suitable for ADs	16(28%)	14(24%)	22(38%)	6(10%)	0.41 (0.05 to 0.77)	12
Tension during early stopping decision making of ADs among key decision makers	12(21%)	26(46%)	14(25%)	4(7%)	0.63 (0.25 to 1.00)	13
Negative attitudes towards ADs among some funding panel board members	22(37%)	22(37%)	8(13%)	8(13%)	0.97 (0.60 to 1.34)	14

Note: The number of participants in the denominator varies due to the exclusion of respondents who were not able to answer certain items.

Appendix 4.10: Supplementary summary data on cross-sector perceptions of concerns towards ADs use in confirmatory trials

Concern	Perceived level of concern					Relative concern parameter (95% CI)	Rank
	Not at all	Slightly	Somewhat	Moderately	Extremely		
UK CTUs							
Efficacy early stopping of trials	7(28%)	4(16%)	4(16%)	4(16%)	6(24%)	-0.35(-0.76 to 0.07)	1
Robustness of AD methodology to influence policy decision making when trials are stopped early	2(8%)	8(32%)	8(32%)	5(20%)	2(8%)	-0.26(-0.67 to 0.15)	2
Non-inferiority early stopping of trials	9(36%)	2(8%)	5(20%)	4(16%)	5(20%)	-0.15(-0.56 to 0.26)	3
Fear of introducing operational bias	3(12%)	11(44%)	4(16%)	4(16%)	3(12%)	-0.10(-0.51 to 0.31)	4
Impact of ADs on secondary trial objectives when trials are stopped early	2(8%)	12(48%)	6(24%)	3(12%)	2(8%)	-0.01(-0.43 to 0.40)	5
Acceptability of the findings from ADs by the research community or Regulators to change practice	5(20%)	9(36%)	6(24%)	3(12%)	2(8%)	0.13(-0.29 to 0.55)	6
Potential change in the population during the course of an adaptive trial and its impact on interpretation of the findings	7(28%)	9(36%)	2(8%)	5(20%)	2(8%)	0.23(-0.20 to 0.66)	7
Futility stopping of trials for futility	10(40%)	4(16%)	8(32%)	2(8%)	1(4%)	0.51(0.07 to 0.95)	8
Private sector							
Early stopping of trials for non-inferiority	2(15%)	3(23%)	3(23%)	4(31%)	1(8%)	-0.39(-0.99 to 0.20)	1
Impact of ADs on secondary trial objectives when trials are stopped early	2(15%)	5(38%)	1(8%)	4(31%)	1(8%)	-0.24(-0.84 to 0.37)	2
Fear of introducing operational bias	1(8%)	8(62%)	-	2(15%)	2(15%)	-0.22(-0.82 to 0.39)	3
Early stopping of trials for efficacy	3(23%)	3(23%)	2(15%)	5(38%)	-	-0.16(-0.76 to 0.45)	4
Potential change in the population during the course of an adaptive trial and its impact on interpretation of the findings	3(23%)	4(31%)	2(15%)	3(23%)	1(8%)	-0.09(-0.70 to 0.52)	5
Robustness of AD methodology to influence policy decision making when trials are stopped early	3(23%)	4(31%)	2(15%)	4(31%)	-	-0.00(-0.62 to 0.61)	6
Acceptability of the findings from ADs by the research community or Regulators in order to change practice	3(23%)	5(38%)	2(15%)	3(23%)	-	0.22(-0.40 to 0.85)	7
Early stopping of trials for futility	7(54%)	1(8%)	3(23%)	2(15%)	-	0.87(0.18 to 1.55)	8
Public Funders							
Robustness of AD methodology to influence policy decision making when trials are stopped early	7(10%)	8(12%)	23(34%)	15(22%)	9(13%)	-0.46(-0.73 to -0.19)	1
Acceptability of the findings from ADs by the research community or Regulators in order to change practice	5(7%)	12(18%)	18(26%)	19(28%)	7(10%)	-0.43(-0.70 to -0.16)	2
Impact of ADs on secondary trial objectives when trials are stopped early	7(10%)	17(25%)	20(29%)	14(21%)	4(6%)	-0.09(-0.35 to 0.18)	3
Non-inferiority early stopping of trials	13(19%)	15(22%)	13(19%)	11(16%)	9(13%)	-0.07(-0.34 to 0.20)	4

Fear of introducing operational bias	10(15%)	14(21%)	22(32%)	11(16%)	4(6%)	0.04(-0.23 to 0.31)	5
Potential change in the population during the course of an adaptive trial and its impact on interpretation of the findings	10(15%)	21(31%)	17(25%)	9(13%)	4(6%)	0.17(-0.11 to 0.44)	6
Efficacy early stopping of trials	13(21%)	20(33%)	14(23%)	10(16%)	4(7%)	0.25(-0.02 to 0.52)	7
Futility early stopping of trials	22(32%)	17(25%)	10(15%)	7(10%)	5(7%)	0.58(0.29 to 0.86)	8

Appendix 5.1: Adaptive design-related search terms and strategy

1. **Original search terms**
    - adaptive
    - adaptive design
    - adaptive dose finding
    - adaptive dose escalation
    - adaptive randomisation
    - adaptive treatment switching
    - Bayesian adaptive
    - biomarker adaptive
    - continuous reassessment
    - drop the loser
    - enrichment
    - flexible design
    - group sequential
    - interim analysis
    - internal pilot
    - MAMS
    - multi-stage
    - play the winner
    - pick the winner
    - seamless
    - seamless II/III or seamless 2/3
    - sample size reestimation or sample size re-estimation

2. **Other possible search terms - some of these might be captured in the above or too general, but were used in the literature**.
    - accumulating data
    - active learning
    - adaptive allocation
    - adaptive learning
    - adaptive sample size methods
    - bayesian
    - dose finding
    - dose selection
    - futility
    - go/no go
    - internal pilot
    - novel
    - pragmatic design
    - preplanned
    - reanalysis
    - response-adaptive
    - sample size re-assessment
    - sample size review
    - stopping rule
    - two stage adaptive design
    - vanguard phase

3. **Independent search terms were applied to the ClinicalTrials.gov during the scoping exercise.**

Search Terms	Number of identified trials via ClinicalTrials.gov	Inclusion Decision Comments
Novel	18492	No (too sensitive, impractical and result in too many irrelevant trials)
Dose finding	575	No (use 'adaptive dose finding' instead)
Adaptive	291	Yes
Interim	202	Yes
Interim analysis	103	No (captured using the term 'interim')
Interim analyses	103	No (captured using the term 'interim')
MAMS	81	Yes
Dose selection	41	Yes
Bayesian	35	Yes
Adaptive design	33	No (captured using the term 'adaptive')
SSR	27	No - not relevant
Futility	24	Yes
Enrichment	20	Yes
Stopping rule	19	Yes
Seamless	13	Yes
Group sequential	12	Yes (with no speech marks)
Go/no go	11	Yes
Preplanned	9	Yes
Bayesian adaptive	8	No (captured using the term 'bayesian')
Adaptive randomisation	7	No (captured using the term 'adaptive')
Multi-stage/ multiple stage/ Multiple arm	6	Yes
Active learning	5	Yes
Adaptive dose finding	3	No (captured using the term 'adaptive')
Accumulating data	3	Yes
Response adaptive	2	No (captured using the term 'adaptive')
Continuous reassessment	2	Yes
Reanalysis	1	Yes
Pick the winner	1	Yes
Internal pilot	1	Yes
Drop the loser	1	Yes
Dose escalation	1248	Yes (with no speech marks)
Adaptive allocation	1	No (captured using the term 'adaptive')
Vanguard phase	0	No
Two stage adaptive design	0	No
Seamless II/III or seamless 2/3	0	No
Sample size reassessment	0	No
Sample size review	0	No
Pragmatic design	0	No
Play the winner	0	No
Flexible design	0	No
Biomarker adaptive	0	No
Adaptive treatment switching	0	No

Adaptive sample size methods	0	No
Adaptive learning	0	No
Sample size adjustment	38	Yes
Sample size re-estimation	2	Yes
Sample size modification	0	No (in speech marks)

4. Final Search terms

(Adaptive) OR (Interim) OR (Dose selection) OR (Bayesian) OR (Futility) OR (Enrichment) OR (Stopping rule) OR (Seamless) OR (Group sequential) OR (Go/no go) OR (Preplanned) OR (MAMS) OR (Multi-stage) OR (Multiple stage) OR (Multiple arm) OR (Active learning) OR (Accumulating data) OR (Continuous reassessment) OR (Reanalysis) OR (Pick the winner) OR (Internal pilot) OR (Drop the loser) OR (Dose escalation) OR (Sample size adjustment) OR (Sample size re-estimation)

Appendix 5.2: A list of case studies of confirmatory adaptive designs found in the literature

Trial registration number	Type of adaptation	Additional information (Title and/or References)
NCT01225276	Seamless 2/3 AD 'Pick-the-winner'	Safety and Efficacy Study of Three Different Dosages of NewGam in Patients With CIDP (POINT)
NCT00518687	Group sequential	Efficacy, Immunogenicity, and Safety of a Single Dose of V710 in Adult Patients Scheduled for Cardiothoracic Surgery (V710-003 AM2) (Fowler et al., 2013)
ISRCTN29161170	Group sequential	CRISP trial (Rogers et al., 2014)
NCT00059306	Group sequential	Secondary Prevention of Small Subcortical Strokes Trial (SPS3) study (Benavente et al., 2011)
NCT00095576	Group sequential	Investigation of V520 in an HIV Vaccine Proof-of-Concept Study (V520-023) (Buchbinder et al., 2008)
ISRCTN38366450	Group sequential	BALTI-2 trial (Gates et al., 2013)
NCT00047632	Group sequential	Safety and Efficacy of Interferon Gamma-1b Plus Chemotherapy for Ovarian and Peritoneal Cancer
NCT00574275	Group sequential	Aflibercept Compared to Placebo in Term of Efficacy in Patients Treated With Gemcitabine for Metastatic Pancreatic Cancer (VANILLA)
NCT00283842	Group sequential Treatment selection	Study Evaluating Desvenlafaxine Succinate Sustained-release (DVS SR) in Adult Outpatients With Pain Associated With Diabetic Peripheral Neuropathy (Allen et al., 2014)
NCT01209702	Seamless AD	A Study of RoActemra/Actemra (Tocilizumab) in Patients With Ankylosing Spondylitis Who Have Failed Treatment With NSAIDs
NCT00428597	Group sequential	A Study Of Sunitinib Compared To Placebo For Patients With Advanced Pancreatic Islet Cell Tumors (Cheng et al., 2013)
NCT00242879	Seamless 2/3 AD Dose selection	A Dose Ranging Study Of GW640385 Boosted With Ritonavir (Rtv) In Comparison To A RTV-Boosted Protease Inhibitor (PI) In HIV-1 Infected PI-Experienced Adults
NCT01566630	Seamless AD Dose selection	Safety and Efficacy of RLX030 in Pregnant Women With Pre- Eclampsia
NCT00612742	SSR	Safety and Efficacy of LibiGel® for Treatment of Hypoactive Sexual Desire Disorder in Postmenopausal Women (BLOOM) (White et al., 2012)
NCT00164736	Group sequential Treatment selection	Breastfeeding, Antiretroviral, and Nutrition Study (van der Horst et al., 2009)
NCT00463567	Seamless 2/3 AD Dose selection	26 Week Efficacy, Safety and Tolerability Study of Indacaterol in Patients With Chronic Obstructive Pulmonary Disease (COPD) (Donohue et al., 2010)

NCT00860288	Seamless 2/3 AD Dose selection	Efficacy and Long-Term Safety of Vildagliptin as Add-on Therapy to Metformin in Patients With Type 2 Diabetes
NCT01061736	Operational Seamless 2/3 AD Dose selection	Evaluation of SAR153191(REGN88)(Sarilumab) on Top of Methotrexate in Rheumatoid Arthritis Patients (RA-MOBILITY)
NCT00098293	Group sequential Treatment selection	Trial of Maraviroc (UK-427,857) in Combination With Zidovudine/Lamivudine Versus Efavirenz in Combination With Zidovudine/Lamivudine (MERIT)
NCT00666224	Group sequential	Evaluate Early Glatiramer Acetate Treatment in Delaying Conversion to Clinically Definite Multiple Sclerosis of Subjects Presenting With Clinically Isolated Syndrome (PreCISe)
NCT01069939	Group sequential	Comparative Efficacy & Safety Study of D961H Versus Placebo for the Prevention of Gastric and Duodenal Ulcers With Low-dose Aspirin (Sugano et al., 2014)
NCT00677807	Seamless AD	Safety, Tolerability and Efficacy of Indacaterol in Patients With Moderate-to-severe Chronic Obstructive Pulmonary Disease (COPD) (Chapman et al., 2011)
NCT00594399	Seamless AD	Veterans Enhanced Fitness Study
NCT01166542	Two-stage AD Treatment selection	Efficacy Study of REOLYSIN® in Combination With Paclitaxel and Carboplatin in Platinum-Refractory Head and Neck Cancers
NCT01149655	Group sequential Treatment selection	Efficacy & Safety Study of Oral Aripiprazole in Adolescents With Schizophrenia (ATTAIN 266)
NCT01497938	SSR	Outpatient Study to Evaluate Safety and Effectiveness of the Low Glucose Suspend Feature (ASPIRE) (Klonoff et al., 2013)
NCT00740051	Group sequential	A Randomised, db, Placebo-controlled Study of BI 1356 for 18 Weeks Followed by a 34 Week Double-blind Extension Period (Placebo Patients Switched to Glimepiride) in Type 2 Diabetic Patients for Whom Treatment With Metformin is Inappropriate
NCT00450580	Group sequential	HIV-1 Infection Study of Once a Day Versus Twice a Day Protease Inhibitor in Antiretroviral Treatment Naive Adults (Hughes et al.)
NCT00260676	Group sequential	Protective Ventilatory Strategy in Potential Organ Donors (Mascia et al., 2010)
NCT01328938	Seamless 2/3 AD Treatment selection	GCPGC in Chemotherapy-induced Neutropenia
NCT00874419	Group sequential	Erlotinib Versus Gemcitabine/Carboplatin in Chemo-naive Stage IIIB/IV Non-Small Cell Lung Cancer Patients With Epidermal Growth Factor Receptor (EGFR) Exon 19 or 21 Mutation (ML20981)
NCT00321178	Group sequential	BURULICO Drug Trial Study Protocol: RCT SR8/SR4+CR4, GHANA
NCT01096082	Seamless 2/3 AD	Safety and Efficacy of Lithium Carbonate in Patients With Spinocerebellar Ataxia Type 3
NCT00490139	Group sequential Treatment selection	ALTTO (Adjuvant Lapatinib And/Or Trastuzumab Treatment Optimisation) Study; BIG 2-06/N063D

NCT00324805	Group sequential	Chemotherapy With or Without Bevacizumab in Treating Patients With Stage IB, Stage II, or Stage IIIA Non-Small Lung Cancer That Was Removed By Surgery
NCT01694836	Group sequential	Depigoid Birch 5000 Longterm Study in Adults and Adolescents
NCT01182441	SSR	Evaluation of the WATCHMAN LAA Closure Device in Patients With Atrial Fibrillation Versus Long Term Warfarin Therapy (PREVAIL)
NCT01002417	Operational Seamless 2/3 AD Dose selection	MCS in the Treatment of Lower Urinary Tract Symptoms (MCS_LUTS)
ISRCTN47823388	Seamless SSR	Triple Antiplatelets for Reducing Dependency after Ischaemic Stroke (TARDIS)
ISRCTN52968807	Group sequential	Persephone: duration of herceptin with chemotherapy 6 versus 12 months
ISRCTN01151335	Group sequential	Pressure RElieving Support SUrfaces: a Randomised Evaluation 2 (PRESSURE 2)
NCT01905657	Seamless 2/3 AD Group sequential Treatment selection	Study of Two Doses of MK-3475 (Pembrolizumab) Versus Docetaxel in Previously-Treated Participants With Non-Small Cell Lung Cancer (MK-3475-010/KEYNOTE-010)
NCT01852110	Seamless 2/3 AD Dose selection	Efficacy and Safety of MK-7622 as Adjunct Therapy to Donepezil in Participants With Alzheimer's Disease (MK-7622-012)
ISRCTN 4911786	Seamless 2/3 AD Treatment selection	Physiotherapy Rehabilitation for Osteoporotic Vertebral Fracture (PROVE trial)
NCT01812369	Group sequential	Perioperative Chemotherapy for Patients With Locally Advanced Bladder Cancer (VESPER)
NCT01641939	Seamless 2/3 AD Dose selection	A Study of Trastuzumab Emtansine Versus Taxane in Patients With Advanced Gastric Cancer
NCT01735669	Group sequential	Open Randomized Controlled Trial to Evaluate the Efficacy and Safety of Remifentanil Versus Nitrous Oxide in External Cephalic Version at Term in Singleton Pregnancy in Breech Presentation (REMIVER)
NCT01091636	Seamless 2/3 AD Group sequential	Intraoperative Hyperthermic Intraperitoneal Chemotherapy With Ovarian Cancer
NCT01641016	Operational seamless 2/3 AD	Short-cycle therapy (SCT) (5 days on/2 days off) in young people with chronic human immunodeficiency virus (HIV) infection: an open, randomised, parallel group, multicentre phase II/III trial; BREATHER (PENTA 16)
NCT01222559	Group sequential	Efficacy and Safety Study of co.Don Chondrosphere to Treat Cartilage Defects
NCT01908192	Seamless 2/3 AD SSR	Adaptive Phase II Study to Evaluate the Safety & Efficacy of Sodium Benzoate as an Add-on Treatment for Schizophrenia in Adolescents

ISRCTN79705874	SSR	Debt counselling for depression in primary care (Decoder)
NCT01752985	Seamless Dose selection	Study to Evaluate the Effects of BMS-813160 on Protein Loss in the Urine of Subjects With Type 2 Diabetes and Diabetic Kidney Disease
NCT01998958	Seamless Dose selection	A Study to Evaluate the Safety and Efficacy of Intranasal Esketamine in Treatment-resistant Depression (SYNAPSE)
NCT00176852	Seamless Treatment selection	Stem Cell Transplant for Hemoglobinopathy
ISRCTN88609453	Group sequential	Glycerine Trinitrate for Retained Placenta (GOT-IT)
NCT00532194	Seamless	An RCT of Concurrent and Maintenance Cediranib in Women With Platinum-sensitive Relapsed Ovarian Cancer (ICON6)

Appendix 6.1: Summary data of compliance in the reporting of general CONSORT 2010 checklist items

CONSORT checklist item	Completeness in reporting of general CONSORT items					
	Partial/Complete /Not applicable	Absent	Complete	Partly complete	Cannot assess	Not Applicable
(#1a) Randomised trial in title	53(78%)	15(22%)	43(63%)	10(15%)	-	-
(#1b) Structured summary	67(99%)	1(1%)	63(93%)	4(6%)	-	-
(#2a) Background and rationale	68(100%)	-	68(100%)	-	-	-
(#2b) Objectives/hypotheses	58(85%)	10(15%)	55(81%)	3(4%)	-	-
(#3a) Trial design and allocation ratio	67(99%)	1(1%)	12(18%)	55(81%)	-	-
(#3b) Changes to methods	26(38%)	8(12%)	17(25%)	4(6%)	34(50%)	5(7%)
(#4a) Eligibility criteria	67(99%)	1(1%)	66(97%)	1(1%)	-	-
(#4b Settings/Locations	65(96%)	3(4%)	39(57%)	26(38%)	-	-
(#5) Interventions	68(100%)	-	62(91%)	6(9%)	-	-
(#6a) Predefined outcomes	68(100%)	-	56(82%)	12(18%)	-	-
(#6b) Changes to outcomes	23(34%)	5(7%)	4(6%)	-	45(66%)	14(21%)
(#7a) Determining sample size	66(97%)	2(3%)	66(97%)	-	-	-
(#7b) Explain interim analysis and stopping	62(91%)	6(9%)	35(51%)	27(40%)	-	-
(#8a) Randomisation methods	36(52%)	32(47%)	35(51%)	1(1%)	-	
(#8b) Type of randomisation	58(85%)	10(15%)	57(84%)	1(1%)	-	-
(#9) Allocation concealment mechanism	18(26%)	50(74%)	15(22%)	3(4%)	-	-
(#10) Randomisation implementation	28(41%)	40(59%)	8(12%)	20(29%)	-	-
(#11a) Blinding	51(91%)	6(9%)	11(16%)	16(24%)	-	35(51%)
(#11b) Similarity of interventions	65(96%)	3(4%)	12(18%)	5(7%)	-	48(71%)
(#12a) Statistical methods	66(97%)	2(3%)	54(79%)	12(18%)	-	-
(#12b) Additional analyses	39(57%)	29(43%)	19(28%)	4(6%)	-	16(24%)
(#13a) Participants flow	65(96%)	3(4%)	58(85%)	7(10%)	-	-
(#13b) Losses and exclusions	63(93%)	5(7%)	55(81%)	8(12%)	-	-
(#14a) Recruitment/Follow up dates	66(97%)	2(3%)	56(82%)	10(15%)	-	-
(#14b) Why stopped early	68(100%)	-	46(68%)	-	-	22(32%)
(#15) Baseline data	66(97%)	2(3%)	66(97%)	-	-	-

(#16) Numbers analysed	61(90%)	7(10%)	57(84%)	4(6%)	-	-
(#17a) Outcomes and estimation	66(97%)	2(3%)	36(53%)	30(44%)	-	-
(#17b) Binary outcome presentation	58(85%)	10(15%)	3(4%)	-	-	55(81%)
(#18) Ancillary analyses	57(89%)	11(16%)	28(41%)	12(18%)	-	17(25%)
(#19) Harms	65(96%)	3(4%)	63(93%)	2(3%)	-	-
(#20) Limitations	48(72%)	19(28%)	37(54%)	12(18%)	-	-
(#21) Generalisability	67(99%)	1(1%)	67(99%)	-	-	-
(#22) Interpretation	68(100%)	-	65(96%)	3(4%)	-	-
(#23) Registration	42(62%)	26(38%)	38(56%)	4(6%)	-	-
(#24) Full trial protocol	15(22%)	53(78%)	15(22%)	-	-	-
(#25) Funding and other support	62(91%)	6(9%)	61(90%)	1(1%)	-	-

Appendix 7.1: Unrestricted SSR results for 3CPO trial

Figure (7.1-A) shows the observed mortality for the pooled data and the SOT arm compared to mortality assumed in the SOT at the design stage. The research team assumed a 15% and 9% mortality within 7 days (a pooled mortality of 12%), in the SOT and CPAP or NIPPV arms. Assuming that a total recruitment of at least 300 participants (100 per arm) is required before performing a SSR, the observed median (IQR) pooled mortality was 8.6% (8.3% to 8.9%), ranging from 7.2% to 9.7%. Thus, the median (IQR) overestimation in the pooled mortality was 3.5% (3.1% to 3.7%), with a maximum of 4.8%. The pooled mortality estimates appear to be unstable and inconsistent during the initial recruitment of approximately the first 200 participants. The estimation of SOT mortality is done only using participants in the control arm.

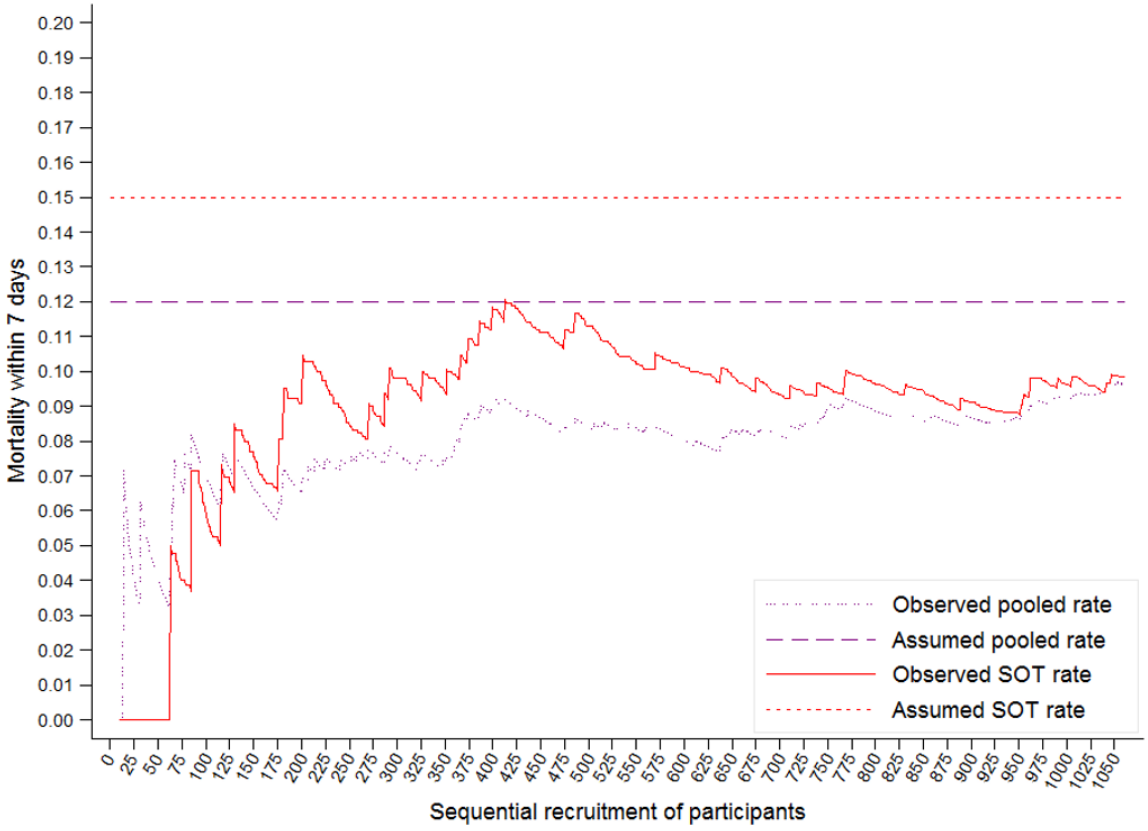

Figure 7.1-A. Uncertainty around assumed mortality for 3CPO trial.

Figure 7.1-B displays the re-estimated sample sizes. Here, the pooled mortality at the interim was used by invoking an approximate formula on equation (2:4). A 2 on the denominator of equation (2:4) is replaced by 1.5 since the allocation ratio is technically 2 to 1 (CPAP or NIPPV to SOT) for the primary endpoint. Assuming a 12% pooled mortality, 6% absolute difference as clinically relevant to detect, 80% power and a 5% two-sided

type I error; approximately 1,036 participants (~345 per arm) would be required. For illustrative purposes, the planned sample sizes using the SOT and pooled mortality using versions of equations (2:3) and (2:4), respectively are presented in Figure (7.1-B). As evident, the re-estimated sample sizes are much lower than the planned because the pooled mortality was markedly overestimated.

The total re-estimated sample size has a median (IQR) of 768 (746 to 795), with a minimum and maximum of 654 and 860, respectively. This is assuming that SSR is performed based on the primary outcome data of at least 300 participants in total (100 per group). Similarly, the median (IQR) overestimation in the sample size is 269 (242 to 291), assuming the study preserves 80% power as planned. However, if the research team (imagining this was prospectively undertaken) chose to continue with the trial as planned, the trial would have approximately 99% power based on re-estimated pooled mortality.

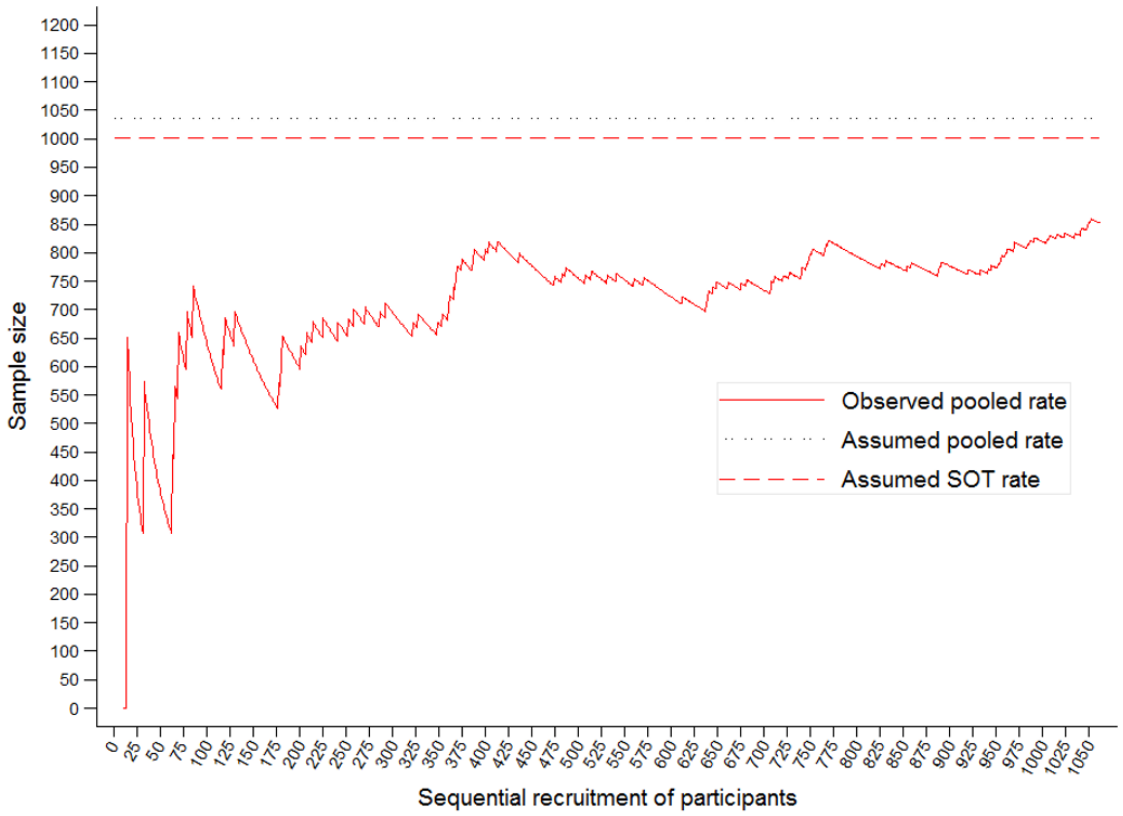

Figure 7.1-B. Pattern of re-estimated total sample size for 3CPO trial

One of the fundamental question raised during this retrospective application of SSR is whether the assumed 6% absolute difference in mortality is still a clinically relevant effect to detect in view of overestimated pooled or SOT mortality. The results presented so far assume that the 6% absolute mortality difference is a fixed

clinically relevant effect regardless of the underlying pooled or control mortality rate. However, this may be unrealistic when the observed pooled mortality is close to the boundary space of proportions. For instance, it is questionable whether the assumed change of 6% from 15% to 9% is the same as from 9% to 3% under the observed mortality rate. Some investigators may be willing to consider an effect size that remain constant regardless of the observed underlying mortality. For an assumed constant OR 0.56 (equivalent to an RR of 0.60), Figure (7.1-C) illustrates the changing pattern of the RRs estimated based on observed pooled and SOT mortality rates.

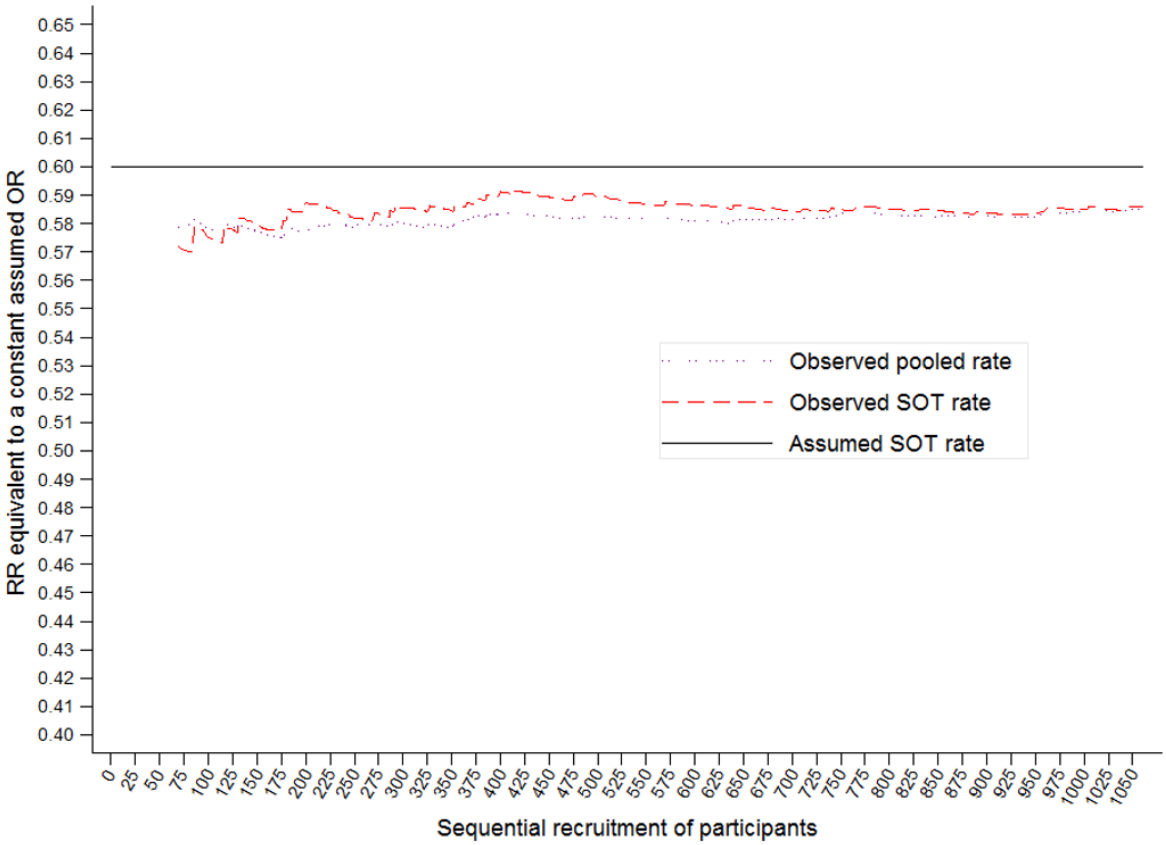

Figure 7.1-C. Changes in Risk Ratio for an assumed constant Odds Ratio for 3CPO trial.

In this case, if the research team seek to detect a constant effect on an OR scale regardless of uncertainty in the pooled or control mortality rate, this would result in a marked increase in sample size, approximately by a mean (IQR) of 486 (411 to 531) participants. This means that the OR sought remained the same at 0.56, but the observed RR slightly changed from 0.60 to 0.58. This is because the assumed risk difference of 6% (15% to 9%) changed to approximately 3.6% (8.6% to 5%). The increase in the mean sample size under the OR of 0.56 is due to the increase in the OR variance when the event rates are rare.