Adjudicating Outcomes in Stroke Trials

Peter John Godolphin MSc. BSc. (Hons)

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

July 2019

Table of Contents

Index of tables 11

Index of figures 16

List of abbreviations 18

Abstract 20

Acknowledgements 23

The Adjudicating Outcomes in Stroke Trials (AOST) Collaboration 24

Publications 27

Peer-reviewed journal articles 27

Conference oral presentations 28

Conference poster presentations 28

Chapter 1 – Introduction 30

1.1 Background 31

1.2 Stroke 31

1.3 Outcome Assessment in Clinical Trials 35

1.3.1 Detection bias 36

1.3.2 Performance bias 37

1.3.3 Non-differential misclassification 38

1.4 Central Adjudication in Clinical Trials 39

1.5 Methods of Author Contact to Receive Unpublished Trial Data for a Systematic Review 46

1.6 Methods of Adjudication in Clinical Trials 49

1.7 Adjudication in Cardiovascular Trials..... 53

1.8 Evidence of Adjudication in Stroke Trials 56

1.9 Research Question..... 60

1.10 Aims of this thesis..... 61

Chapter 2 – Methods 63

2.1 Introduction..... 63

2.2 Systematic Reviews 63

2.2.1 Search Strategies 63

2.2.2 Electronic Databases..... 64

2.2.3 Identifying Grey Literature 65

2.2.4 Handsearching and checking citations of included articles..... 66

2.2.5 Languages to include 66

2.2.6 Managing records and de-duplication..... 67

2.2.7 Screening records 67

2.2.8 Data extraction 67

2.2.9 Narrative synthesis 68

2.10 Summary of systematic review methods 69

2.3 Meta-analysis 69

2.3.1 Fixed effect meta-analysis 70

2.3.1.1 Inverse variance method 71

2.3.1.2 Mantel-Haenszel method 72

2.3.1.3 Peto's method for odds ratios 73

2.3.2 Assessing heterogeneity 74

2.3.3 Random effects meta-analysis 75

2.3.4 Exploring heterogeneity 77

2.3.5 Publication bias 77

2.3.6 Methodological meta-analysis 79

2.3.7 Summary of meta-analysis methods 80

2.4 Measures of agreement 81

2.4.1 Percent agreement 82

2.4.2 Kappa 85

2.4.3 Intraclass correlation coefficient 91

2.4.4 Limits of agreement 93

2.4.5 Summary of measure of agreement methods 93

2.5 Methods to establish whether treatment effect estimates differ for outcomes assessed by different raters 94

2.5.1 Test of homogeneity 94

2.5.2 Ratio of treatment effects 94

2.5.3 Summary of methods to establish whether treatment effect estimates differ for outcomes assessed by different raters 96

Chapter 3 – Outcome assessment by central adjudicators versus site investigators in randomised stroke trials: A systematic review and meta-analysis 97

Abstract 99

3.1 Introduction 100

3.2 Methods 101

3.2.1 Review protocol 101

3.2.2 Eligibility Criteria 101

3.2.3 Defining the Search Strategy 102

3.2.3.1 Scoping review 103

3.2.3.2 Final Search Strategy 106

3.2.4 Study Outcomes 106

3.2.5 Selection of Studies 106

3.2.5.1 Checking screening process 1 (titles and abstracts) 107

3.2.5.2 Checking screening process 2 (full texts) 108

3.2.6 Data Extraction 108

3.2.6.1 Designing the data extraction form 109

3.2.6.2 Double data extraction 110

3.2.6.3 Obtaining missing data 110

3.2.7 Assessment of risk of bias 110

3.2.8 Statistical Analysis 111

3.3 Results 114

3.3.1 Search Results 114

3.3.2 Characteristics of the potentially eligible trials 116

3.3.3 Characteristics of adjudication in included trials 116

3.3.4 Risk of bias 123

3.3.5 Impact of adjudication on primary treatment effect 128

3.3.6 Impact of central adjudication on the number of events reported 128

3.4 Discussion 139

Chapter 4 – A nested randomised trial investigating short email with attachment versus long email without attachment when contacting authors to request unpublished data for a systematic review 143

Abstract 144

4.1 Introduction 145

4.2 Methods 146

4.2.1 Rationale 146

4.2.2 CONSORT 147

4.2.3 Participants 147

4.2.4 Trial Design 149

4.2.5 Randomisation 150

4.2.6 Plan for eliciting response 151

4.2.7 Outcome Measures 152

4.2.8 Statistical Analysis 153

4.3 Results 155

4.3.1 Characteristics of included studies 155

4.3.2 Impact of email format on response rate 157

4.3.3 Impact of email format on response time 158

4.3.4 First response outcome and eventual collaboration 159

4.4 Discussion 161

Chapter 5 – Outcome assessment by central adjudicators in randomised stroke trials: simulation of differential and non-differential misclassification 166

Abstract 167

5.1 Introduction 168

5.2 Methods 169

5.2.1 Differential misclassification using real trial data 169

5.2.1.1 Data collection 169

5.2.1.2 Misclassification for binary outcomes 170

5.2.1.3 Misclassification for ordinal outcomes 171

5.2.1.4 The proportion of misclassification necessary to alter the estimated treatment effect 172

5.2.1.5 Statistical analysis 172

5.2.2 Non-differential misclassification using hypothetical trial data .. 173

5.2.2.1 Data generation 174

5.2.2.2 Characteristics to vary 174

5.2.2.3 Misclassifying events 175

5.3. Results: 176

5.3.1 Differential misclassification using real trial data 176

5.3.2 Non-differential misclassification using hypothetical trial data .. 190

5.4. Discussion 195

Chapter 6 – Cost-benefit of central adjudication in randomised stroke trials 199

Abstract 200

6.1 Introduction 201

6.2 Methods 202

6.2.1 Data Collection 202

6.2.2 Benefit of Adjudication 202

6.2.3 Cost of Adjudication 202

6.2.4 Statistical Analysis 203

6.3 Results 205

6.4 Discussion 210

Chapter 7 – Central adjudication of Serious Adverse Events in the Efficacy of Nitric Oxide in Stroke (ENOS) Trial 212

Abstract 213

7.1 Introduction 214

7.2 Methods 215

7.2.1 Efficacy of Nitric Oxide in Stroke (ENOS) Trial 215

7.2.1.1 Summary of ENOS Methods 215

7.2.1.2 Summary of ENOS Results 215

7.2.1.3 Adjudication of Investigator reported Serious Adverse Events in ENOS 216

7.2.2 Statistical Methods 217

7.3 Results 220

7.3.1 Agreement on SAE diagnosis 220

7.3.2 Agreement on SAE classification 224

7.3.3 Agreement on likely causality of SAEs 229

7.3.4 Time taken to adjudicate 232

7.3.5 Effect of central adjudication on ENOS safety results 233

7.4 Discussion 237

Chapter 8 – Discussion 241

8.1 Summary of findings 241

8.2 Findings in the context of previous research 245

8.3 Limitations of this research 246

8.4 Direction of future work 248

8.5 Reflection on this thesis 251

8.6 Conclusions 253

References 254

Appendix A – PRISMA Checklist 283

Appendix B – Full search strategy 286

Appendix C – Data Extraction Form 289

Appendix D – Data Extraction Manual 293

Appendix E – Missing Information Log 296

Appendix F – Data Sharing and Use Agreement 297

Appendix G: CONSORT Checklist 302

Appendix H: Short Invitation email 306

Appendix I: Long invitation email 307

Appendix J: Adjudicating Outcomes in Stroke Trials Protocol ... 310

Appendix K – Cost of Adjudication Questionnaire 324

Index of tables

Table 1 Members of the AOST Collaboration 25

Table 2 Summary of contributions of AOST Collaborators..... 26

Table 3 Example binary outcome..... 39

Table 4 Example binary outcome after non-differential misclassification 39

Table 5 Example binary outcome after differential misclassification ... 39

Table 6 Example binary outcome after differential misclassification ... 39

Table 7 Modified Rankin Scale (mRS) 58

Table 8 Cell counts for the outcome of stroke in trial i 70

Table 9 Agreement between central adjudicators and site investigators on stroke..... 82

Table 10 Agreement between central adjudicators and site investigators on functional outcome where mRS=2..... 83

Table 11 Agreement between central adjudicators and site investigators in a hypothetical trial of 1000 participants..... 84

Table 12 Indices of agreement between central adjudicators and site investigators on stroke for hypothetical trial..... 85

Table 13 Interpretation of the kappa statistic (from Landis and Koch[161]) 86

Table 14 Joint proportions of agreement between central adjudicators and site investigators on a categorical outcome with k categories..... 87

Table 15 Agreement between central adjudicators and site investigators on functional outcome in a hypothetical trial of 100 participants..... 88

Table 16 Mean square expectations for mixed-model and estimate of ρ 92

Table 17 Assessment of stroke by central adjudicators and site investigators 95

Table 18 Example search strategies 103

Table 19 Scoping review of a random sample of 100 records 104

Table 20 Estimation of workload 105

Table 21 Characteristics of included trials and potentially eligible trials 117

Table 22 Summary of included trials 119

Table 23 Characteristics of adjudication in included trials 122

Table 24 Risk of bias of included trials 124

Table 25 Characteristics of included trials 156

Table 26 Author position in main results paper for primary and secondary contacts of included trials 157

Table 27 Primary outcome: response from contacted authors 157

Table 28 Sensitivity analysis. Logistic regression model for primary outcome adjusted for variables with baseline imbalance 158

Table 29 Time to response (days) 158

Table 30 Secondary outcomes 160

Table 31 Summary of parameters used in the simulation of non-differential misclassification 175

Table 32 Summary of trials included in simulation study 177

Table 33 Agreement between central adjudicators and site investigators using original trial data 179

Table 34 Functional outcome by treatment group using central adjudicator and site investigator data in REVASCAT 180
Table 35 Agreement on functional outcome between central adjudicators and site investigators in REVASCAT..... 181
Table 36 Functional outcome by treatment group using central adjudicator and site investigator data in TARDIS 182
Table 37 Agreement on incidence and severity of stroke between central adjudicators and site investigators in TARDIS 183
Table 38 Number and proportion of participants required to be differentially misclassified to alter estimated treatment effect..... 184
Table 39 Agreement between central adjudicators and site investigators on primary outcome after differential misclassification . 185
Table 40 Functional outcome by treatment group using central adjudicator and misclassified site investigator data in REVASCAT ... 186
Table 41 Agreement on functional outcome between central adjudicators and example misclassified site investigators in REVASCAT (data from one of 1000 simulations) 187
Table 42 Functional outcome by treatment group using central adjudicator and misclassified site investigator data in TARDIS 188
Table 43 Example agreement on incidence and severity of stroke between central adjudicators and misclassified site investigators in TARDIS (data from one of 1000 simulations) 189
Table 44 Number and proportion of non-differentially misclassified events required such that treatment effect is no longer significant at 5% level 191

Table 45 Characteristics of included trials..... 206

Table 46 Time taken for adjudication and dealing with disagreements 207

Table 47 Cost of dealing with disagreements between adjudicators. 208

Table 48 Cost of adjudication per corrected event by included trial .. 209

Table 49 Characteristics of Investigator reported serious adverse events and certainty of adjudicator's diagnosis 221

Table 50 Agreement on Diagnosis of serious adverse events between central adjudicators and site investigators..... 222

Table 51 Agreement on classification of serious adverse events between central adjudicators and site investigators 225

Table 52 Agreement on classification of cardiovascular serious adverse events between central adjudicators and site investigators..... 226

Table 53 Agreement on classification of central nervous system serious adverse events between central adjudicators and site investigators . 227

Table 54 Agreement on classification of respiratory serious adverse events between central adjudicators and site investigators..... 228

Table 55 Agreement on likely causality of serious adverse events between central adjudicators and site investigators 229

Table 56 Agreement on likely causality of SAEs between central adjudicators and site investigators – Dichotomous analysis 1..... 231

Table 57 Agreement on likely causality of SAEs between central adjudicators and site investigators – Dichotomous analysis 2..... 231

Table 58 Time taken for the adjudication of each serious adverse event 232

Table 59 Number of patients with serious adverse events during follow-up at day 90 classified by site investigators and central adjudicators: GTN vs no GTN and Continue vs Stop..... 234

Table 60 Impact of central adjudication on ENOS safety analysis: Ratio of Odds Ratios for serious adverse event classifications..... 235

Table 61 Agreement between site investigators and central adjudicators on the number of patients with serious adverse events during follow-up at day 90 236

Index of figures

Figure 1 Types of stroke 32

Figure 2 Partial flow diagram from Cochrane Review[60] 45

Figure 3 Fixed effect meta-analysis 71

Figure 4 Example funnel plot 79

Figure 5 Flow diagram 115

Figure 6 Analysis of 18 comparisons from 15 included trials, comparing the effect size for the primary outcome based on whether assessment was by central adjudicators or not 129

Figure 7 Meta-analysis of RTE in included studies, using a random effects model 130

Figure 8 Meta-analysis of RTE by intervention type, using a random effects model 131

Figure 9 Meta-analysis of RTE by adjudication risk of bias, using a random effects model 132

Figure 10 Meta-analysis of RTE by blinding status of site investigators, using a random effects model 133

Figure 11 Meta-regression with number of participants randomised fitted as a covariate 134

Figure 12 Meta-regression with number of trial sites fitted as a covariate 135

Figure 13 Meta-analysis of RTE by funding source, using a random effects model 136

Figure 14 Meta-analysis of RTE, using a fixed effect model 137

Figure 15 Meta-analysis of change in number of reported events in included studies, using a random effects model 138
Figure 16 Plan to elicit response..... 152
Figure 17 Kaplan-Meier curve for time to response for both intervention groups 159
Figure 18 Amount of non-differential misclassification required such that treatment effect (relative risk=0.67) is no longer significant at 5% level for various sample sizes and overall event rates..... 192
Figure 19 Amount of non-differential misclassification required such that treatment effect (relative risk=0.82) is no longer significant at 5% level for various sample sizes and overall event rates..... 193
Figure 20 Amount of non-differential misclassification required such that treatment effect (relative risk=0.90) is no longer significant at 5% level for various sample sizes and overall event rates..... 194
Figure 21 Agreement on serious adverse event diagnosis between central adjudicators and local investigators 223
Figure 22 Agreement on likely causality of serious adverse events between central adjudicators and site investigators 230

List of abbreviations

AC	Adjudication committee
AOST	Adjudicating Outcomes in Stroke Trials
CENTRAL	Cochrane Central Register of Controlled Trials
CI	Confidence interval
CNS	Central nervous system
CONSORT	CONsolidated Standards of Reporting Trials
CT	Computerised tomography
ECG	Electrocardiogram
ENOS	Efficacy of Nitric Oxide in Stroke
EMA	European Medicines Agency
FDA	Food and Drug Administration
GTN	Glyceryl trinitrate
GLASSY	GLOBAL LEADERS Adjudication Sub-StudY
ICC	Intraclass correlation coefficient
INQUIRE	Increasing Quality In Patient-oriented academic clinical Research
IQR	Inter-quartile range
ln	Natural logarithm
MeSH	Medical Subject Heading

MI	Myocardial infarction
mmHg	Millimetres of mercury
mmol	Millimoles
mRS	modified Rankin Scale
MRI	Magnetic resonance imaging
mRS	modified Rankin Scale
OR	Odds Ratio
PRISMA	Preferred Reporting Items for Systematic Review and Meta Analyses
RCT	Randomised controlled trial
REWARD	Reduce research waste and reward diligence
ROR	Ratio of Odds Ratios
RTE	Ratio of treatment effects
SAE	Serious adverse event
SAR	Serious adverse reaction
SD	Standard deviation
SE	Standard error
SWAT	Study within a trial
TIA	Transient Ischaemic Attack

Abstract

Central adjudication in randomised trials refers to the evaluation of outcome data by independent experts who are typically part of an event or outcome adjudication committee. Alternatively, local site investigators can assess outcomes, but these evaluations are often discarded in favour of the centrally adjudicated outcome. Central adjudication is thought to improve the precision of treatment effect estimates by reducing random error (non-differential misclassification), and in open-label studies adjudication has the potential to limit systematic error (differential misclassification) as adjudicators can always be blinded to treatment allocation.

However, secondary analysis of trial data suggests that adjudication makes no meaningful difference to the endpoints of stroke and functional outcome. In addition, adjudication is a time-consuming and often expensive process. Given that central adjudication is common in trials investigating stroke, research is needed to establish what the benefits and costs are of adjudicating outcomes in stroke trials.

In a systematic review, 15 randomised stroke trials (69,560 participants) were identified that had their primary outcome assessed by both central adjudicators and site investigators. The primary outcomes included were stroke (8 out of 15 trials, i.e. 8/15, 53%), a composite including stroke (6/15, 40%) and functional outcome after stroke (1/15, 7%). Overall, there was no evidence of any difference in treatment effect

estimates based on data from central adjudicators and site investigators (pooled Ratio of Treatment Effects=1.02, 95% C.I:[0.95, 1.09]).

This thesis also investigated whether two different approaches for contacting authors had any impact on the probability of receiving a response when trying to acquire data for a systematic review in a methodological trial. This nested randomised trial found that a short email with a protocol attached elicited a similar response compared to a longer email without a protocol.

To explore circumstances where central adjudication would change the treatment effect estimate, five of the trials identified in the review were included in a further simulation study. Differential misclassification was introduced into each study via simulation until the treatment effect estimate was altered. For trials with a binary outcome, between 2.1% and 6% of participants needed to be differentially misclassified before this situation ensued.

In addition, hypothetical trials were simulated with a binary outcome and varying sample size (1000-10000), overall event rate (10-50%), and treatment effect (0.67-0.90). Non-differential misclassification was introduced until the treatment effect was non-significant at 5% level. For these hypothetical trials, extensive non-differential misclassification was needed before the treatment effect became non-significant; trials with an overall event rate close to 50% and a larger sample size needed the highest proportion of random error before this occurred.

Nine of the trials included in the review provided data on the cost of adjudication. These costs included adjudicators' time, direct payments to adjudicators and co-ordinating centre costs. The number of events corrected after adjudication was the measure of benefit used. The mean cost per event corrected by adjudication was £2295.10.

To investigate whether these findings were similar for adjudication of safety data, a case study was carried out using data from the Efficacy of Nitric Oxide in Stroke Trial. Serious adverse events were reported by site investigators who were not blinded to treatment allocation. Central blinded adjudicators reviewed the investigators' report and used evidence available to confirm or re-categorise the classification of event. Repeating the main trial safety analysis with investigator reported events showed that adjudication had no effect on the main trial safety conclusions.

To conclude, these studies have shown that central adjudication of the primary outcome in stroke trials does not alter treatment effect estimates. However, for studies without adequate blinding, a small amount of systematic error has the potential to alter the primary analysis and in this circumstance, adjudication is important. Given that the cost of central adjudication is not trivial, the potential advantages of adjudication may not outweigh cost and time disadvantages in stroke trials with blinded outcome assessment.

Acknowledgements

I would like to thank Professor Alan Montgomery and Professor Philip Bath who helped conceive this project and have provided me with consistent support throughout. Their ideas, knowledge and input have been invaluable in the completion of this thesis. I would like to express special thanks to Professor Alan Montgomery for his guidance and motivation, which have been of paramount importance.

I would also like to thank the Adjudicating Outcomes in Stroke Trials collaboration. Without their help, this research would not have been possible. I would like to acknowledge my colleagues at Nottingham Clinical Trials Unit, who have kept me entertained, especially during the write-up of this thesis. Special thanks go to Dr Christopher Partlett, Dr Reuben Ogollah, Trish Hepburn, Wei Tan, Lucy Bradshaw, and Mara Ozolins for their encouragement, friendship and advice. I would also like to thank Dr Matthew Grainge for his guidance throughout.

The work in this thesis was funded by a National Institute for Health Research Doctoral Research Fellowship (DRF-2016-09-057). I would like to thank the National Institute for Health Research for their support.

Finally, I would like to express my gratitude to my family and friends. Special mentions go to my Dad for his understanding and patience during these last three years, to my Mum for her exceptional proof reading ability and sound advice, and to Francesca for her undying support, constant encouragement, and ability to keep me motivated, without which I would have struggled to complete this thesis.

The Adjudicating Outcomes in Stroke Trials (AOST) Collaboration

The majority of the research presented in this thesis was undertaken as part of the AOST collaboration. The members of the AOST collaboration are listed in Table 1. The majority of the collaborators provided data from randomised trials that they either led or were part of the research team. The five members of the collaboration who did not provide data were part of Nottingham Clinical Trials Unit during their involvement in the project.

The areas in which all members of the collaboration in this research programme contributed are given in Table 2. This table indicates where each member of the collaboration led a specific element of the project, made a major contribution to or was involved in a supporting role for each component of this research. I would like to thank all members of the collaboration for making this research project possible, and for their contribution in assisting with each element of the research.

Throughout the remainder of this thesis, I will use the first person singular voice in active sentences for consistency. I recognise that the research described in this thesis is collaborative work, and that there were many contributors for the elements of this project. Therefore, I would like to acknowledge the input from both my supervisors and all members of the collaboration in this work.

Table 1 Members of the AOST Collaboration

Name	Initials	Trial involved with	Affiliation
Ale Algra	AA	ESPRIT	Utrecht University
Pierre Amarenco	PA	SOCRATES	Paris Diderot University
Hisatomi Arima	HA	PROGRESS	Fukuoka University
Henry JM Barnett	HJMB	NASCET	-
Philip M Bath	PMB	TARDIS	University of Nottingham
Eivind Berge	EB	HAEST	Oslo University Hospital
Leo H Bonati	LHB	ICSS	University College London
Martin M Brown	MMB	ICSS	University College London
John Chalmers	JC	PROGRESS	University of New South Wales
Christopher Chen	CC	VITATOPS	National University of Singapore
Antoni Dávalos	AD	REVASCAT	Hospital Germans Trias i Pujol
Robert Dineen	RD	TARDIS	University of Nottingham
Lelia Duley	LD	-	University of Nottingham
J Donald Easton	JDE	POINT	University of California, San Francisco
John W Eikelboom	JWE	VITATOPS	McMaster University
Misha Eliasziw	ME	NASCET	Tufts University
Jordan J Elm	JJE	POINT	Medical University of South Carolina
Scott Evans	SE	SOCRATES	Harvard University
Mary Farrant	MF	POINT	University of California, San Francisco
Allan J Fox	AJF	NASCET	University of Toronto
Robert P Giugliano	RPG	ENGAGE AF-TIMI 48	Harvard University
Peter J Godolphin	PJG	-	University of Nottingham
John Gregson	JG	ICSS	London School of Hygiene and Tropical Medicine
Jacoba P Greving	JPG	ESPRIT	Utrecht University
Graeme J Hankey	GJH	VITATOPS	The University of Western Australia
Naohisa Hosomi	NH	J-STARS	Hiroshima University
S Claiborne Johnston	SCJ	SOCRATES	The University of Texas at Austin
Stephen C Knipp	SCK	CABACS	Universitätsklinikum Essen
Masayasu Matsumoto	MM	J-STARS	Hiroshima University
Alan A Montgomery	AAM	-	University Of Nottingham
Yoji Nagai	YN	J-STARS	Kobe University Hospital
Bruce Neal	BN	PROGRESS	University of Sydney
Christopher Partlett	CP	-	University Of Nottingham
Emily Patsko	EP	-	University of Leicester
Annamarei Ranta	AR	FASTEST	University of Otago
Per Morten Sandset	PMS	HAEST	Oslo University Hospital
Joaquín Serena	JS	REVASCAT	Hospital Josep Trueta
Nicola Sprigg	NS	TARDIS	University Of Nottingham
Christian Weimar	CW	CABACS	Universitätsklinikum Essen
Lisa Woodhouse	LW	TARDIS	University Of Nottingham

Table 2 Summary of contributions of AOST Collaborators

Task	Lead contribution	Major contribution	Supporting contribution
Obtaining funding	PJG, AAM, PMB	-	-
Conceptualisation	PJG, AAM, PMB	-	-
Designing study protocols	PJG	AAM, PMB,	LD, NS
Designing analysis plans	PJG	AAM, PMB	CP
Designing data extraction form	PJG	AAM	PMB, LD
Designing cost questionnaire	PJG, AAM	PMB	-
Scoping review and designing search strategy	PJG	-	AAM, PMB
Screening titles and abstracts	PJG	EP	AAM
Screening full texts	PJG	EP	AAM
Extracting data and assessing risk of bias	PJG, EP	-	AAM
Collecting data for response rates trial	PJG	-	-
Providing individual patient data and/or aggregate trial results	-	PMB, AA, EB, MMB, JC, ME, JG, JPG, GJH, NH, SCJ, JS, CW, RPG	SCK, PMS, LHB, YN, MM, HJMB, AJF, MF, JDE, JJE, BN, HA, AD, PA, SE, NS, RD, JWE, CC
Providing cost data	-	PMB, LW, GJH, CW, EB, AA, AR, NH, HA, ME	-
Writing simulation programs	PJG	-	-
Data entry and data cleaning	PJG	-	-
Data analysis and interpretation	PJG	AAM, PMB	-
Writing AOST publications	Chapter 3: PJG Chapter 4: PJG Chapter 5: PJG Chapter 6: PJG Chapter 7: PJG	Chapter 3: AAM, PMB, LD, EB Chapter 4: AAM, PMB Chapter 5: AAM, PMB, CP Chapter 6: AAM, PMB Chapter 7: AAM, PMB, NS	Chapter 3: AA, MMB, JC, ME, JG, JPG, GJH, NH, SCJ, JS, CW, PMS, EP Chapter 4: - Chapter 5: EB, PMS, ME, JS, MMB, Chapter 6: LW, GJH, CW, EB, AA, AR, NH, HA, ME Chapter 7: LW

Publications

This section contains a list of outputs that arose from work in this thesis. This list comprises journal articles and conference presentations.

Peer-reviewed journal articles

1. PJ Godolphin, AA Montgomery, LJ Woodhouse, D Bereczki, E Berge, R Collins, E Diez-Tejedor, J Gommans, KR Lees, S Ozturk, S Phillips, S Pocock, K Prasad, S Szatmari, Y Wang, PM Bath, N Sprigg. Central adjudication of serious adverse events did not affect trial's safety results: Data from the Efficacy of Nitric Oxide in Stroke (ENOS) trial. PLoS ONE 2018, 13(11):e0208142. https://doi.org/10.1371/journal.pone.0208142

2. PJ Godolphin, PM Bath, AA Montgomery. Short email with attachment versus long email without attachment when contacting authors to request unpublished data for a systematic review: a nested randomised trial. BMJ Open 2019, 9:e025273. http://dx.doi.org/10.1136/bmjopen-2018-025273

3. PJ Godolphin, PM Bath, A Algra, E Berge, MM Brown, J Chalmers, L Duley, M Eliasziw, J Gregson, JP Greving, GJ Hankey, N Hosomi, SC Johnston, E Patsko, A Ranta, PM Sandset, J Serena, C Weimar, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in stroke trials: A systematic review and meta-analysis. Stroke 2019, 50:2187-2196. https://doi.org/10.1161/STROKEAHA.119.025019

4. PJ Godolphin, PM Bath, C Partlett, E Berge, MM Brown, M Eliasziw, PM Sandset, J Serena, AA Montgomery. Central adjudication of outcomes is most valuable in unblinded trials: a simulation study. [Under Review].

5. PJ Godolphin, PM Bath, A Algra, E Berge, J Chalmers, M Eliasziw, GJ Hankey, N Hosomi, A Ranta, C Weimar, L Woodhouse, AA Montgomery. Cost-benefit of outcome adjudication in nine randomised stroke trials. [Under Review].

Conference oral presentations

1. PJ Godolphin, PM Bath, AA Montgomery. Different methods of author contact in order to receive unpublished data for a Systematic Review. 2018. Oral presentation at Link ’18 Conference.

2. PJ Godolphin, PM Bath, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in randomised stroke trials: A systematic review and meta-analysis. 2019. Oral presentation at 5th International Clinical Trials Methodology Conference.

Conference poster presentations

1. PJ Godolphin, L Woodhouse, D Bereczki, AA Montgomery, PM Bath, N Sprigg. Assessing the impact of independent adjudication of serious adverse events on the safety results of the efficacy of nitric oxide in

stroke (ENOS) trial. Trials 2017, 18(Suppl 1):P166. Poster presentation at 4th International Clinical Trials Methodology Conference.

2. PJ Godolphin, PM Bath, AA Montgomery. A comparison of central adjudication and site assessment of outcomes in randomised stroke trials: the Adjudicating Outcomes in Stroke Trials' collaboration (AOST). 2018. Poster presentation at 4th European Stroke Organisation Conference.

3. PJ Godolphin, PM Bath, AA Montgomery. Randomised trial of different methods of author contact in order to receive data for a systematic review of randomised stroke trials. 2018. Poster presentation at 4th European Stroke Organisation Conference.

4. PJ Godolphin, PM Bath, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in randomised stroke trials. 2019. Poster presentation at 40th International Society for Clinical Biostatisticians Conference.

5. PJ Godolphin, PM Bath, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators in randomised stroke trials: simulation of differential and non-differential misclassification. 2019. Poster presentation at 5th International Clinical Trials Methodology Conference.

Chapter 1 – Introduction

Publications contributing to this chapter:

1. PJ Godolphin, AA Montgomery, LJ Woodhouse, D Bereczki, E Berge, R Collins, E Diez-Tejedor, J Gommans, KR Lees, S Ozturk, S Phillips, S Pocock, K Prasad, S Szatmari, Y Wang, PM Bath, N Sprigg. Central adjudication of serious adverse events did not affect trial's safety results: Data from the Efficacy of Nitric Oxide in Stroke (ENOS) trial. PLoS ONE 2018, 13(11):e0208142. https://doi.org/10.1371/journal.pone.0208142

2. PJ Godolphin, PM Bath, AA Montgomery. Short email with attachment versus long email without attachment when contacting authors to request unpublished data for a systematic review: a nested randomised trial. BMJ Open 2019, 9:e025273. http://dx.doi.org/10.1136/bmjopen-2018-025273

3. PJ Godolphin, PM Bath, A Algra, E Berge, MM Brown, J Chalmers, L Duley, M Eliasziw, J Gregson, JP Greving, GJ Hankey, N Hosomi, SC Johnston, E Patsko, A Ranta, PM Sandset, J Serena, C Weimar, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in stroke trials: A systematic review and meta-analysis. Stroke 2019, 50:2187-2196. https://doi.org/10.1161/STROKEAHA.119.025019

1.1 Background

This chapter describes the subject area of the thesis, including a brief overview of stroke (incidence, diagnosis, common outcomes and clinical trials), how outcomes are assessed in clinical trials at local sites and centrally through adjudication, central adjudication in cardiovascular trials and more specifically randomised trials investigating stroke and methods to contact authors of clinical trials to acquire their data for a systematic review. The final two sections of this chapter highlight gaps in the current literature and describe the research question this thesis aims to answer, and the general aims of this thesis.

1.2 Stroke

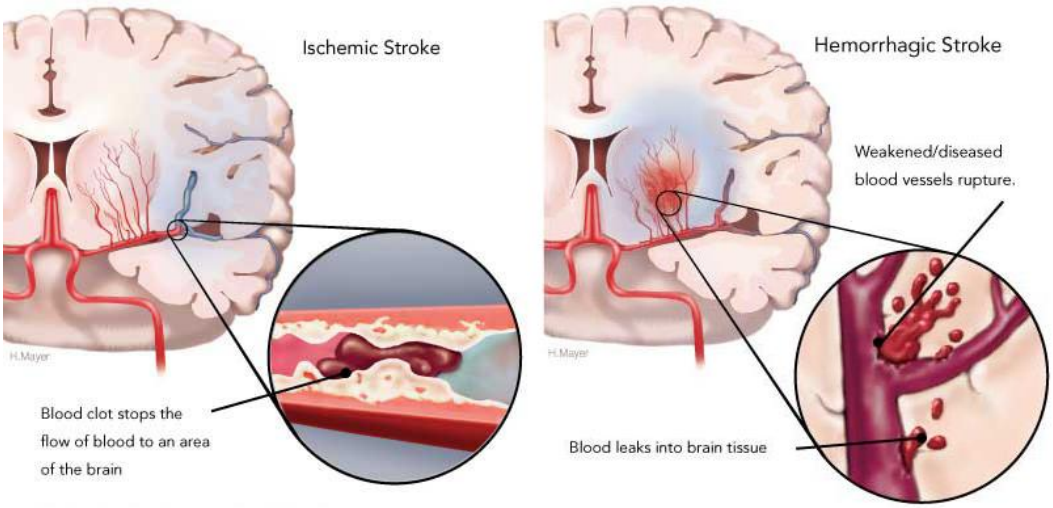
There are over 150,000 cases of stroke a year in the UK[1]. Indeed, stroke is the fourth highest cause of death in the UK[2] and the second largest cause of death globally[3]. Stroke represents a growing health problem worldwide, with the global burden predicted to double by 2030[4].

To tackle this growing problem, there has been an increase in the amount of research funding for stroke, with the UK government and charities more than doubling the amount of funding allocated to stroke research between 2008 and 2012 to £56M[5]. Increased funding allows research into effective interventions and therapies to help prevent and manage stroke, and measures to improve rehabilitation for stroke survivors. Due to clinical trials being considered one of the best means

of obtaining evidence to determine efficacy of a new treatment[6], there has been a steady increase in the number of clinical trials carried out across all disease areas[7], with a large number undertaken for stroke specific trials[8].

There are two main types of stroke, ischaemic and haemorrhagic (Figure 1). Ischaemic strokes are more prevalent and make up approximately 87% of all stroke, leaving the remaining 13% as various forms of haemorrhage[9]. An Ischaemic stroke occurs when there is an occlusion in one of the arteries that supply blood to the brain, leading to a loss of neurological function[10]. A haemorrhagic stroke is caused when one of the intracerebral arteries ruptures, causing bleeding directly into the brain tissue[11]. Patients with a haemorrhagic stroke often complain of a headache, nausea and vomiting, but these symptoms do not distinguish between ischaemic and haemorrhagic stroke with any certainty[11].

Figure 1 Types of stroke



Original images from https://www.heartandstroke.ca/stroke/what-is-stroke/types-of-stroke

Stroke is primarily diagnosed using patient history and physical examination, and timely diagnosis is important due to acute interventions for stroke having a narrow time period to be effective[9]. Neuroimaging is used to confirm diagnosis of stroke and also to differentiate between ischaemic and haemorrhagic stroke[12]. A computerised tomography (CT) scan is used to rule out stroke mimics, that is conditions that present like a stroke but are in fact non-stroke disorders such as seizures or migraines, and to identify haemorrhages, but CT scans do not always detect ischaemic stroke[9]. Many varied interventions are available to treat acute stroke (such as venous thrombolysis with alteplase, and mechanical thrombectomy[13, 14]), to reduce the chance of first or subsequent stroke, and to improve patient's mobility or assist in the rehabilitation process after patients have suffered a stroke.

Clinical trials are often designed to compare the efficacy of a number of experimental treatments under test, and any potential benefits tend to be measured alongside the side-effects or hazards. To determine benefits and hazards, a between-group measure of effect is estimated for some outcome that the trial collects. In stroke trials, key outcomes commonly collected are occurrence of stroke, type of stroke, disability level a specified time period after stroke onset, and mobility measures. Stroke trials can be categorised into three main types; acute, prevention and rehabilitation. Acute stroke trials are those which aim to treat participants shortly after stroke onset to improve their eventual outcome. Stroke prevention trials can be categorised into primary and

secondary prevention trials, where primary prevention studies aim to prevent participants who are at risk of a stroke from having their first stroke, and secondary prevention trials involve participants who have had a stroke previously, as nearly a third of stroke survivors suffer recurrence[15]. Rehabilitation trials are often more diverse and cover a range of outcomes and interventions in an effort to improve a variety of areas such as mobility, vision, swallowing and mood[16].

Due to the increase in research spending and number of clinical trials undertaken, there is a current drive to ensure that trials are run efficiently so that research funding is not wasted. A study by Yordanov et al.[17] found that there was avoidable research waste caused through inadequate planning in the majority of 2711 randomised trials reviewed. Thus, research waste is an area that requires due attention. The Lancet's Reduce research waste and reward diligence (REWARD) campaign was founded in 2014 and aims to maximise research potential[18]. The campaign focuses on five main areas: priority setting[19]; research design, conduct and analysis[20]; regulation and management[21]; accessibility[22] and reporting[23], to ensure that "we commit to playing our part in increasing value and reducing waste in research"[18].

In addition, a group of researchers have formed Increasing Quality In Patient-oriented academic clinical Research (INQUIRE), a framework of clinical research quality to assist in "avoiding waste and increasing value in clinical research"[24]. There are other initiatives, such as Trial Forge, which provides an interactive website with methodological

research on improving trial efficiency that is easily accessible[25, 26]. The goal of this platform is to increase the evidence base for trial design to gain small improvements in each of the individual trial processes, in order to make a considerable improvement overall in trial efficiency. Treweek et al.[25] state: “What has been lacking is a coordinated attempt to tackle inefficiency in clinical trials”.

1.3 Outcome Assessment in Clinical Trials

Throughout this thesis, the term site investigator will be used for the person(s) involved in a trial who originally assess the outcome(s) at the participating trial site. Examples of site investigators are research nurses, clinicians or outcome assessors.

In a clinical trial, assessors at each participating research site or centre typically make a decision about a specific outcome, for example occurrence of stroke. The assessor will often have interacted with the participant, have read clinical notes, seen other medical data such as CT or magnetic resonance imaging (MRI) scans, and have witnessed symptoms to determine their decision. As such, in some trials (e.g. open trials), the site investigator may be unblinded to treatment assignment[27] and may exhibit bias, even if it is subconscious. It is difficult to test for blinding, with uncertainty around the best method to determine whether blinding has been adequate[28], and studies do not commonly report whether their trials are at risk of unblinding[29]. Therefore, having an independent blinded assessment of key trial outcomes could assist in reducing potential biases[27].

There is the potential for unblinded site-assessors to exaggerate treatment effect estimates[30], with subjective outcomes most prone to bias. This exaggeration is caused through differential misclassification, where the site assessors systematically allocate participants erroneous outcomes influenced by their knowledge of treatment assignment[31]. Subjective outcomes are often defined as those that are “dependent on assessor judgement”[32]. Examples of such outcomes are quality of life, degree of pain and even measures such as type of stroke or cause of death that require some decision making from the assessor. Studies investigating whether unblinded site-assessors exhibited bias on subjective outcomes found that they exaggerated effect estimates by 36% for binary outcomes[30], 27% for time-to-event outcomes[33] and 68% for measurement scale outcomes[34]. Unblinded site investigators are most likely to exhibit detection bias, whilst performance bias is also a possibility. The following subsections explain what is meant by these different kinds of bias and how they can occur in clinical trials.

1.3.1 Detection bias

Detection bias denotes systematic differences in how outcomes are assessed between treatment arms. This is often caused by inadequate blinding of outcome assessors, assessors employing different measurement techniques or tools, or differences in outcome definitions or time points where outcomes are collected[35]. These differences have the potential to impact on the validity of treatment effect estimates. Blinding of outcome assessors can reduce the chance that assessors know which intervention participants received. This helps to ensure that

it is the intervention that affects the trial outcome rather than knowledge of treatment allocation. There is greatest risk of detection bias for subjective outcomes, where assessors' opinions have greater potential to be influenced by knowledge of the trial intervention[36]. Not all studies can be designed in a way so that outcome assessors are blinded, although even in an open trial many trial designs aim to ensure that the key endpoints are collected and/or evaluated by blinded assessors[37]. In some randomised trials, self-reported outcomes are collected, and in this case if participants know the treatment assignment, it is impossible to avoid the possibility of detection bias.

1.3.2 Performance bias

Performance bias refers to the differences in the care that is given to participants in each treatment arm, other than the trial intervention[36]. Sufficient blinding of participants and providers is vital to protect against performance bias. Ensuring that participants and providers are unaware of treatment allocation guards against the possibility that the level of attention, amount or type of additional therapies received, or supplementary testing is systematically different between treatment arms. Blinding is not always possible, since in some trials it is obvious which intervention participants have received, for example surgery trials without sham treatment. Furthermore, even in trials which appear to blind participants and/or personnel it can become apparent which treatment has been administered based on side effects or tastes/smell of the trial intervention. Therefore, in most randomised studies there is

a risk that performance bias could be evident and have some impact on the estimated between group measure of effect[38].

1.3.3 Non-differential misclassification

For randomised trials with sufficient blinding around outcome assessment, detection and performance bias should not occur. In this case, site assessors will not differentially misclassify the trial outcome(s)[39]. However, it is still possible that non-differential misclassification might occur. This misclassification is caused by random error, and occurs equally in each trial arm. For large, multicentre trials, non-differential misclassification is often prevalent, with varied practice across trial sites and varied levels of experience between the site investigators at the different trial sites. The amount of misclassification may be higher for international studies, since language and cultural differences could cause outcomes to be assessed in a different fashion[40].

Examples of differential and non-differential misclassification of a binary outcome are given in Tables 3-6. Table 3 shows observed data, with no indication of any difference in outcome between treatment and control. Table 4 shows an example situation after non-differential misclassification, with random error resulting in reduced proportions with events affecting both groups in the same way. Table 5 gives an example situation of differential misclassification, here there is systematic error, with only participants in the treatment group misclassified so that the benefit of treatment is enhanced. Table 6 shows an alternative example of differential misclassification, with

events reduced in the treatment arm and events increased in the control arm, causing even further exaggerated effect compared with Table 5.

Table 3 Example binary outcome

	Treated (n=100)	Control (n=100)
No event	75	75
Event	25	25

Table 4 Example binary outcome after non-differential misclassification

	Treated (n=100)	Control (n=100)
No event	85	85
Event	15	15

Table 5 Example binary outcome after differential misclassification

	Treated (n=100)	Control (n=100)
No event	85	75
Event	15	25

Table 6 Example binary outcome after differential misclassification

	Treated (n=100)	Control (n=100)
No event	85	65
Event	15	35

1.4 Central Adjudication in Clinical Trials

Central adjudication in clinical trials refers to the re-evaluation of clinical trial data by independent experts in order to provide information about key outcomes or safety issues. The independent experts are often referred to as the adjudication committee, endpoint committee or clinical events committee. The use of adjudication in clinical trials has become commonplace in many clinical areas, including oncology[41] and cardiovascular disease, with over 80% of cardiovascular trials reported to include adjudication[42]. Independent adjudication can be conducted

by a single expert or a panel of experts, and the adjudicators typically assess primary outcomes determined by the site assessors to decide whether or not they agree, or to categorise them in a specific way. Safety outcomes are often adjudicated to ensure they are valid and can be classified by severity.

Central adjudicators can always be blinded to treatment assignment and therefore implementing adjudication in a clinical trial protects against detection bias and helps control for differential misclassification[43]. Reducing bias in clinical trials is vital to ensure that trial results are valid and that decisions about a particular intervention are correct. This is especially important in trials with few events, where a small degree of differential misclassification can have a large impact on study conclusions[44, 45]. There are guidelines in place from both the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) which suggest that adjudication is particularly valuable "when the intervention is not delivered in a blinded fashion"[46] or when "endpoints are complex to assess and/or include subjective components"[47].

In addition, adjudication provides a standardised way of assessing outcomes. As described previously, in a large international multicentre trial there could be a large number of site assessors. These assessors can differ in experience, ability and thoroughness. In trials where outcome assessors are blinded to treatment, some assessors may become unblinded and therefore at risk of bias. Ensuring that a small pool of central experts analyse all trial data not only helps to guard

against bias, it also potentially enables a more accurate evaluation of the trial outcome by experienced and blinded adjudicators. This should reduce random error and make treatment effect estimates more precise, theoretically meaning that sample sizes can be reduced (or at least not inflated) for trials that employ appropriate adjudication. A think tank organised at the FDA's White Oak headquarters in 2016 summarised that central adjudication is undertaken to limit bias and can help maintain standardised definitions[48]. The consortium mentioned that "adjudication… may be less critical in small feasibility studies or if endpoints rely on objective metrics rather than on the interpretation of complex data that require high levels of clinical judgement"[48]. Moreover, an overview of the use of adjudication in EU and US drug approvals found that approximately 20-30% of the primary endpoints were classified as subjective, and therefore most likely to receive benefit from central adjudication[49].

Furthermore, apart from the measureable benefits that outcome adjudication has for clinical trials, there are 'softer' advantages such as reassurance that adjudication can have for clinical trial staff. Two clinical trials commenting on the adjudication process in their studies, specifically mentioned that the assurance an adjudication committee brings is "of substantial importance"[50, 51]. Additionally, adjudication introduces a level of quality control to detect and correct outcomes from poorly trained or poorly performing investigators[52]. This can be particularly useful in a lengthy clinical trial, and indeed many stroke trials are lengthy, lasting many years and recruiting thousands of

participants[53]. Countries or regions that require further training can be identified or additional measures to help certain trial sites improve their outcomes assessment can be implemented. Moreover, adjudication itself could in fact result in improved site outcome assessment. If site assessors know that adjudication is being undertaken, then site performance may increase due to a policing effect from the adjudication. In this situation, the site assessors are aware that their outcome assessment will be checked and so may perform more carefully. On the other hand, site assessors may become less strict or careful, as they know that their outcome will not be the final outcome used in the trial. Trying to identify the impact of adjudication on these 'softer' outcomes is challenging and may be better investigated using qualitative methods, but it is important to keep them in mind when deciding on the usefulness of the adjudication process overall.

Furthermore, whilst it is the assumption that adjudication is the 'gold standard' and when the adjudicated outcome differs from the outcome assessed by the site investigator that the adjudicated outcome is the correct one, it is often the case that adjudicators only have access to limited patient data, and never physically interact with the patients, unlike many site assessors[43]. The view held by some is that, although experts in their field, the opinions of the adjudicators may hold no more validity than those of the hospital staff who have overseen the patient and made a decision based on a number of factors, rather than just numerical data or scans[44]. The level of information available to adjudicators may have a large impact on how accurate the adjudicator's

outcome assessment really is. Additionally, if adjudicators are simply relying on site assessment of outcomes, then biased reports and clinical notes may mean that adjudication does not reduce the bias seen in a trial. Whilst this would not be the case for more objective evidence, such as cranial scans, it is a possibility for a large number of commonly adjudicated outcomes.

Moreover, adjudication is a costly procedure and can result in a significant delay in a clinical trial. The data are sent to an adjudicator who reviews it and sends back their decision; this is then entered and combined with the site information[43, 44, 54]. If there is disagreement between the adjudicator and site, then a second adjudicator is often asked to evaluate the data to determine the final outcome. A variety of methods are used to deal with disagreements[55]. In some trials there are multiple stages in the process to deal with disagreements. In others, group adjudication is implemented, with all the adjudicators convening to discuss the conflicting outcomes with the aim of reaching a consensus decision[54, 56]. Group adjudication is potentially an even greater expense, when taking into account adjudicator time, travel costs, room hire and other factors such as duplication of effort in having each outcome assessed by every adjudicator[43].

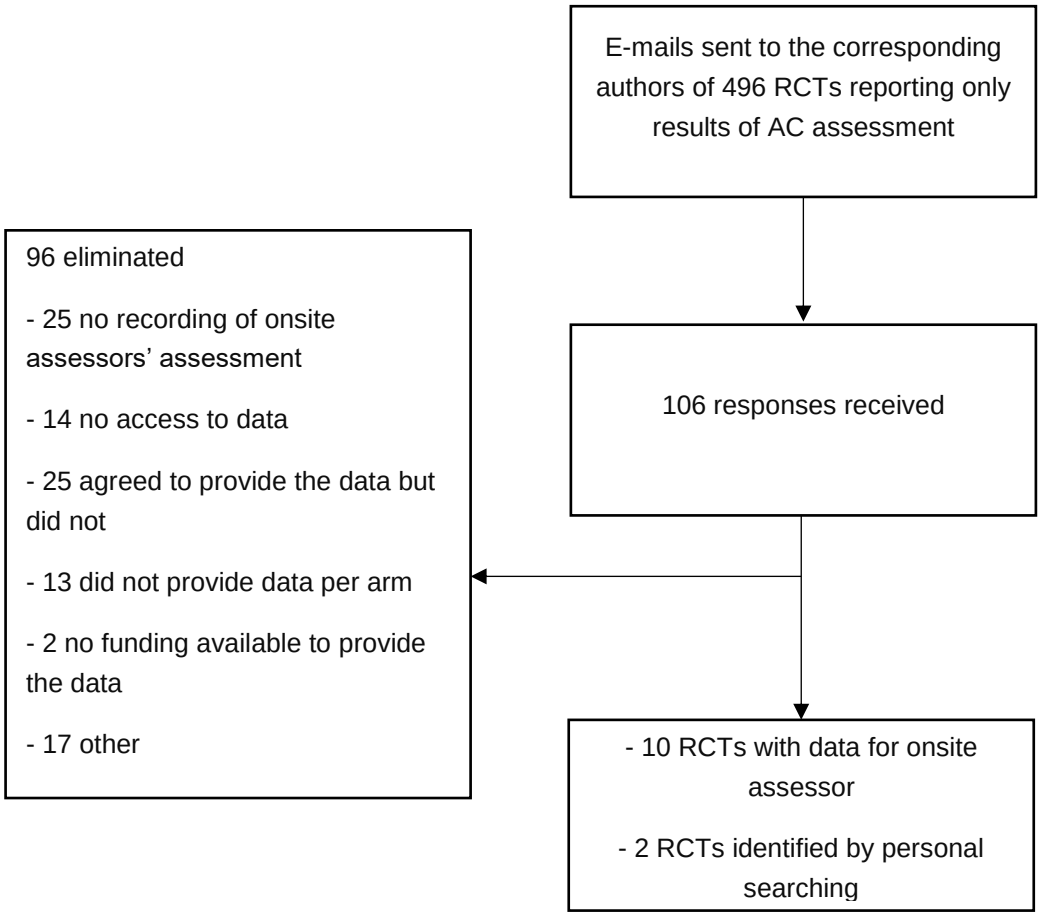
In medical research that involves many thousands of patients, the adjudication process is known to be complex, time-consuming and expensive[57]. Adjudication of bleeding outcomes in an international thromboprophylaxis trial was estimated to cost £80,000 for 1,737 suspected events[58, 59], with these costs only being those associated

with the co-ordinating centre (e.g. preparing the events for adjudication, designing and implementing an online platform for adjudication, filing etc.). There are additional costs to consider, including direct costs paid to adjudicators and the cost of the time that it takes the adjudicators to carry out their adjudications as well as dealing with disagreements. In a large cardiovascular trial, adjudication was carried out on all suspected cases of nonfatal myocardial infarction, nonfatal stroke and cardiovascular death at an estimated cost of over £500,000 for 7,911 diagnoses[51]. To justify this level of spending, particularly in publicly funded trials, adjudication must provide sufficient benefit. This benefit can be measured in several ways, and, as described earlier there may be additional benefits or problems that adjudication brings that are unmeasurable. One way to gain an understanding of the usefulness of adjudication is to investigate if the results from a randomised trial would have differed had no adjudication taken place.

A Cochrane systematic review[60] compared central adjudication of subjective binary outcomes with site outcome assessment and investigated its effect on treatment estimates. The review found 47 studies (275,078 patients) eligible for inclusion in a meta-analysis and the conclusion was that, on average, there was no difference in treatment effect estimates between adjudicators and site assessors (Ratio of Odds Ratios: 1.00, 95% C.I: 0.97 to 1.04). It could be hypothesised that subjective outcomes are most prone to be ‘corrected’ by adjudication, and so this result, showing that adjudication of these events did not change the trial results on average, could give weight to

the idea that adjudication of objective outcomes may have even less impact. No statistically significant interaction by blinding status was found, although the Ratio of Odds Ratios point estimate of 1.08 suggested that adjudication in trials with unblinded site assessment was more likely to affect trial conclusions, this was not conclusive (95% C.I: 0.94 to 1.23). Therefore, this evidence contributes to the argument that adjudication may not provide sufficient benefit in clinical trials to warrant the cost.

Figure 2 Partial flow diagram from Cochrane Review[60]



RCT refers to randomised controlled trial, AC refers to Adjudication Committee

During this Cochrane review, the review team found 496 studies that reported only the results of the adjudicated outcome, and not the outcome of the site-assessor. After contacting the corresponding authors of the 496 randomised trials, 12 provided sufficient data to be included in the review (Figure 2[60]). Thus, it appears that the site-assessed outcomes are not routinely published, and contacting the authors of the randomised trials to receive these data is necessary in order to evaluate central adjudication.

1.5 Methods of Author Contact to Receive Unpublished Trial Data for a Systematic Review

Systematic reviews are often considered to provide the highest possible level of evidence[61]. However, meta-analyses, the statistical components of a systematic review, often rely on the acquisition of unpublished summary results or further data, as described previously for the Cochrane review. It can be a challenge to obtain these data, especially if the meta-analysis requires individual patient data[62, 63]. This results in there being insufficient studies to pool together[64]. A large amount of missing data in a systematic review can threaten the validity of the study[65].

Data sharing is not yet standard practice in the field of biomedical research, although it may soon become commonplace[66], and many researchers currently struggle to acquire the underlying datasets used in journal articles[67]. There are many reasons why there may be difficulties in acquiring data; authors may not wish to share their data[68],

they may be too busy to respond to or deal with requests[64], they may have moved institution and be uncontactable, the data could have been lost or destroyed, or authors may not have access to the data[69]. Author inclination to share data could be related to the strength of evidence and perceived quality of statistical results from their study[70]. Research studies that report borderline evidence of an intervention effect may be less likely to take part in data sharing, which would result in publication bias for a systematic review[71].

Furthermore, for a systematic review that requires unpublished data for its meta-analysis, like the Cochrane review described previously, data availability bias may occur[71]. This situation arises if only a subset of the studies asked to provide data actually agree to the request. As seen in the Cochrane review, less than 3% of trials approached actually provided data to allow their inclusion in the review. There is the possibility that trials that do provide data show different results compared to those that do not provide data. Thus, it is imperative to attempt to receive the maximum amount of data possible, and therefore the method of author contact is important to facilitate this.

Various methods are currently employed to encourage authors to share data for a systematic review. Some example of these are financial incentives and including data sharers as authors in study publications. A randomised trial investigating the effectiveness of financial incentives found that data sharing may depend on specific trial characteristics (e.g. sample size and treatment effect) but that financial incentives do not help to facilitate data sharing[72, 73]. Academic incentives for making

data available could be introduced, such as invitation into group collaboration, or having data sharing recognised to an equivalent level as scientific presentations or journal publications[74]. One method to assist in data sharing is providing researchers' with a data-sharing agreement. A randomised trial of 247 researchers found that those who received a data-sharing agreement were more likely to share data, compared with those who did not. The study found that receiving a data-sharing agreement improves participants' willingness to share their data set by 24%[75].

Whilst there are methods in place to improve the chance of receiving data, there appears to be no consensus about the best way to approach authors to contact them with a request for data. The corresponding author should be the most appropriate person to contact when attempting to acquire unpublished clinical trial data, but contacting these authors may be challenging as they could be uncontactable or have little time to deal with requests[76]. Therefore, contact with the corresponding author needs to be clear and provide them sufficient information to be able to make a decision on whether or not they can and/or wish to assist the review team by providing data.

The initial interaction with any author may have a large impact on whether they collaborate, or even on whether they reply at all. The first step when attempting to acquire data is to open up a channel of communication with the data custodian, assumed in this thesis to be the author of the journal article, with email representing a cost-effective and common way of contacting fellow researchers[77]. Thus, an important

and interesting outcome to assess is whether an initial response is received, before considering if subsequent interaction with the author will affect their willingness to share unpublished data with the systematic reviewer.

1.6 Methods of Adjudication in Clinical Trials

There are many different ways that adjudication is currently implemented in randomised trials. One of the first considerations to be taken into account is the number of adjudicators to use. A review of adjudication-related processes by Stuck et al.[55] found that nearly 50% of trials used three adjudicators, whilst approximately 11% used six or more. However, it was also common for trials to have only two adjudicators reviewing each case. A study which investigated how adjudication committee size impacted on the trial results of an international multicentre surgical trial found that reducing the size of the adjudication committee from six to four had minimal impact on trial findings[78]. This conclusion is supported in an article by Walter et al.[54], in which there was excellent agreement between adjudicators, with the article suggesting that results would have remained consistent if the number of adjudicators had been reduced from five to three or even down to two. Both of these studies mentioned that a large amount of resource, in terms of direct cost and time-associated costs, could have been saved with a reduced committee size[54, 79]. A simulation study by Kahan et al.[56] investigated various adjudication methods and summarised that whilst increasing the number of adjudicators assessing each outcome will reduce bias, “the extent of this benefit

depends on the true ICC"[56]. Here the intraclass correlation coefficient (ICC) provides a measure of the agreement between adjudicators.

Whilst adjudication committee size differed between trials, adjudicators are predominantly blinded to treatment arm[42, 55, 60], but were not always blinded to site assessment[60]. Enabling adjudicators to know site assessment could introduce unnecessary bias into the adjudication process, and could potentially jeopardise the blinded nature of the adjudicators. However, the majority of the adjudicators were independent from the trial[42, 55, 60], providing some reassurance that adjudication was not implemented to change trial results in favour of a novel intervention.

The independence of adjudicators has been called into question by a report carried out by Serebruany and Fortmann[45] which claimed that the majority of the adjudicators employed in a cardiovascular trial had strong links to the trial sponsor and primary investigator. The adjudication process in this trial did in fact change the trial results from neutral to favour the trial intervention, and the report suggests that adjudication could have influenced the trial outcomes as well as commenting that independence can and should be a feature of all adjudication committees. Additionally, Serebruany[80] suggested that the delays in the adjudication process in a cardiovascular trial "lack reasonable explanation" and that these delays could have been introduced in an effort to influence the results of clinical trials. Moreover, Serebruany and Atar[81] questioned the findings of three randomised trials in which adjudication had changed the results in each case in

favour of the new intervention, and suggested that independent audits should take place when there is large disagreement between site-assessors and central adjudicators. Therefore, bot independence and blinding of the adjudicators must be taken into account when assessing how valid the adjudication process is, and also how valid the results of a trial are.

Another key consideration when determining the effectiveness of an adjudication committee, is to identify what information about trial participants was provided to the committee. Results of tests and procedures that were performed in the trial were provided to just over half of the adjudication committees in the trials identified in a review by Stuck et al.[55]. A separate review by Ndounga Diakou et al.[60] found that standard case report forms were given to adjudication committees in 55% of trials and in 15% of trials adjudicators were provided with all or some elements of participants' medical files.

Two reviews that reported the characteristics of adjudication in clinical trials found that the decision process used by adjudicators was poorly documented[42, 55]. A description of how a consensus was achieved was reported for just over half of the trials included in one of the reviews[42], with the whole committee reviewing each case for 41% of the trials included. The other review established that a full consensus decision was the most common method of a committee reaching a decision, but that having two independent adjudicators assessing each case was still common[55]. If two adjudicators adjudicated the same event, then either the adjudicators agreed a joint consensus if they originally disagreed, or

a third adjudicator was asked to make the final decision after original disagreement[55]. In addition, some adjudication processes involve a degree of quality control that is separate to adjudicating in duplicate as described above[82]. This quality control procedure is less time consuming then adjudicating in duplicate, but could lead to similar inaccuracies to the process of having each case adjudicated by only a single adjudicator.

The final thing to consider when evaluating the best method of adjudication is to assess what outcomes were adjudicated and who provided these to the adjudication committee. Most trials that use adjudication have adjudicators assess suspected events identified by the site investigators[42, 55, 60]. Other methods that were sparsely reported include using a computer algorithm to identify suspected events, adjudicating all patients or adjudicating all events[60]. Whilst adjudicating events that are identified by site investigators occurred for nearly all the trials identified by each review, this appears to be the method most prone to bias. Adjudicators are only able to confirm or refute the outcomes and are unable to add any additional outcomes that may have been missed by site assessors, causing under-reporting[48]. Granger et al. suggest a possible alternative, where adjudicators also review “events that are ‘triggered’ by some feature that raises the suspicion of an event”[43], and a few of the trials identified in one review did use a computer algorithm to do this. Whilst this method, or the completely unbiased method of adjudicating all participants, should yield more reliable results, it is also far more

resource intensive. There does not appear to be consensus as to whether the gains are worth the extra resource and time needed to implement any of these potentially ‘superior’ methods.

It is important to note that there have been technological developments that could reduce the cost of the adjudication process, or provide an alternative way of applying adjudication. A generic outcome adjudication database model has been developed and trialled in a number of randomised trials. This adjudication module provides automated co-ordination of the adjudication procedure. However, whilst this process may reduce the cost of the adjudication process, it may not help to identify additional events or be suitable for trials with a more complex adjudication procedure[83]. The use of registry-based studies has benefitted from central adjudication to identify complex outcomes[84], but for trials with simpler outcomes, the use of public registers could be a quicker and cheaper alternative to central adjudication and this approach could be adopted in randomised trials[85].

1.7 Adjudication in Cardiovascular Trials

Adjudication appears to be particularly prevalent in cardiovascular randomised trials. As mentioned previously, over 80% of cardiovascular trials are reported to include adjudication[42]. For venous thromboembolism trials, outcomes were adjudicated in the majority of prevention and acute treatment trials (70% and 65% respectively)[55].

There has been a substantial amount of research on the effect of adjudication of cardiovascular outcomes[86, 87], in particular adjudication of myocardial infarction[43, 44, 50, 79, 88-90]. Mahaffey et al. describe the central adjudication in a large cardiovascular trial that randomised 10,948 participants with acute coronary syndromes to receive either a cyclic heptapeptide or placebo[79, 89]. The primary outcome for the trial was mortality or myocardial infarction assessed 30 days post randomisation, and was evaluated by site investigators and two central adjudicators. Computer algorithms were developed to systematically identify variables that would suggest occurrence of a myocardial infarction. In total, the primary outcomes for 5005 participants with suspected myocardial infarction were adjudicated by the central adjudicators, and this process found an additional 816 participants with a myocardial infarction that had been missed by site assessors. Mahffey et al. comment that "adjudication of suspected nonfatal MI end-point events is important to provide independent, unbiased, standard, systematic assessments"[89].

However, a meta-analysis of 10 cardiovascular trials carried out at McMaster University between 1993 and 2006 that had adjudicated their primary outcome found that adjudication had no significant effect on study conclusions[44]. The majority of the trials included had blinded outcome assessment and the outcomes were commonly assessed major clinical events, such as myocardial infarction or stroke, and Pogue et al.[44] question "whether the efforts and costs (of adjudication) justify the gains". It appears that adjudication may be an added level of

complexity that is not needed in all clinical trials. Hata et al.[50] evaluated the usefulness of an adjudication committee in a large cardiovascular trial, which had a composite primary outcome of macrovascular and microvascular events (cardiovascular death, nonfatal stroke, nonfatal myocardial infarction, new or worsening nephropathy and retinopathy) adjudicated. All site-assessed suspected primary outcomes and deaths were assessed by the adjudication committee. Central searches were also undertaken in an attempt to identify additional primary outcomes that had been missed by site investigators. In total, site assessors reported events for 2443 participants, with 85% of these participants confirmed to have an event after adjudication. An additional 48 events were added through searching of the trial database. However, analysis using either adjudicated or unadjudicated data resulted in similar treatment effect estimates for all outcomes in question. Hata et al. conclude that the "adjudication process used … had no discernible effect on the main findings of the trial"[50].

As well as the primary outcome of trials undergoing assessment by adjudicators, safety outcomes such as serious adverse events are also regularly adjudicated[91]. In cardiovascular trials, central adjudication is often used to determine cause of death[51, 92-94] and adverse events such as bleeding[58]. In an orthopaedic trial, self-reported cardiovascular serious adverse events such as stroke and myocardial infarction were adjudicated[95]. The study found that adjudication was vital to the trial results and that this impacted on the interpretation of the trial findings, with only 58% of self-reported strokes verified by adjudicators.

Whilst there is conflicting evidence about the usefulness of adjudication in cardiovascular trials, there are guidelines in place from both the FDA and EMA which recommend the use of an adjudication committee to assess cardiovascular outcomes in diabetes phase 2 and 3 trials[96, 97]. Therefore, a strong level of evidence would be needed to make the case that adjudication of cardiovascular outcomes is unnecessary, and currently there does not appear to be the evidence base to question this guidance. Thus, in the majority of trials that include cardiovascular primary or safety outcomes, these undergo adjudication.

1.8 Evidence of Adjudication in Stroke Trials

Whilst there is a plethora of research on adjudication in cardiovascular trials, as described above, one cardiovascular area that has lacked as much attention is the impact of adjudication on stroke outcomes. Occurrence of stroke, type of stroke, bleeding outcomes, cause of death and functional outcome after stroke are commonly adjudicated in stroke trials. Ninomiya et al.[51] evaluated the effects of the end point adjudication process on a stroke trial, where adjudicators analysed type of stroke and cause of death. Site investigators and central adjudicators had good agreement on type of stroke (kappa, κ=0.53, 95% C.I: 0.49 to 0.57), with highest agreement being for diagnosis of haemorrhagic stroke (94%). There was also excellent agreement between adjudicators and local site investigators for cause of death, with agreement being 94% and 93% for Cardiovascular and Cancer death respectively. Ninomya et al. investigated whether adjudication had any impact on the estimated hazard ratios for the effect of treatment on

occurrence of stroke and found almost identical estimates using adjudicated and unadjudicated data. The study concluded that “the end point adjudication process … had no substantive impact on the effect estimates obtained from the trial” and that “with clinical research costs escalating, it is a requisite on researchers to ensure that scarce resources are applied as effectively as possible” as well as asking for “formal quantitative estimates of the likelihood of an adjudication process influencing trial conclusions”[51].

In addition to the study by Ninomiya et al., a post hoc analysis of a secondary prevention stroke trial by Easton et al. was recently published[98]. This study had a primary endpoint that includes stroke, myocardial infarction and death. All endpoints were assessed by a blinded adjudication committee. The primary outcome was almost unchanged when comparing the adjudicated data to the site investigator data, with hazard ratios and 95% confidence intervals being 0.89 (0.78 to 1.01) and 0.88 (0.78 to 1.00) respectively. The crude agreement between adjudicators and investigators was 91% for ischaemic stroke and 88% for major haemorrhages. The study concluded that “double-blinded treatment outcome studies with a stroke endpoint, or a composite of stroke and death, may not benefit from adjudication”[98].

In many acute stroke trials, the primary outcome used is functional outcome measured on the modified Rankin scale (mRS). The mRS is a single item ordinal rating scale that runs from 0 to 6, with 0 referring to ‘no symptoms at all’, 3 representing ‘moderate disability; requiring some

help, but able to walk without assistance' and 6 indicating 'death' (Table 7)[99]. In recent years functional outcome has received attention in an attempt to improve its reliability using adjudication[100]. The mRS is usually scored in a face to face consultation between clinician and patient, and this can be adjudicated through a video recording of the meeting[101]. This video footage can then be adjudicated by blinded experts, who also score the assessment. A simulation of the adjudication of mRS in stroke trials found that increasing the reliability of mRS scores using adjudication could result in up to 20% reductions in sample size[102]. Whilst reducing sample size would result in faster trials, there are also economic considerations, and improving the reliability of measurements would lead to extensive cost-saving[103].

Table 7 Modified Rankin Scale (mRS)

Score	Description
0	No symptoms at all
1	No significant disability despite symptoms; able to carry out all usual duties and activities
2	Slight disability; unable to carry out all previous activities, but able to look after own affairs without assistance
3	Moderate disability; requiring some help, but able to walk without assistance
4	Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance
5	Severe disability; bedridden, incontinent and requiring constant nursing care and attention
6	Dead

Adjudication of stroke type has been explored in an international acute stroke trial, with the authors investigating its impact on functional outcome[52]. The authors found that using unadjudicated outcome as a

covariate in the primary analysis model gave identical point estimate and confidence interval. Furthermore, increasing the amount of misclassification by 20 times the amount witnessed in the trial through simulation still gave similar results to those obtained using the adjudicated data. They found that central adjudication of diagnosis would have only changed the trial findings for a subgroup analysis by the adjudicated diagnosis, if misclassification was high (10 times the amount seen in the trial) and there was a strong treatment-diagnosis interaction[52].

In addition to the research described above, one other study has investigated central adjudication of the mRS in an acute stroke trial[104]. In this study, local site assessment was carried out through face-to-face interviews, with site assessors blinded to treatment allocation. Central adjudication used both structured phone interviews and video recordings of the site-assessors interviews. The study found that odds ratios from the primary analysis model were similar when using the outcome determined by site investigators or the outcome provided by the central adjudicators.

As mentioned previously, £56M of government and charity funding was allocated to stroke research in the UK during 2012, and it is the aim of the Stroke Association to double to research funding assigned to stroke by 2020. Furthermore, the REWARD campaign through the European Stroke Organisation Trials Network Committee mentions stroke research as one area where improvements in efficiency can be made[105]. This campaign aims to “identify sources of waste in the way

stroke research is chosen, designed, conducted, analysed, regulated, managed, disseminated, and reported". Therefore, it is important to identify the scenarios where adjudication plays an important role, to ensure new trials produce high quality evidence, and where adjudication is not needed, so that resources are used in an efficient manner.

1.9 Research Question

As described previously, adjudication of many cardiovascular outcomes has received a large amount of attention in the literature, but there is a lack of evidence about the importance of central adjudication of the primary outcome in stroke trials. Furthermore, no studies that I am aware of have attempted to pool adjudication data from multiple stroke trials. Currently, there a large number of different outcomes in stroke trials that are adjudicated as well as different types of stroke trials that involve an adjudication process. Therefore, it is important to understand if, and when, adjudication of a trial's primary outcome is beneficial for studies investigating stroke.

Furthermore, it would be desirable to establish in what circumstances adjudication of outcomes has the potential to most impact on study conclusions and when it is not an important process for stroke trials. Moreover, there has been little to no assessment of the cost of adjudication in stroke trials, and no work on identifying the most effective and most cost-benefit methods of adjudication. Based on the

gaps in the literature identified, this thesis will aim to address the following research question:

What are the benefits and costs of adjudicating outcomes in stroke trials?

1.10 Aims of this thesis

In order to address this research question, the aims of this thesis are as follows:

1. Describe which types of stroke trials have adjudicated outcomes, what outcomes are adjudicated and how adjudication has been conducted
2. Investigate the effects on the primary results of stroke trials based on analysis of outcomes assessed by central adjudicators compared with outcomes assessed by site investigators
3. Evaluate when central adjudication of the primary outcome of a stroke trial would change the primary treatment effect estimate
4. Measure the cost of central adjudication in randomised stroke trials and evaluate this cost alongside the number of events corrected by adjudication

In addition, this thesis will also explore:

5. Whether the format of emails to contact authors has any effect on the response rate when attempting to acquire unpublished data for a systematic review

The next chapter provides a description and review of the methods used in this thesis to address the research question and meet the aims listed above, as well as giving a rationale for why these methods were implemented.

Chapter 2 – Methods

2.1 Introduction

This chapter introduces the methodologies used in this thesis and gives a critique of these methods, providing a rationale for why the methods were selected for this research programme. The chapter will cover systematic reviews, meta-analysis, measures of agreement including percent agreement and kappa, and methods to establish whether treatment estimates differ from each other including the ratio of treatment effects and tests of homogeneity. A more detailed description of the specific methods used in each chapter of this thesis will then be presented in the methods sections of Chapters 3-7.

2.2 Systematic Reviews

A systematic review involves a methodical search of the literature to identify all of the evidence in order to answer a research question or evaluate a specific hypothesis[106]. A systematic review follows a set of eligibility criteria in order to determine which records to include, as well as making the review both reproducible and updateable. If the systematic search of the literature identifies the majority of relevant articles, then bias will be reduced and the review will produce reliable findings[107].

2.2.1 Search Strategies

Defining the correct search strategy is paramount to the success of the review. Search strategies should be comprehensive enough to identify

all of the research around a specific topic (review sensitivity), but also precise enough to ensure that the reviewer(s) do not spend excessive time screening irrelevant records (review specificity). Search strategies can be built using Boolean operators (e.g. AND, OR, NOT, AND NOT). A search strategy is often developed with assistance of an information specialist or librarian, and search strategies can be tested using a scoping search to ensure that the search strategy has adequate sensitivity and specificity[108].

2.2.2 Electronic Databases

Once developed and tested, the search strategy can be implemented using electronic databases. Database searching is a time-consuming process, partly due to each database having its own unique search syntax. Multiple databases are often recommended to identify all of the literature to answer the systematic review research question[109, 110]. For reviews that search for randomised trials, MEDLINE, Embase and the Cochrane Central Register of Controlled Trials (CENTRAL) are advised as the minimum set of databases to search[106]. However, searching a large number of databases is a laborious process, as search strategies need to be translated and the number of potentially irrelevant records increases[111]. Bramer et al. recommends searching in at least Embase, MEDLINE, Web of Science and Google Scholar to “guarantee adequate and efficient coverage”[111].

Through searching electronic databases, in addition to published material such as journal articles and books, conference proceedings can be retrieved. Depending on the review question, the conference

abstract may contain enough information to determine inclusion. A substantial proportion of conference abstracts are never published, and so identifying and retrieving them may help to reduce publication bias[112].

2.2.3 Identifying Grey Literature

Grey literature refers to materials that are not controlled by commercial publishing organisations and thus publishing the materials are not of paramount importance[113]. Examples of grey literature are internal reports, working papers and newsletters. Given that these materials are not regularly published, this often leads to grey literature being missed from searches that exclusively use electronic databases to identify records.

Therefore, to identify grey literature, alternative methods need to be implemented to find relevant studies. These methods do not differ extensively from normal search methods and include looking in databases of grey literature, databases of media reporting, targeted email requests to key contacts, internet search engines (e.g. Google Scholar), relevant funder and third sector websites, and use of social media (e.g. twitter)[113, 114]. The time taken to identify this additional material is not trivial, with a study by Saleh et al.[115] finding that searching for grey literature takes up approximately 27% of the entire search time for a review. Haddaway et al. suggest that a simpler and more time-efficient approach is to search Google Scholar for the first 200-300 records, and that this can identify sufficient grey literature[116].

2.2.4 Handsearching and checking citations of included articles

Handsearching involves searching through resources that are not indexed in electronic databases (e.g. medical journals or conference abstract booklets). In addition to handsearching, the citations of already included articles can be assessed to determine whether any are relevant for the review. Both of these practices can add value to a review, and in particular handsearching has been shown to identify reports of randomised trials that would otherwise be missed[117, 118], but with limited time and resource, using an appropriate search strategy in a suitable electronic database will identify the majority of eligible records[117].

2.2.5 Languages to include

In systematic reviews, studies can be restricted by language[119]. Many reviews limit eligible articles to those only written in English[120, 121], and whilst this does exclude a number of potentially eligible and relevant records, it does not appear to bias findings[120]. Thus, if there is a lack of time and/or resources then restricting the language to English will reduce the number of records that need to be assessed, without missing a large number of eligible records. However, non-English articles should be recorded so an assessment into the number of potentially eligible records missed through language restrictions can be made.

2.2.6 Managing records and de-duplication

Records can be managed using reference manager software, with common software packages including EndNote, Mendeley and RefWorks. A search strategy that involves multiple databases will inevitably lead to duplicate records being identified. These duplicates should be removed prior to screening. An advantage of reference manager software packages is in the provision of automated de-duplication tools. This auto-searching can be effective[122], but should be complemented with hand-searching methods to identify and remove duplicates before the commencement of title and abstract screening[123].

2.2.7 Screening records

Systematic reviews can be labour intensive, with reviews commonly taking over a year to complete[124]. A large portion of this time is taken up with screening records. The first stage of screening is either title and abstract screening or title screening alone. Screening titles alone can be more efficient than screening titles and abstracts[125]. However, if reference manager software is used then there is less benefit from screening titles only, and screening titles and abstracts together can be a time-efficient process[126]. It is important to have independent reviewers screening records, with one reviewer assessing a subset of the records if there are time or resource constraints on the review[126].

2.2.8 Data extraction

Once screening is complete, data extraction follows in which the reviewers collect information from each included study. Errors in data

extraction are common, with these errors having the potential to bias the review findings[127]. To limit data extraction errors, data should be double-extracted for all included studies by two independent reviewers[128].

2.2.9 Narrative synthesis

After data extraction, it is common to carry out a narrative synthesis of the included studies. The synthesis involve a textual summary of the study findings and a description of differences between the various study characteristics[129]. The extent of the narrative can range from a brief summary of the included studies to a more detailed synthesis of the data. A detailed synthesis can include[130]:

- An initial synthesis of the findings from the included studies
- An exploration of relationships in the data extracted
- Information about the strength of evidence for the review question

Narrative syntheses have been criticized as they can introduce bias into the review findings through lack of transparency[131]. However, if quantitative synthesis is not possible, a narrative synthesis does provide an alternative approach that enables some evaluation of the research question. In general, where possible, quantitative synthesis is advised to ensure findings are lacking in bias and are as robust as possible[132].

2.10 Summary of systematic review methods

In summary, a number of different electronic databases including a minimum of MEDLINE, Embase and CENTRAL as well as grey literature searching using Google scholar should provide suitable coverage for a review aiming to identify randomised trials and secondary analyses of randomised trials. A two-stage screening process of titles and abstracts followed by full texts should be time-efficient given the large number of records anticipated. A subset of these titles and abstracts and full texts should be checked to ensure adequate agreement before proceeding with the remainder of the screening. Data extraction should be performed in duplicate, but, given the quantitative nature of the primary analysis of stroke trials, provided there are sufficient studies, data should be appropriate to be pooled in a quantitative synthesis i.e. a meta-analysis.

2.3 Meta-analysis

As described in the previous section, a systematic review is the process of retrieving all the evidence about a specific research question in order to produce a robust answer to the question. Meta-analysis is the approach of quantitatively synthesising data from multiple studies. If data are not appropriate for quantitative synthesis then only a narrative synthesis can be performed.

In this section, meta-analysis methods for binary data will be summarised since the majority of primary outcomes in stroke trials are binary (e.g. stroke or a composite including stroke). This section will

follow terminology from Deeks et al.[133] Meta-analysis methods used to synthesise odds ratios (OR) will be described here. Meta-analysis methods relating to synthesising risk differences, risk ratios or continuous variables can be found here[133]. Table 8 displays the 2x2 table for the binary outcome of stroke, from the i^{th} trial ($i = 1, \dots, n$) that is to be synthesised in a meta-analysis.

Table 8 Cell counts for the outcome of stroke in trial i

	Stroke	No stroke	Total
Intervention	a_i	b_i	n_{1i}
Control	c_i	d_i	n_{2i}

The OR for the i^{th} study and standard error (SE) for the natural logarithm (ln) of the OR are:

$$OR_i = \frac{a_i d_i}{b_i c_i}$$

$$SE[\ln(OR_i)] = \sqrt{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}}$$

A meta-analysis for all n included studies will estimate a pooled effect that uses each odds ratio. There are two main approaches to estimate this pooled effect, using a fixed effect meta-analysis or a random-effects meta-analysis.

2.3.1 Fixed effect meta-analysis

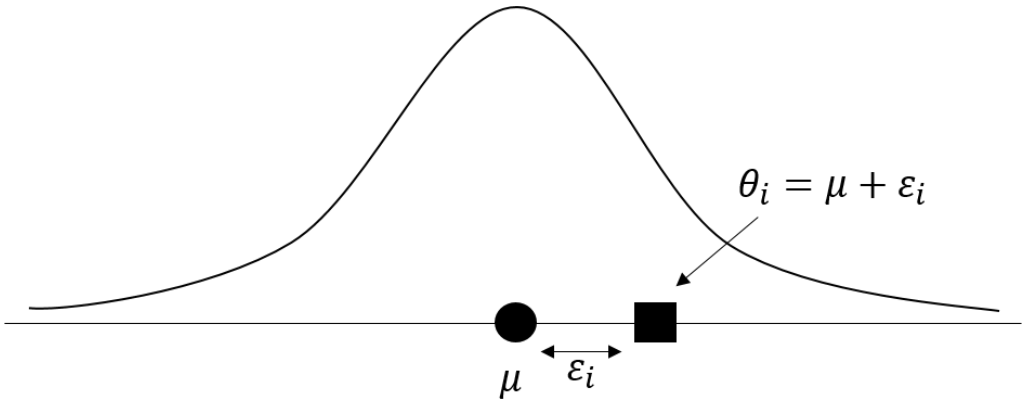
A fixed effect meta-analysis makes the assumption that all n included studies are estimating the same treatment effect, μ, referred to as the pooled effect[134]. In reality, the i^{th} trial produces a treatment effect, θ_i

that deviates from the pooled effect by an amount, defined as the within-study error ε_i. The model assumes:

$$\theta_i = \mu + \varepsilon_i$$

Each θ_i is assumed to be normally distributed with mean μ, and variance τ^2. A graphical representation of the fixed effect assumption is displayed in Figure 3.

Figure 3 Fixed effect meta-analysis



A variety of methods are available to estimate μ, with three methods suitable for binary outcomes discussed next.

2.3.1.1 Inverse variance method

In the inverse variance method, the effect sizes (θ_i's) are combined using a weighted average of the individual treatment effects. Thus:

$$\mu = \frac{\sum_{i=1}^{n} w_i \theta_i}{\sum_{i=1}^{n} w_i}$$

The weights are the reciprocals of the variances for the individual studies:

$$w_i = \frac{1}{SE(\theta_i)^2}$$

The consequence of this weighting is to give more weight to larger studies that typically have lower standard errors when computing the pooled effect. The standard error for the pooled estimate is specified as:

$$SE(\mu) = \frac{1}{\sqrt{\sum_{i=1}^{n} w_i}}$$

Using the standard error, 95% confidence intervals can be computed for the pooled effect in the normal way. The inverse variance method is an attractive approach to weighting the individual treatment effects since it can be used to combine any type of data (continuous, binary, ordinal etc.), as long as there are standard errors available[133]. This gives the approach wide application for use.

2.3.1.2 Mantel-Haenszel method

An alternative to the inverse variance method is one proposed by Mantel and Haenszel[135]. This method is preferable for trials when event rates are low or trial samples sizes are small, and gives similar estimates to the inverse variance method otherwise. The estimate of μ is given by the same formula as in the inverse variance method, but the Mantel-Haenszel method uses an alternative approaching to weighting the estimates. This approach is only suitable for binary outcomes. Consider the generic example given in Table 8, these weights are given by:

$$w_i = \frac{b_i c_i}{n_{1i} + n_{2i}}$$

The standard error for the natural logarithm of the pooled estimate, here an odds ratio, is:

$$SE[\ln(\mu)] = SE[\ln(OR)] = \sqrt{\frac{1}{2}\left(\frac{E}{R^2} + \frac{F+G}{R \times S} + \frac{H}{S^2}\right)}$$

with:

$$R = \sum_{i=1}^{n} \frac{a_i d_i}{n_{1i} + n_{2i}}, \qquad S = \sum_{i=1}^{n} \frac{b_i c_i}{n_{1i} + n_{2i}}, \qquad E = \sum_{i=1}^{n} \frac{(a_i + d_i) a_i d_i}{(n_{1i} + n_{2i})^2},$$

$$F = \sum_{i=1}^{n} \frac{(a_i + d_i) b_i c_i}{(n_{1i} + n_{2i})^2}, \qquad G = \sum_{i=1}^{n} \frac{(b_i + c_i) a_i d_i}{(n_{1i} + n_{2i})^2},$$

$$H = \sum_{i=1}^{n} \frac{(b_i + c_i) b_i c_i}{(n_{1i} + n_{2i})^2}$$

2.3.1.3 Peto's method for odds ratios

A third method, which is only available to pool odds ratios, was suggested by Yusuf et al.[136] This method is similar to the inverse variance method, with μ, here a pooled OR, given by:

$$\mu = OR = \exp\left(\frac{\sum_{i=1}^{n} w_i \ln(OR_i)}{\sum_{i=1}^{n} w_i}\right)$$

Here, each OR_i is calculated using the approximate Peto method[133, 136]. Unlike the other two methods, Peto's method can produce biased pooled estimates if there is severe imbalance between treatment arms or if the observed treatment effects are very large[137]. Thus, Peto's method may not be suitable for some situations.

2.3.2 Assessing heterogeneity

All fixed effect approaches discussed so far assume that the treatment effect is the same for all the included trials, with any difference due to chance alone. This is a strong assumption, and it is possible that there are differences in the treatment effect between trials due to differences in the trial characteristics (e.g. patients recruited, disease setting, intervention etc.). These differences are referred to as the heterogeneity of the studies.

Cochran developed a test to determine if the included studies were sufficiently homogenous to be pooled in a fixed effect meta-analysis[138]. However, the test is criticised due to its relatively low power to detect variation between studies[139]. Furthermore, the test does not quantify the heterogeneity between the included studies, and only produces an estimate for the level of significance of the overall heterogeneity. For completeness, the heteregoenity statistic (Q) for the inverse variance and Mantel- Haenszel method is given by:

$$Q = \sum_{i=1}^{n} w_i(\theta_i - \mu)^2$$

Under the hypothesis of homogeneity, the Q statistic follows a chi-squared distribution with $n - 1$ degrees of freedom, and this test is used to determine statistical significance.

An alternative measure that quantifies the heterogeneity was suggested by Higgins and Thompson in 2002[140]. The measure, known as I^2, uses

the Q statistic to determine the extent of heterogeneity and is calculated as follows:

$$I^2 = \begin{cases} \dfrac{Q-(n-1)}{Q} & \text{for } Q > (n-1) \\ 0 & \text{for } Q \leq (n-1) \end{cases}$$

If Q is less than or equal to the degrees of freedom $(n-1)$, then I^2 is constrained at the minimum value. Otherwise I^2 is bounded above by 1. Thus, I^2 is contained within the range $0-1$. For ease, I^2 is usually given as a percentage, with higher values indicating a more heterogeneous sample of studies. A very rough guide to interpret I^2 is given by the Cochrane handbook[141], with 0-40% indicating that heterogeneity may not be important, 30-60% suggesting moderate heterogeneity, 50-90% representing substantial heterogeneity and 75-100% referring to considerable heterogeneity. However, Higgins et al. urge caution with these categories and specify that the observed value of I^2 should be interpreted using the confidence interval for I^2 as well as taking into account the magnitude and direction of treatment effect for the included studies[142].

2.3.3 Random effects meta-analysis

If sufficient heterogeneity is present, then a fixed effect meta-analysis may not be suitable and can bias the pooled effect. In this circumstance, a random-effects meta-analysis is appropriate. The random effects model was proposed by DerSimonian and Laird[143] in 1986. This model relaxes the assumption that each study has a

common treatment effect, and instead denotes the treatment effect of the i^{th} study by μ_i. The random-effects model is as follows:

$$\theta_i = \mu_i + \varepsilon_i, \qquad \mu_i = \mu + \eta_i$$

Here, η_i is the error that accounts for between-study variability, with $\eta_i \sim N(0, \tau^2)$ and $\theta_i \sim N(\mu, \tau^2 + \sigma_i^2)$. τ^2 is estimated as:

$$\tau^2 = \frac{Q - (n-1)}{\sum_{i=1}^{n} w_i - \left(\frac{\sum_{i=1}^{n} {w_i}^2}{\sum_{i=1}^{n} w_i}\right)}$$

As in Section 2.3.2, Q is the statistic associated with heterogeneity and $\tau^2 = 0$ if $Q < (n-1)$, with w_i calculated using the inverse variance method. The treatment effect of the i^{th} study is given weighting:

$$w_i' = \frac{1}{SE(\theta_i)^2 + \tau^2}$$

and then the pooled treatment effect and standard error for the random effects model are calculated using these new weights as:

$$\mu = \frac{\sum_{i=1}^{n} w_i' \theta_i}{\sum_{i=1}^{n} w_i'}, \qquad SE(\mu) = \frac{1}{\sqrt{\sum_{i=1}^{n} w_i'}}$$

When $\tau^2 = 0$, the weights given above simplify to those used in the fixed effect inverse variance model, that is $w_i = w_i'$. However, if $\tau^2 > 0$, then the weights in the random effect model will be lower than those in the equivalent fixed effect model. This means that a random effects model with $\tau^2 > 0$ gives more weight relatively to smaller studies[133].

Choosing between the random effects and fixed effect model is of vital importance when undertaking a meta-analysis, both in terms of

interpretation of the pooled estimate and in obtaining the value of the estimate itself. The assumptions of the fixed effect meta-analysis are strong, and likely to be invalid in many meta-analyses, so the random effects model appears the preferred choice in general[144]. The random effects model allows greater generalisability of results, relaxes the assumption around common effect size, has a greater likelihood of fitting the observed sampling distribution and when $\tau^2 = 0$ gives exactly the same result as the fixed effect model[144]. A common and reasonably sensible approach would be to use a random effects meta-analysis, and fit a fixed effect meta-analysis as a sensitivity analysis[145].

2.3.4 Exploring heterogeneity

In addition to expressing heterogeneity using I^2, it can be useful to explore specific characteristics of the included studies that contribute to the observed heterogeneity. Two methods include subgroup analysis (for categorical variables, or continuous variables that are categorised) and meta-regression (for continuous variables). Both methods should be pre-specified, and they allow for an investigation into the heterogeneity between studies (if it exists)[146].

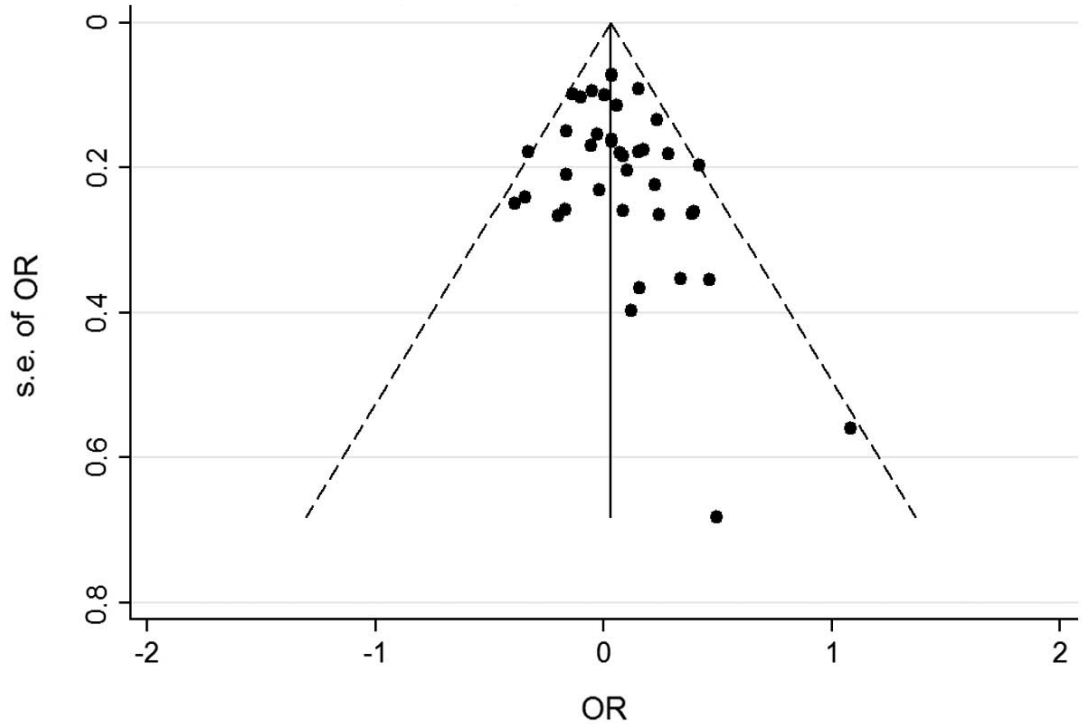
2.3.5 Publication bias

Publication bias exists in a meta-analysis and can lead to a pooled estimate that differs considerably from the underlying true value of the estimate. It is due to the scientific culture of greater weight being given to positive results, including publication in high impact journals and media coverage, and less weight being given to negative findings,

some of which are never published let alone in high impact journals[147]. This bias leads to, often smaller, and usually negative, studies being missed from the meta-analysis, which will impact the estimated pooled effect.

To identify publication bias, a funnel plot can be drawn, which plots individual effect estimates against the inverse of the precision of that estimate (directly linked to sample size)[148]. This leads to a funnel shape, with a small number of large studies at the top and a greater number of smaller studies that should be widely dispersed around the pooled effect. If there is bias present, then the funnel plot will not be symmetrical[149]. An example funnel plot is given in Figure 4. There appears to be missing studies in the bottom left section of the graph (small negative studies). However, with the addition of one study in this area, it would become challenging to identify publication bias, and thus these graphs are considerably subjective.

Figure 4 Example funnel plot



Original image from https://stats.stackexchange.com/questions/193260/minimal-requirements-for-funnel-plots
OR refers to odds ratio and is given on the log scale, s.e. refers to standard error

2.3.6 Methodological meta-analysis

The methods described above have been developed for meta-analyses of intervention studies. For meta-analysis of methods studies, sometimes termed as meta-epidemiological studies[150], whilst most of the methods described are still applicable, there are a number of key differences.

For example, assessment of publication bias using methods such as the funnel plot or other statistical tests not mentioned in this thesis, is not possible since these methods have not been translated into meta-epidemiological research[151]. Furthermore, meta-epidemiological studies are less common and so a meta-analysis of these methods

articles will often find fewer eligible studies to include, leading to lower statistical power[152].

In 2017, Murad and Wang[153] proposed reporting guidelines for meta-epidemiological studies, which have adapted elements from the PRISMA guidelines to allow for the specifics of a meta-analysis of methods articles. A checklist, similar to the PRISMA checklist is also provided, to help ensure that meta-epidemiological research is clearly reported.

2.3.7 Summary of meta-analysis methods

In summary, given the strong assumptions around the fixed effect meta-analysis, and potential heterogeneity if treatment effect estimates differ considerably between adjudicators and site investigators for some trials and not others, then a random effects meta-analysis would be appropriate to select as the choice for the primary analysis. A fixed effect meta-analysis using the inverse variance method (due to its ease of combining many different types of data) would be used as a sensitivity analysis. Heterogeneity can be assessed using I^2 and this can be explored using subgroup analysis (for categorical or dichotomised continuous variables) and meta-regression (continuous variables). Assessing publication bias in a methodological meta-analysis is challenging, and without clearly established methods, it is not be possible to assess this statistically or graphically. However, description of the trials included and other trials which are identified but do not have data available, allows a comparison between available and

unavailable data, enabling some evaluation into the possibility of publication or data availability bias.

2.4 Measures of agreement

The main aim of this thesis is to determine the costs and benefits of adjudicating outcomes in stroke trials. To assess the benefits, it is important to evaluate how different the adjudicators' outcome is to the site investigators. To quantify this difference, it is useful to assess the agreement between both groups using various measures of agreement. To determine how well raters, in this case adjudicators and site investigators, agree, it is assumed that the central adjudicators' outcome is gold standard. This assumption is made throughout the thesis, meaning that any disagreement is assumed to indicate that the site investigators are in error. Whilst this may not always be the case, it is an essential assumption to the assessment of the benefit of adjudication. Further comment around this assumption is made in Chapter 8.

In the research discussed in this thesis, measurements (brain scans, video recording etc.) are taken from patients which are then assessed in duplicate by site investigators and central adjudicators. This section will focus on the methods to quantify agreement between these two groups of raters, making the assumption that the central adjudicators' rating is gold standard. Throughout this section, two outcomes will be considered that are commonly adjudicated, stroke (Yes/No, binary) and

functional outcome measured on the modified Rankin Scale (mRS, ordinal 7-point scale, 0-6).

2.4.1 Percent agreement

Throughout this and subsequent subsections, terminology from Fleiss et al.[154] will be used. Consider the general case in which central adjudicators and site investigators both assess the outcome of Stroke (binary). The proportions of patients where both raters agree and disagree on their diagnosis of stroke is given in Table 9.

Table 9 Agreement between central adjudicators and site investigators on stroke

	Central Adjudicators		
Site Investigators	**Stroke**	**No Stroke**	**Total**
Stroke	a	b	p_1
No Stroke	c	d	q_1
Total	p_2	q_2	1

Cells contain proportions and not counts

Here, a is the proportion of patients where both central adjudicators and site investigators agree that the patient has had a stroke, and d is the proportion of patients where both agree that the patient has not had a stroke. These diagonal elements in the table, where both raters agree, are shaded. The off diagonal elements, b and c, indicate where the raters disagree. The overall proportion of agreement can be calculated as:

$$p_0 = a + d.$$

This proportion is usually given as a percentage. Hence, this crude measure of agreement is commonly termed as percent agreement.

Suppose instead that the two groups of raters assess functional outcome (ordinal 7-point scale). Overall agreement is given by summing the diagonal elements of the 7x7 table, in a similar fashion to the binary outcome. To measure the agreement on each specific category of the ordinal scale, for example mRS=2, the 7x7 table first needs to be collapsed into a 2x2 table (see Table 10), and then the same process can be followed as with the binary outcome.

Table 10 Agreement between central adjudicators and site investigators on functional outcome where mRS=2

	Central Adjudicators		
Site Investigators	**mRS=2**	**mRS≠2**	**Total**
mRS=2	a	b	p_1
mRS≠2	c	d	q_1
Total	p_2	q_2	1

Cells contain proportions and not counts

However, suppose the outcome that is being assessed is rare, for example in a population where less than 5% of the sample are expected to have a stroke, then p_0 will be inflated by d. An alternative measure of agreement is given by:

$$p_s = \frac{2a}{2a + b + c} = \frac{a}{\overline{p}}$$

This measure can give a substantially different measure of agreement than that indicated by p_0 if the outcome is rare, see example in Fleiss et al.[154]. Here, $\overline{p} = (p_1 + p_2)/2$, and $\overline{q}$ can be similarly defined as: $\overline{q} = \frac{q_1 + q_2}{2} = d + \frac{b+c}{2}$. A further alternative measure of agreement, λ_r, was suggested by Goodman and Kruskal[155] is defined as:

$$\lambda_r = \frac{(\mathrm{a} + \mathrm{d}) - \overline{q}}{1 - \overline{q}} = \frac{2a - (b + c)}{2a + (b + c)}$$

The statistic λ_r has a maximum value of 1 when there is perfect agreement. Both λ_r and p_s ignore d. Similar indices exist that can be defined disregarding a. For example, $p_s' = \frac{2d}{2d+b+c}$. Rogot and Golberg[156] suggest that both p_s and p_s' can be used together in an index of crude agreement and define:

$$A = \frac{p_s + p_s'}{2} = \frac{a}{p_1 + p_2} + \frac{d}{q_1 + q_2}$$

To illustrate how these measures of agreement differ, a hypothetical example of central adjudicators and site investigators assessing the binary outcome of stroke for 1000 participants in a trial is considered (Table 11). Here, the proportions are given rather than cell counts for ease of demonstration.

Table 11 Agreement between central adjudicators and site investigators in a hypothetical trial of 1000 participants

	Central Adjudicators		
Site Investigators	**Stroke**	**No Stroke**	**Total**
Stroke	100 (0.1)	50 (0.05)	150 (0.15)
No Stroke	20 (0.02)	830 (0.83)	850 (0.85)
Total	120 (0.12)	880 (0.88)	1000 (1)

Cells contain numbers of participants (proportion)

Table 12 uses the hypothetical example in Table 11 to demonstrate how the different measures of crude agreement provide varied results. Given the ease of interpretation of percent agreement, p_0, it is a good choice of index to use. However, it is clear that p_0 overestimates the agreement between the adjudicators and site assessors if the outcome

is uncommon, which is the case in many stroke trials. Furthermore, all the indices used thus far do not account for the possibility that some raters may have been unsure and guessed. Therefore, the agreement presented in Table 12 needs to be interpreted alongside an estimate of agreement by chance, as the real level of concordance is potentially far lower[157, 158]. Thus, if used, percent agreement should be accompanied by an additional measure that provides this estimate.

Table 12 Indices of agreement between central adjudicators and site investigators on stroke for hypothetical trial

Index	p_0	p_s	λ_r	p_s'	A
Agreement	0.93	0.74	0.48	0.96	0.85

2.4.2 Kappa

A statistic that measures agreement over that expected by chance was first given by Cohen in 1960[159]. This statistic, kappa (κ), ranges from agreement worse than chance (κ<0), agreement equal to or better than chance (κ≥0) to perfect agreement (κ=1). Cohen's kappa can be defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Here, P_o is the observed value of the index (the proportion that agree) and P_e is the value of the index expected due to chance alone. Kappa is the ratio of the agreement more than chance ($P_o - P_e$) and the maximum possible excess ($1 - P_e$). Fleiss et al.[154] show that κ can be rewritten as:

$$\kappa = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}$$

For the example in Table 11, this gives $\kappa = 0.70$.

Kappa is regularly used in medical research[160], and is commonly presented alongside percent agreement as a more robust measure of agreement between raters. Due to its widespread use, Landis and Koch[161] categorised kappa to quantify different levels of agreement (Table 13). These interpretations are used to allow kappa to be interpreted by a wide audience, and were meant as a guide and not as strict thresholds. Thus, the categories could be modified based on the outcome being rated or the raters that are involved. Stricter boundaries are often used[158].

Table 13 Interpretation of the kappa statistic (from Landis and Koch[161])

Kappa (κ)	**<0.00**	**0.00-0.20**	**0.21-0.40**	**0.41-0.60**	**0.61-0.80**	**0.81-1.00**
Agreement	Poor	Slight	Fair	Moderate	Good	Excellent

The example above has explored kappa for a binary outcome, or an ordinal or nominal outcome that is dichotomised in order to determine agreement on a specific category of the outcome. In most cases, for an outcome with $k > 2$ categories, it is of interest to determine overall agreement between raters, in this thesis central adjudicators and site investigators. Thus, as with percent agreement, instead of transforming the agreement table into 2x2 format, it is desirable to keep it in kxk format. A generic example for an outcome with k categories is given in Table 14.

Table 14 Joint proportions of agreement between central adjudicators and site investigators on a categorical outcome with k categories

	Central Adjudicators				
Site Investigators	**1**	**2**	**…**	$\boldsymbol{k}$	**Total**
1	p_{11}	p_{12}	$\cdots$	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	$\cdots$	p_{2k}	$p_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{k}$	p_{k1}	p_{k2}	$\cdots$	p_{kk}	$p_{k.}$
Total	$p_{.1}$	$p_{.2}$		$p_{.k}$	1

Kappa is defined as before, but in the general case the observed agreement (as with percent agreement) is given by:

$$P_o = \sum_{i=1}^{k} p_{ii}$$

Similarly, agreement expected due to chance alone can be defined as:

$$P_e = \sum_{i=1}^{k} p_{i.}p_{.i}$$

Table 15 gives an example when central adjudicators and site investigators both assess functional outcome measured on the mRS. Here, crude percent agreement is 72% and $\kappa = 0.67$, which would be interpreted as moderate to good agreement between central adjudicators and site investigators.

Table 15 Agreement between central adjudicators and site investigators on functional outcome in a hypothetical trial of 100 participants

	Central Adjudicators							
Site Investigators	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
0	4	0	0	0	0	0	0	4
1	1	14	2	3	0	0	0	20
2	0	0	13	3	1	0	0	17
3	0	1	1	8	2	5	0	17
4	0	0	0	3	4	3	0	10
5	0	0	0	0	3	9	0	12
6	0	0	0	0	0	0	20	20
Total	5	15	16	17	10	17	20	100

Cell values are number of participants, row and column labels are levels on mRS

The standard error for kappa, and therefore the 95% C.I for the point estimate, can be computed. Furthermore, hypothesis tests that test the null hypothesis that kappa is not equal to a specific value can also be carried out. These are not given in this thesis, in part because kappa will be used as a descriptive measure of agreement between central adjudicators and site investigators. Thus, testing whether this agreement is better than chance is not of interest as it is already known from previous work that high agreement is seen between these raters. Of interest is the strength of agreement between the two groups of raters, and testing whether there is excellent agreement ($\kappa > 0.8$) is secondary to whether trial results differ when the centrally adjudicated outcome is used instead of the site investigators' assessment of the outcome.

All measures of agreement that have been described so far have treated all categorical outcome as nominal. Whilst this is sensible for unordered categorical outcomes, and obviously valid for binary outcomes, for ordinal outcomes, like functional outcome, the ordered nature of the data is not taken into account. Kappa and percent agreement both treat disagreement between one level of the outcome as the same as disagreement between three, four or five levels of the outcome. For an ordinal outcome, it may be of interest to weight any disagreement based on the number of levels of difference between the ordinal outcomes given by the raters. Cohen[162] produced an extension of his original kappa measure to enable weighting. This is termed weighted kappa. The weights w_{ij} are constrained to lie between 0 and 1, with perfect agreement assigned a weighting of 1 and disagreement given lesser weighting to be determined by whoever is measuring the agreement, providing that symmetry is maintained when assigning weightings (i.e. disagreement for two categories is 0.5 for all levels of the outcome). Thus:

$$w_{ii} = 1, \qquad 0 \leq w_{ij} < 1 \text{ for } i \neq j, \qquad w_{ij} = w_{ji}$$

P_o and P_e are defined similarly to before, but now take into account the weighting that is specified. Thus:

$$P_{o(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ii}$$

$$P_{e(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.} p_{.i}$$

Weighted kappa, κ_w, is then defined in exactly the same way as kappa but uses the proportions that take weights into account. Therefore:

$$\kappa_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}}$$

The choice of weighting can have a large impact on the value of the weighted kappa statistic, so clinical guidance or a logical argument is advisable to determine appropriate weights[163]. Common choices of weights include linear and quadratic. The quadratic choice penalises disagreement by more categories by a far larger extent than linear weights, so the choice of weights is important and should reflect the outcome.

In the example displayed in Table 15, weighted kappa would give a better reflection of the agreement as the mRS is an ordinal outcome. Using linear weights, $\kappa = 0.83$ and quadratic weights, $\kappa = 0.92$. Both values of kappa are considerably higher than the unweighted version that was 0.67.

In addition to kappa (commonly referred to as Cohen's Kappa), Fleiss[164] proposed an extension for situations where there are more than two groups of raters (commonly referred to as Fleiss's Kappa). This statistic is discussed in detail in this paper by Fleiss[164], but is not discussed in this thesis. Agreement between outcomes in this thesis will only occur between two groups of raters (central adjudicators and site investigators) and so the Fleiss extension is not required in this research project.

2.4.3 Intraclass correlation coefficient

In the example above measures of agreement are described for categorical outcomes. However, these approaches cannot be used for continuous outcomes. The majority of outcomes that are adjudicated are categorical. To my knowledge there are no continuous outcomes adjudicated in stroke trials, but ordinal outcomes, such as the mRS, could be treated as continuous. Thus, it is useful to consider a number of measures of agreement for continuous outcomes.

As before, consider two groups of raters (central adjudicators and site assessors) that rate the same group of m participants. The intraclass coefficient (ICC) compares the variability of each group of raters outcome with the total variability that is observed for all participants. Shrout and Fleiss[165] discuss six different variants of the ICC to account for different situations. In this thesis one variant of the ICC is described, since the others are unlikely to be suitable when determining agreement between central adjudicators and site investigators. All other variants are described in detail by Shrout and Fleiss[165].

The ICC variant covered in detail in this thesis is ICC(3,1). This makes the assumption that each participant is rated by the same k raters, who are the only k raters used in the study. In addition, this ICC provides an assessment of agreement for a single outcome and not an average rating. For simplicity, ICC(3,1) will be referred to as ICC for the remainder of this section.

A two-way mixed model is used to determine ICC, with participants treated as random effects and the raters as fixed effects. Let x_{ij} denote the i^{th} observation on the j^{th} participant, with $1 \leq i \leq k, 1 \leq j \leq m$. Then:

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}$$

Here, μ is the overall mean of the observations, a_i is the departure from μ of the mean of the i^{th} raters outcome value and b_j is the difference between the j^{th} participants outcome value and the overall mean, with $b_j \sim N(0, \sigma_B^2)$. The difference in the outcome value of rater i when assessing participant j is given by $(ab)_{ij}$. The residual error of the model is denoted as e_{ij}, with $e_{ij} \sim N(0, \sigma_E^2)$. For this ICC, a_i is constrained so that $\sum_{i=1}^{k} a_i = 0$ and $\sum_{i=1}^{k} (ab)_{ij} = 0$ and $\theta_A^2 = \sum_{i=1}^{k} {a_i}^2/(k-1)$[165, 166]. Using this model, the expected mean square can then be used to calculate the estimate of ICC, which has true unknown value ρ (Table 16).

Table 16 Mean square expectations for mixed-model and estimate of ρ

Source of variation	Degrees of freedom	Mean Square	Expected Mean Square
Between participants	$m-1$	MS_B	$k\sigma_B^2 + \sigma_E^2$
Within participants	$m(k-1)$	MS_W	$\theta_A^2 + k/(k-1)\sigma_{AB}^2 + \sigma_E^2$
Between raters	$k-1$	MS_O	$n\theta_A^2 + k/(k-1)\sigma_{AB}^2 + \sigma_E^2$
Error	$(m-1)(k-1)$	MS_E	$k/(k-1)\sigma_{AB}^2 + \sigma_E^2$
Statistic	**Mean square formula**		$\boldsymbol{\rho}$
ICC	$\frac{MS_B - MS_E}{MS_B + (k-1)MS_E}$		$\frac{\sigma_B^2 - \sigma_{AB}^2/(k-1)}{\sigma_B^2 + \sigma_{AB}^2 + \sigma_E^2}$

Using the example given in Table 15 for functional outcome, and treating this outcome as continuous, ICC can be estimated as $\rho = 0.92$.

2.4.4 Limits of agreement

An alternative to ICC, would be to determine limits of agreement between raters. A common method is to use a Bland-Altman plot, in which the difference between raters is plotted against the mean of the two ratings[167]. Horizontal lines can be added to the graph at the mean difference and upper and lower limits of the 95% confidence interval for the mean difference, referred to as the limits of agreement.

However, it is often not advised to treat an ordinal variable with few categories (like mRS) as continuous[168]. A review of the use of mRS in stroke trials found that is was treated either as an ordinal variable, or dichotomised at 1 or 2 and treated as binary, so it seems appropriate to follow these approaches in this thesis[169].

2.4.5 Summary of measure of agreement methods

For this thesis, exploring measures of agreement between central adjudicators and site investigators, kappa is a suitable and robust statistic to choose. Kappa is superior to the other measures of agreement discussed in the present subsection due to the fact that kappa provides a measure of agreement over that due to chance and the data collected in this thesis will most likely be binary or categorical (ordinal). However, kappa can be hard to interpret, and therefore percent agreement will be presented in addition to kappa, to provide a measure of crude agreement that is both easy for the reader to interpret and also for the reader to calculate (if they so wish). For ordinal variables, such as the mRS, weighted kappa should be used, with weightings agreed with clinical input.

2.5 Methods to establish whether treatment effect estimates differ for outcomes assessed by different raters

In addition to measuring agreement between raters, it is of interest to establish whether differences in assessing an outcome produce treatment effect estimates that differ between raters. For a randomised trial, this issue is arguably more important than whether the agreement between raters is good or poor. Two methods to determine whether treatment estimates differ are discussed in this section.

2.5.1 Test of homogeneity

A test of homogeneity tests whether the treatment effect estimate produced from the outcome assessed by the first rater is significantly different to the treatment effect estimate produced from the outcome assessed by the second rater. There are various ways to achieve this test of homogeneity, ranging from Zelen's test of odds ratios[170], to a Wald's test following estimation of a statistical model[171] or Pitman's test of equal variances[172]. However, all of the aforementioned tests simply provide an assessment of the statistical significance and do not quantify the difference (if there is a difference).

2.5.2 Ratio of treatment effects

An alternative to a test of homogeneity, that does quantify the difference between treatment effect estimates, is to take the ratio of treatment effects (RTEs). The most common ratio of treatment effects is

the ratio of odds ratios (RORs), and was developed by Zacks and Solomon[173].

Table 17 Assessment of stroke by central adjudicators and site investigators

Site investigators	**Stroke**	**No stroke**
Intervention	a_1	b_1
Control	c_1	d_1
Central adjudicators		
Intervention	a_2	b_2
Control	c_2	d_2

Consider the generic example in Table 17. The ROR is given as:

$$ROR = \frac{OR_{site}}{OR_{adj}} = \frac{a_1 d_1}{b_1 c_1} \Big/ \frac{a_2 d_2}{b_2 c_2}$$

To determine the variance of the ROR, the natural logarithm of both variances are added. Thus:

$$SE[\ln(ROR)] = \sqrt{\frac{1}{a_1} + \frac{1}{b_1} + \frac{1}{c_1} + \frac{1}{d_1} + \frac{1}{a_2} + \frac{1}{b_2} + \frac{1}{c_2} + \frac{1}{d_2}}$$

Here, an $ROR > 1$ indicates that the effect estimates based on the central adjudicators assessment of stroke lead to more beneficial treatment effect estimates compared to those based on the site investigators assessment.

The ROR can be thought of as a subset of the RTE. The RTE extends the ROR to allow additional outcomes, or outcomes measured with a different approach (e.g. risk ratios), to be processed[150, 174]. The RTE follows a similar process to the ROR and the standard error can be calculated, enabling the variability around the RTE to be quantified

using a 95% confidence interval. This measure is superior to tests of homogeneity, and also permits use in a meta-analysis as RTEs can be calculated for each included study, provided that data is available for both site investigators and central adjudicators.

2.5.3 Summary of methods to establish whether treatment effect estimates differ for outcomes assessed by different raters

In summary, to determine whether treatment effect estimates differ for the same outcome assessed by site investigators and central adjudicators, the RTE is an appropriate statistic that quantifies how much effect estimates differ. The RTE would be the expected statistic for the comparison between treatment estimates between site investigators and central adjudicators rather than the ROR due to the fact that a large number of different outcomes, including ones not suitable to be analysed using odds ratios, are adjudicated in stroke trials. However, for additional analyses, such as a comparison of the number of events reported before and after adjudication or comparison between trials that report odds ratios only, the RTE will be the same as the ROR, and so the estimate will be presented as the ROR as this is more descriptive of the statistic.

The next chapter describes a systematic review and meta-analysis of randomised stroke trials. The meta-analysis uses the RTE as the statistic to synthesise included studies and provides evidence to answer the research question posed by this thesis.

Chapter 3 – Outcome assessment by central adjudicators versus site investigators in randomised stroke trials: A systematic review and meta-analysis

Publications contributing to this chapter:

1. PJ Godolphin, PM Bath, A Algra, E Berge, MM Brown, J Chalmers, L Duley, M Eliasziw, J Gregson, JP Greving, GJ Hankey, N Hosomi, SC Johnston, E Patsko, A Ranta, PM Sandset, J Serena, C Weimar, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in stroke trials: A systematic review and meta-analysis. Stroke 2019, 50:2187-2196. https://doi.org/10.1161/STROKEAHA.119.025019

2. PJ Godolphin, PM Bath, AA Montgomery. A comparison of central adjudication and site assessment of outcomes in randomised stroke trials: the Adjudicating Outcomes in Stroke Trials' collaboration (AOST). Poster presentation at 4th European Stroke Organisation Conference.

3. PJ Godolphin, PM Bath, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in randomised stroke trials. 2019. Poster presentation at 40th International Society for Clinical Biostatisticians Conference.

4. PJ Godolphin, PM Bath, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in randomised stroke trials: A systematic review and meta-analysis. 2019. Oral presentation at 5th International Clinical Trials Methodology Conference.

Abstract

Background: In randomised stroke trials, central adjudication of a trial's primary outcome is regularly implemented. This chapter compares outcomes assessed by central adjudicators with outcomes assessed by site investigators in stroke trials to evaluate whether trial results are altered following adjudication.

Methods: Randomised stroke trials where the primary outcome was assessed by both site investigators and central adjudicators were included in the review. Information about the adjudication process and the treatment effect for the primary outcome assessed both by central adjudicators and by site investigators was extracted from each trial. In each case, a ratio of these treatment effects (RTE) was calculated, with an RTE > 1 indicating that adjudication resulted in a more beneficial treatment effect. A random-effects meta-analysis model was fitted to estimate a pooled effect.

Results: Fifteen trials comprising 69,560 participants were included. The primary outcomes for these trials were stroke (8/15, 53%), a composite event including stroke (6/15, 40%) and functional outcome after stroke (1/15, 7%). The majority of site investigators were blind to treatment allocation (9/15, 60%). On average, there was no difference in treatment effect estimates based on data from central adjudicators and site investigators (pooled RTE=1.02, 95% C.I: [0.95, 1.09]).

Conclusions: There was no evidence to suggest that adjudication of the primary outcome in stroke trials had any impact on trial conclusions.

3.1 Introduction

The previous chapters have described the current literature, given a clear rationale for the investigation in this thesis and have critiqued and described the methods used in this research. This chapter is the first of the thesis that aims to answer the research question, namely: what are the benefits and costs of adjudicating outcomes in stroke trials?

This chapter describes a systematic review and meta-analysis that combine all the evidence on central adjudication of the primary outcome of a stroke trial, to establish whether this affects the primary result of each trial. This is the first step in the investigation into providing evidence to answer the research question, and will help to determine whether trials in which the primary outcome is not that adjudicated as likely to obtain different results. The ratio of treatment effects (RTEs) will be the statistic used to measure whether the trial results change.

The aims of this chapter are to (1) describe which type(s) of stroke trials have adjudicated outcomes, what outcomes are adjudicated and how adjudication has been conducted; (2) investigate the effects on the primary results of stroke trials when using outcomes assessed by central adjudicators compared with outcomes assessed by site investigators; (3) evaluate whether these effects differ by trial characteristics; and (4) establish whether adjudicating outcomes in stroke trials changes the number of events reported or the primary outcome.

3.2 Methods

The systematic review and meta-analysis were performed following the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) statement[175]. The PRISMA checklist is given in Appendix A.

3.2.1 Review protocol

The protocol for the review was developed with input from an experienced systematic reviewer and member of the Cochrane pregnancy and childbirth group, Lelia Duley. The protocol was registered on PROSPERO on 6th February 2017 shortly after the initial search had taken place. The version on PROSPERO was updated after the search was updated in November 2018 (see Section 3.2.3.2). This update was completed on 7th December 2018. Both versions of the protocol can be accessed on the PROSPERO website: www.crd.york.ac.uk/prospero/display_record.php?RecordID=56731.

3.2.2 Eligibility Criteria

The systematic review aims to identify all randomised stroke trials that have adjudicated their primary outcome. Therefore, articles were eligible for inclusion if they (1) described a randomised trial; (2) described a stroke trial, defined as one where the participants were either being treated for stroke, or being given an intervention to prevent stroke; and (3) the primary outcome of the trial had undergone assessment by site investigators and central adjudicators, with the trial providing data to calculate treatment effects for the primary outcomes based on central adjudication and site investigators separately.

The definition for a ‘stroke trial’ was chosen in consultation with clinical trialists and stroke researchers and aimed to give a clear definition from which to determine eligibility. This definition should encompass acute, prevention and rehabilitation stroke studies. There were no restrictions on the year of publication for studies to be eligible. Animal studies and/or articles not written in English were not eligible. These articles were recorded but excluded.

3.2.3 Defining the Search Strategy

In order to identify appropriate trials, a search strategy needed to be implemented, that was sensitive enough to capture all eligible studies, but was specific enough so as not to contain an inordinate number of unsuitable studies (e.g. animal studies, observational studies etc.). Medical Subject Headings (MeSH) terms were used to build the search strategy initially, with keywords used to complete the strategy. For the MeSH terms, focus and explode options were used together so that the database retrieved articles where each MeSH term or any associated terms were the main focus of the article. The search strategy for the review was developed with advice from a medical librarian based at the University of Nottingham.

The databases selected to be searched were Embase, MEDLINE, PsycINFO, the Cochrane Central Register of Controlled Trials (CENTRAL), Web of Science and Google Scholar. For Embase, MEDLINE and PsycINFO, Ovid (https://ovidsp.ovid.com/) was used to execute the search strategy, whilst for the other three databases the website-specific search engine was used. To find sufficient grey

literature, the first 300 articles from Google Scholar were screened, which is the amount recommended by Haddaway et al.[116].

An example search strategy for MEDLINE and Google Scholar is displayed below in Table 18. The strategies for EMBASE and PsycINFO follow a similar pattern to that of MEDLINE, whilst the strategies for CENTRAL and Web of Science follow a similar pattern to that of Google Scholar.

Table 18 Example search strategies

	Search text
MEDLINE	
1	exp randomised controlled trial/
2	exp animal/ not human/
3	exp stroke/
4	((adjudicat* or endpoint or outcome or review or classification or central or event)adj2(adjudicat* or committee or panel or review)).mp.
5	1 not 2
6	3 and 4 and 5
Google Scholar	
1	(“randomised” OR “randomized”) AND "stroke" AND ("adjudication" OR "outcome assessment" OR "event panel" OR "endpoint committee" OR "review committee" OR "central review” OR “event committee”)

exp refers to explode, / refers to focus, adj2 refers to words that are adjacent to each other, * refers to truncated words, with the search identifying anything after the * in addition to the truncated word

3.2.3.1 Scoping review

A scoping review was carried out to check the eligibility criteria, by assessing the number of eligible studies, and to estimate how specific the review was. To conduct the scoping review, the search strategy was implemented and a random selection of 100 titles and abstracts were inspected. The initial search (carried out on 17th November 2016) identified 4909 records. These comprised 1409 from EMBASE, 221 from MEDLINE, 138 from PsycINFO, 1668 from CENTRAL, 1173 from

Web of Science and 300 from Google Scholar (constrained at an upper limit of 300).

For nearly half of the records (47%) the full text was needed before eligibility could be determined (Table 19). This was predominantly due to the fact that many studies did not report adjudication in the title and abstract, and the full text was needed before it was clear whether the primary outcome was adjudicated centrally.

Following title and abstract screening, the 47 potentially eligible full texts were reviewed with 9 (19%) meeting eligibility criteria. However, from these nine, only one reported both the site and adjudicator assessment of the primary outcome. Thus, for the remaining eight records, it was necessary to obtain additional material (e.g. secondary analysis paper etc.), or to contact the authors directly to retrieve the missing data.

Table 19 Scoping review of a random sample of 100 records

	Title and abstract review (n=100)
Full text needed	47 (47%)
Not eligible	53 (53%)
If not eligible, reasons	
Not an RCT	34 (34%)
Not a stroke trial	19 (19%)
	Full text review (n=47)
Potentially eligible	9 (19%)
No full text available	3 (6%)
Not eligible	35 (74%)
If not eligible, reasons	
Not an RCT	7 (15%)
Not a stroke trial	11 (23%)
No evidence of adjudication	17 (36%)

All data are N(%). RCT refers to randomised controlled trial

Table 20 Estimation of workload

	Records as of 17/11/2016 (n=4909)
Records identified	4909 (100%)
After de-duplication	4400 (90%)
Full text needed	2068 (42%)
Potentially eligible trial	396 (8%)

All data are N(%)

Table 20 gives an estimation of the workload required in the main review, predicted from the scoping review. Assuming there were approximately 10% duplicate records, this estimation suggests that I would need to assess over 2000 full texts to determine eligibility of articles. This potentially exaggerates the number of unique records due to the fact that the search will contain duplicate records from the same trial (protocol paper, baseline characteristics, main results, secondary analyses etc.). However, even taking this potential overestimation into account, the quantity of full texts needed seems excessively high.

Thus, the decision was made to only include systematic reviews if they mentioned adjudication in the title or abstract (as well as mentioning randomised stroke trials), rather than get full texts for systematic reviews that simply describe a review of randomised stroke trials (as was done in the scoping review). This reduces the number of full texts considered, and looking through every systematic review that includes any stroke trial for one that was missed from the search is potentially unnecessary, as the wide reach of the search should identify at least one of the individual journal articles (protocol paper, main results etc.) from that trial. In addition, for records that are clearly secondary analyses of a main trial a similar process was applied, with the full text

only obtained if the title and abstract mentions adjudication (as well as meeting additional eligibility criteria).

3.2.3.2 Final Search Strategy

The final search strategy was unchanged following the scoping review and is displayed in Appendix B. The initial search was carried out from database inception until 17th November 2016 as in the scoping review. These searches were updated to include additional records from 18th November 2016 to 6th November 2018.

3.2.4 Study Outcomes

For each trial, only the primary outcome was included in this review. Secondary or safety outcomes were not included and, whilst interesting, are out of scope of this review. If the stroke trial had more than two trial arms then all comparisons were included, but for factorial trials only one comparison was selected. The correlation between comparisons that used the same control group was accounted for by calculating an adjusted weight for the trial based on this additional correlation. Continuous, binary and categorical (ordinal and nominal) outcomes were eligible, as were subjective and objective outcomes.

3.2.5 Selection of Studies

A two-stage screening process was used comprising initial screening of title and abstract, followed by full-text screening. Records were stored and managed in EndNote X8. In order to identify records, each record was assigned a four digit ID, RXXXX. During the screening process, when it became clear that multiple records described the same trial and

the records were not duplicates, then the trial was allocated a unique ID as well, in the format TXXX.

Before records were screened, duplicate references were identified, recorded and then discarded. A random sample of 100 titles and abstracts were screened in duplicate before screening progressed to the remainder of the records. As described earlier, if it was unclear from the title and abstract whether the record was eligible then the full text was sought. If the article clearly described a secondary analysis of a trial, then the full-text was only obtained if the record mentioned adjudication as well as satisfying eligibility criteria 1 and 2. Similarly, if the article clearly mentioned a systematic review, then the full-text was only retrieved if the record mentioned adjudication as well as satisfying eligibility criteria 1 and 2.

Full text reports were acquired for all records where potential eligibility was unclear. At the full text screening stage, full texts were grouped into trials where possible, as each trial could have multiple unique records, and these were assessed at a trial level for eligibility. In the full text screening, studies had to satisfy all eligibility criteria to be included in the review. A random sample of 50 full texts were first assessed by multiple reviewers, to check the process, before the remainder of the full texts were assessed.

3.2.5.1 Checking screening process 1 (titles and abstracts)

The check of the first stage of screening, titles and abstracts, involved 100 titles and abstracts assessed independently and in duplicate by

Alan Montgomery (AM) and Peter Godolphin (PG). Initially, 30 titles and abstracts were reviewed, with both reviewers agreeing on 24 (80%) and the remaining six agreed on after discussion. The next 30 titles and abstracts were reviewed with only three disagreements, and again these were agreed on after discussion. There were no disagreement on the final 40 records. After these checks, PG reviewed the remainder of the titles and abstracts independently.

3.2.5.2 Checking screening process 2 (full texts)

For the full text review, three reviewers assessed 50 full texts. The first 10 full texts were initially assessed by the first two reviewers (PG, AM), with no disagreements. The third reviewer, Emily Patsko (EP), used the first 10 full texts, and the decision log created for these, as a training tool. Once the third reviewer (EP) agreed with the first 10 full texts, two reviewers (PG, EP) assessed the remaining 40 in duplicate. There were five disagreements amongst the 40 full texts reviewed in duplicate, but all were agreed on following discussion. The remaining full texts were assessed independently by PG.

3.2.6 Data Extraction

Data extraction was carried out at a trial level, rather than a record level. Data from eligible studies were extracted independently by two reviewers (PG, EP) using a data extraction form. Disagreements were resolved by discussion between the reviewers. If agreement could not be reached then one further reviewer (AM) assessed the study, with the majority decision taken.

3.2.6.1 Designing the data extraction form

Three documents were used in the data extraction process:

- Data extraction form – all information to be extracted was listed here, along with record IDs (and evidence behind the decisions made, if applicable). The form is given in Appendix C.
- Data extraction manual – this gave information on how to complete the data extraction form. This sheet is given in Appendix D.
- Missing information log – this log detailed the missing data for every eligible trial. This log is given in Appendix E.

The data extraction form contained information on study eligibility, and if relevant reasons regarding why the study had been incorrectly listed as eligible at the full text screening stage. For each eligible trial, summary information was extracted including trial design, trial setting, intervention type and number of patients randomised. Details of the primary outcome were also extracted, which included the estimated treatment effect from both the adjudicator and site assessed outcome. Information about the characteristics of the adjudication process was collected from included studies. These characteristics included whether the adjudicators were blind to treatment allocation and blind to outcome assessment by site investigators, the number and profession of adjudicators, the information that was provided to the adjudicators, the process for adjudication including decision making and how disagreements between adjudicators were handled, and how cases

were selected for adjudication. Finally, the risk of bias assessment was carried out (see Section 3.2.7 for more details).

3.2.6.2 Double data extraction

Once data had been extracted by both reviewers (PG, EP), the forms were compared. Any disagreements were resolved after internal discussion. This resulted in a single data extraction form for each of the included studies.

3.2.6.3 Obtaining missing data

It was anticipated that outcomes assessed by site investigators would not be reported in trial publications. If this was indeed the case that essential information was not reported, PG emailed the contact author and requested unreported data[176]. See Chapter 4 for details on the methodology for requesting this missing data.

Any authors who subsequently provided information enabling their trial to be included in the systematic review were invited to join a collaboration, the Adjudicating Outcomes in Stroke Trials (AOST) collaboration. A data sharing and use agreement was obtained upon receipt of any data (see Appendix F for template agreement).

3.2.7 Assessment of risk of bias

For each included study, the Cochrane risk of bias tool[177] was used to assess study quality. Additionally, the risk of bias associated with the process of selecting cases for adjudication was assessed, with this termed as “adjudication risk of bias”. If central adjudicators only assess cases identified by site investigators then some bias may remain,

particularly if site investigators are not blind to treatment allocation and potentially have a biased view of the relative effectiveness of the treatments being compared. Four categories were created for adjudication risk of bias: (1) Only cases identified by site investigators not blind to treatment allocation were selected for adjudication (High risk of bias); (2) Only cases identified by site investigators blind to treatment allocation were selected for adjudication (Medium risk of bias); (3) Either all participants, or all suspected cases (e.g. using computer algorithm to identify possible cases) were selected for adjudication (Low risk of bias); (4) It is not clear how cases were selected for adjudication (Unclear risk of bias).

Initially, the adjudication risk of bias domain had only three categories (Low, High, Unclear). For this earlier version, Low included both Low and Medium in the new domain. The choice to increase the number of categories was made because it was felt that there was sufficient difference between the Low and Medium group to warrant separate reporting. This decision was made after data extraction and the new risk of bias information was only extracted by one reviewer (PG). To ensure consistency, the new category was checked alongside the old double extracted category.

3.2.8 Statistical Analysis

Continuous variables were summarised using mean and standard deviation, or median and interquartile range. Categorical variables were described using number (%). Mann-Whitney U tests, Chi-square tests

and Fisher exact tests were used to determine imbalance between included and potentially eligible but excluded studies.

To compare outcome assessment by central adjudicators and site investigators the RTE was calculated for each trial. The RTE was determined as the ratio of the treatment estimate using outcomes assessed by site investigators to the treatment estimate using outcomes assessed by central adjudicators. A RTE > 1 indicated that adjudication resulted in a more beneficial treatment effect. Data were pooled, if appropriate, in a meta-analysis using a DerSimonian and Laird random effects model.

The I^2-statistic was used to indicate the level of heterogeneity between studies. Subgroup analyses and meta-regression were used to investigate heterogeneity. Interaction between the RTE and the following were tested: (1) Risk of bias (high risk/medium risk/low risk/unclear risk); (2) Blinding status of site investigators (blinded/unblinded); (3) Intervention (drug/surgery/other); (4) Sample size of trial (continuous); (5) Number of sites (continuous); Funding source (Academic/commercial). A Sensitivity analysis was carried out using a fixed effect meta-analysis model.

To establish whether central adjudication led to a change in the number of events reported, irrespective of the RTE, the number of events reported by site investigators were compared to the number reported after central adjudication for each trial. An odds ratio (OR) > 1 indicates that central adjudication led to more events reported. For trials that

used the ordinal mRS, this was dichotomised so that a score of 3 and above indicated an event. As before, data were pooled in a meta-analysis using a random effects model. All analyses were performed in Stata version 15.1 or later.

3.3 Results

3.3.1 Search Results

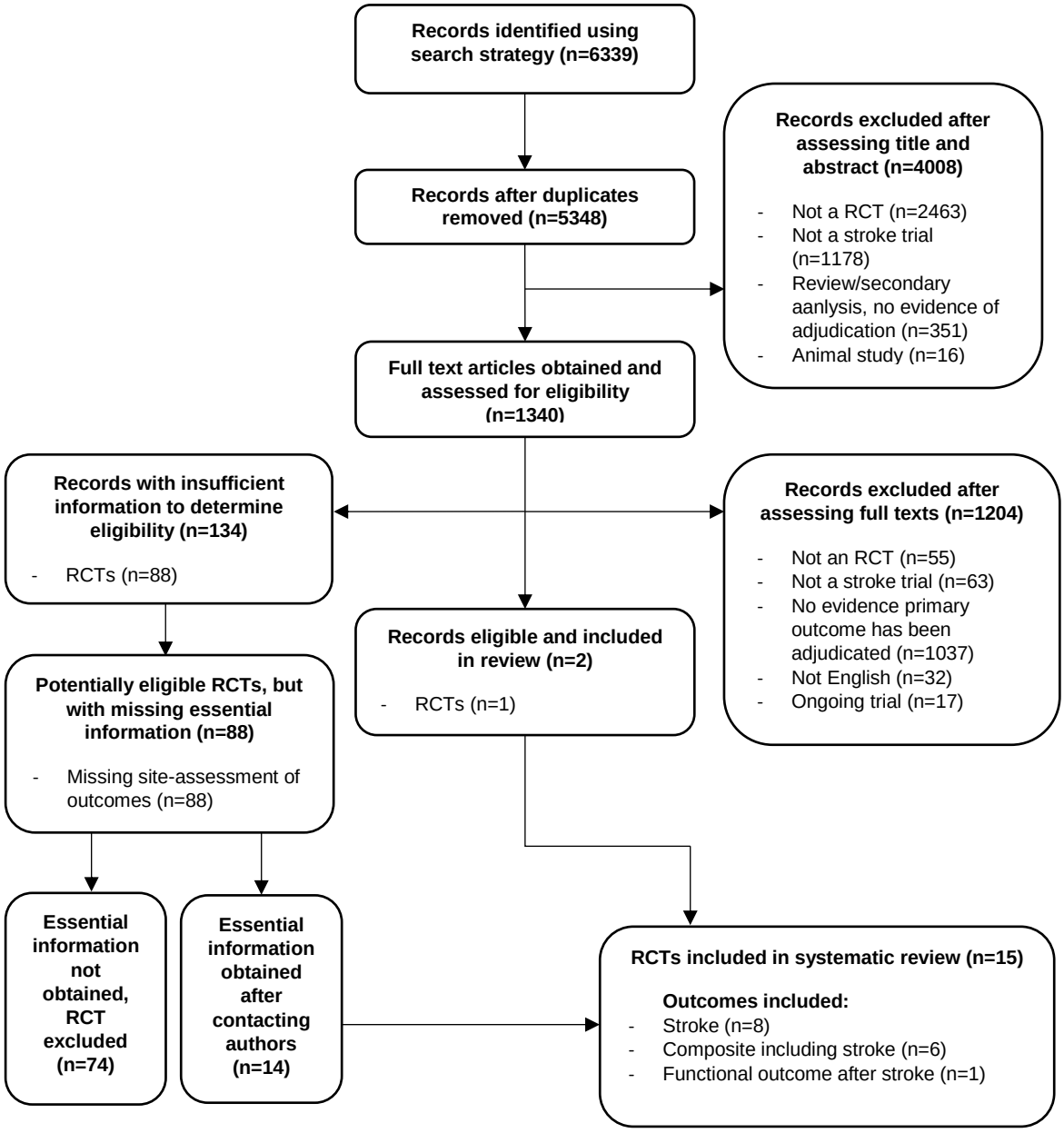
Database searches identified 6339 records (Figure 5). The initial search from database inception to 17th November 2016 found 4909 records (EMBASE – 1409, MEDLINE – 221, PsycINFO – 138, CENTRAL – 1668, Web of Science – 1173, Google Scholar – 300). The updated search from 17th November 2016 until 6th November 2018 yielded an additional 1430 records (EMBASE – 260, MEDLINE – 28, PsycINFO – 111, CENTRAL – 727, Web of Science – 304). Note Google Scholar was not searched again in the update on 6th November 2018 since, by the protocol, the 300 records identified in the initial search were used.

The 6339 records were assessed, and 991 duplicates were identified and removed. The remaining 5348 records had their titles and abstracts screened, with 4008 excluded at this stage. The reasons for exclusion were that: the record did not describe a randomised controlled trial (2463, 61%); the record did not describe a stroke trial as defined earlier (1178, 29%); the record described a systematic review or secondary analysis of a trial and did not mention adjudication (351, 9%); the record described an animal study (16, <1%) (See a flow diagram in Figure 5).

Screening of 1340 full texts indicated one trial (two records), PROGRESS[51, 182], that had reported sufficient information to be included. 1204 records were excluded after full text review, with 55 records not describing a randomised trial (5%), 63 records not

describing a stroke trial (5%), 1037 records providing no evidence that the primary outcome of the stroke trial was centrally adjudicated (86%), 32 records not being written in English (3%), and 17 records describing an ongoing trial (1%). An additional 88 trials (134 records) were potentially eligible, but did not report outcomes assessed by site investigators. Contacting the authors of these studies led to sufficient data from 14 trials. The remaining 74 potentially eligible studies were excluded. Therefore, 15 trials (69,560 participants) were eligible for inclusion[178-193].

Figure 5 Flow diagram



3.3.2 Characteristics of the potentially eligible trials

Table 21 shows the characteristics of the 15 included trials and the 74 potentially eligible studies that were excluded due to insufficient outcome data. The majority of included trials published their main results after 2010, were parallel group, secondary prevention, multicentre and compared two randomised groups. Common primary outcomes included in the review were occurrence of stroke or a composite outcome that included occurrence of stroke (14/15, 93%); site investigators were more likely to have assessed the primary outcome blind to treatment allocation (9/15, 60%). Studies that were potentially eligible but not included due to insufficient essential information were similar, but were from older publications, were more likely to be primary prevention trials, had recruited fewer participants, used fewer trial sites and were less likely to have occurrence of stroke as the primary outcome. A more detailed summary of the 15 included trials is given in Table 22.

3.3.3 Characteristics of adjudication in included trials

For all the included trials, central adjudicators were blind to treatment allocation (Table 23). In 5 trials, adjudicators were blind to outcome assessment by the site investigators. In the remaining 10 trials, either adjudicators were not blind to site investigators assessment or this was not reported. The number of adjudicators involved in each trial ranged from 2 to 28. In over half the studies multiple adjudicators assessed

each event. Disagreements between adjudicators were dealt with in a variety of ways, with additional blinded adjudication most common.

Table 21 Characteristics of included trials and potentially eligible trials

	Included (n=15)	Potentially eligible, but data not received (n=74)	p-value
Year of main trial publication			
1990-2000	2 (13%)	5 (7%)	0.13*
2001-2005	1 (7%)	9 (12%)	
2006-2010	2 (13%)	20 (27%)	
2011-2015	6 (40%)	35 (47%)	
2016-2017	4 (27%)	5 (7%)	
Study design			
Parallel	13 (87%)	70 (95%)	0.27*
Factorial	2 (13%)	4 (5%)	
Type of trial			
Primary prevention	3 (20%)	34 (46%)	0.16*
Secondary prevention	9 (60%)	30 (41%)	
Acute stroke	3 (20%)	10 (14%)	
Number of randomised groups			
2	12 (80%)	61 (82%)	0.86*
3	1 (7%)	7 (9%)	
≥4	2 (13%)	6 (8%)	
Patients randomised			
Mean (sd)	4637 (5764)	3717 (5246)	0.33†
Median [25th, 75th centile]	2885 [449, 6105]	1633 [439, 4576]	
Min, Max	129, 21105	48, 20702	
No. of sites			
Mean (sd)	216 (365)	185 (269)	0.59†
Median [25th, 75th centile]	82 [50, 172]	85 [27, 179]	
Min, Max	4, 1393	1, 1178	
Not reported	0 (-)	5 (7%)	
Intervention			
Drug	9 (60%)	52 (70%)	0.39*
Surgery/procedure	4 (27%)	18 (24%)	
Other	2 (13%)	4 (5%)	
Comparator			
Placebo	2 (13%)	14 (19%)	0.069*
Standard care	10 (67%)	41 (55%)	
Active treatment	0 (-)	14 (19%)	
Surgery/procedure	2 (13%)	5 (7%)	
Other	1 (7%)	0 (-)	
Primary outcome			
Occurrence of stroke	8 (53%)	20 (27%)	0.18*
Composite including stroke	6 (40%)	39 (53%)	
Functional outcome after stroke	1 (7%)	6 (8%)	
Other	0 (-)	9 (12%)	
Blinding status of site investigators			
Blind to primary outcome	9 (60%)	37 (50%)	0.48‡
Not blind to primary outcome	6 (40%)	37 (50%)	

Data are n(%) unless otherwise stated
*p-value from Fisher exact test

†p-value from Mann-Whitney U test
‡p-value from Chi-square test

Table 22 Summary of included trials

Trial name	Population	Intervention	Comparator	Primary outcome	Training of site-investigators	Funding source
CABACS	Patients with high-grade carotid stenosis (≥80%) who required coronary artery bypass grafting (n=129).	Carotid endarterectomy and coronary artery bypass grafting (n=65)	Coronary artery bypass grafting (n=64)	Stroke or death from any cause	Yes	Academic (German Research Council, DFG)
ENGAGE AF-TIMI 48	Patients with moderate-to-high-risk atrial fibrillation (n=21,105).	High-Dose Edoxaban (n=7035) or Low-Dose Edoxaban (n=7034)	Warfarin (n=7036)	Stroke or systolic embolism.	Not reported	Commercial (Daiichi Sankyo Pharma Development)
ESPRIT	Patients with TIA or mild stroke in the past 6 months (n=2739).	Aspirin and dipyridamole (n=1363)	Aspirin alone (n=1376)	Non-fatal stroke, non-fatal MI, major bleeding complication or death from vascular cause.	Yes	Academic (Council of Singapore; European Commission; Janivo Foundation; French Ministry of Health; The Netherlands Heart Foundation; Thrombosis Foundation; UK Stroke Association; University Medical Center Utrecht)
FASTEST	Patients with TIA or stroke (n=291).	Electronic decision support tool (n=172)	Usual care (n=119)	Recurrent stroke	Yes	Academic (Health Research Council of New Zealand)
HAEST	Patients with acute ischaemic stroke and atrial fibrillation (n=449).	Dalteparin (n=224)	Aspirin (n=225)	Recurrent ischaemic stroke	Yes	Academic (Research Council of Norway)

ICSS	Patients with symptomatic carotid stenosis (n=1713)	Stenting (n=855)	Carotid endarterectomy (n=858)	Fatal or disabling stroke	Not reported	Academic (Dutch Research Council; Medical Research Council; European Union; the Stroke Association)
J-STARS	Patients with a history of non-cardioembolic ischaemic stroke and cholesterol level between 4.65 and 6.21 mmol/L (n=1578).	Pravastatin (n=793)	No statins (n=785)	Stroke or TIA	Not reported	Academic (Ministry of Health, Labour and Welfare, Japan)
NASCET	Patients with non-disabling stroke and carotid stenosis of 30-99% in the internal carotid artery. There were three populations mild (<50%, n=1368), moderate (50-69%, n=858) and severe (70-99%, n=659) stenosis.	Carotid endarterectomy. In addition, patients received medical care, including antiplatelet therapy. Mild (n=678), moderate (n=430), severe (n=328).	Medical care, including antiplatelet therapy. Mild (n=690), moderate (n=428), severe (n=331)	Fatal or non-fatal ipsilateral stroke	Yes	Academic (National Institute of Neurological Disorders and Stroke)
POINT	Patients with minor ischaemic stroke or high-risk TIA (n=4881).	Clopidogrel and aspirin (n=2432)	Aspirin alone (n=2449)	Ischaemic stroke, MI or death from ischaemic vascular event	Yes	Academic (National Institute of Neurological Disorders and Stroke)
PROGRESS	Patients with a history of stroke or TIA (n=6105)	Perindopril with the addition of a diuretic at the discretion of treating physician (n=3051)	Matching placebo (n=3054)	Fatal or non-fatal stroke	Yes	Commercial (Servier)
REVASCAT	Patients with acute ischaemic stroke who could be treated within 8 hours (n=206)	Medical therapy (including alteplase if eligible) and thrombectomy (n=103)	Medical therapy (including alteplase if eligible) (n=103)	Functional outcome at 90 days (mRS)	Yes	Commercial (Coviden)
SOCRATES	Patients with nonsevere ischaemic stroke or high-risk TIA (n=13199).	Ticagrelor (n=6589)	Aspirin (n=6610)	Non-fatal stroke, non-fatal MI or death.	Yes	Commercial (AstraZeneca)

SPS3	Patients with MRI-defined symptomatic lacunar infarctions (n=3020).	Blood pressure target <130 mmHg (n=1501)	Blood pressure target 130-149 mmHg (n=1519)	Stroke	Yes	Academic (US National Institutes of Health-National Institute of Neurological Disorders and Stroke)
TARDIS	Patients with acute ischaemic stroke or TIA (n=3096).	Aspirin, clopidogrel and dipyridamole (n=1556)	Aspirin and dipyridamole or clopidogrel alone (n=1540)	Recurrent stroke and TIA	Yes	Academic (British Heart Foundation; National Institute of Health Research)
VITATOPS	Patients with stroke or TIA within the past 7 months (n=8164).	B vitamins (n=4089)	Matching placebo (n=4075)	Not-fatal stroke, non-fatal MI or death from vascular cause.	Not reported	Academic (Australia National Health; Medical Research Council; Singapore National Medical Research Council; Australia National Heart Foundation)

TIA refers to Transient Ischaemic Attack, MI refers to myocardial infarction, mmol/L refers to millimoles per Litre, mRS refers to modified Rankin Scale, mmHg refers to millimetres of mercury

Table 23 Characteristics of adjudication in included trials

	Total (n=15)
Adjudicators blind to treatment allocation	
Yes	15 (100%)
Adjudicators blind to assessment of the primary outcome made by site investigators	
Yes	5 (33%)
No	4 (27%)
Not reported	6 (40%)
Number of adjudicators	
2-4	9 (60%)
5-9	2 (13%)
≥10	4 (27%)
Adjudicators profession*	
Neurologist	14 (93%)
Cardiologist	3 (20%)
Other health professional	5 (33%)
Not a health professional	2 (13%)
Not reported	1 (7%)
Information provided to adjudicators*	
Medical notes	10 (67%)
Original case report forms	11 (73%)
Cranial scans	7 (47%)
Audio recording	1 (7%)
Video footage	1 (7%)
Not reported	3 (20%)
Adjudication method	
Each case was assessed by one adjudicator	5 (33%)
Each case was assessed by >1 adjudicator	9 (60%)
Not reported	1 (7%)
How disagreements were dealt with	
Each event only assessed by single adjudicator, so no disagreements	4 (27%)
Further adjudication	5 (33%)
Consensus decision between adjudicators/committee	4 (27%)
Not reported	2 (13%)
Method used to select cases to adjudicate	
Only cases identified by unblinded site investigators are adjudicated	5 (33%)
Only cases identified by blinded site investigators are adjudicated	5 (33%)
All participants, or all suspected cases identified by blinded site investigators are adjudicated	5 (33%)

All data are N(%)
*Categories are not mutually exclusive

3.3.4 Risk of bias

Adjudicators commonly assessed only cases reported by site investigators. For the ten trials where this occurred, half the site investigators were not blind to treatment allocation. In total, five trials were assessed as low adjudication risk of bias, five as medium adjudication risk of bias and five as high adjudication risk of bias (Table 24). One trial, NASCET, had CT scans and patients' medical histories adjudicated that had been identified by unblinded site investigators, but in NASCET, all participants were adjudicated making this low adjudication risk of bias. REVASCAT also had all participants adjudicated. For the additional three trials with low adjudication of bias, ENGAGE AF TIMI-48, POINT and SOCRATES, each had blinded site assessors, and suspected cases were identified for adjudication.

All studies were high quality according to the Cochrane risk of bias tool, with the majority of the domains rated as low risk of bias. In fact, twelve trials were rated as low risk of bias for all the domains in the Cochrane risk of bias tool, with ICSS, REVASCAT and TARDIS each having one domain each that was high risk of bias.

Table 24 Risk of bias of included trials

Trial name	Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Incomplete outcome data	Selective reporting	Adjudication
CABACS	Sequence was generated as stratified blocks with randomly varying block size (Low risk)	Participants were allocated in a concealed way by web-based central preoperative randomisation before surgery (Low risk)	No blinding since this was an open trial but the outcome (stroke or death) was not likely to be influenced by lack of blinding (Low risk)	All primary outcome events were adjudicated by an independent blinded end point committee (Low risk)	Primary outcome data available for all participants in primary analysis (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed events reported by unblinded site investigators (High risk).
ENGAGE AF-TIMI 48	Sequence was randomly generated with stratification for important prognostic factors (Low risk)	Participants were allocated via central, 24-hour, interactive, computerised response system (Low risk)	The subjects, investigators and staff involved in the treatment were unaware of treatment allocation (Low risk)	Staff involved in clinical evaluation were unaware of treatment allocation. In addition, an independnet clinical events committee adjudcated all efficacy endpoints in a blinded manner (Low risk)	Primary endpoint ascertained for 99.5% of the total 56,346 patient-years of potential follow-up (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed suspected events reported by blinded site investigators as well as suspected events identified through reviewing source information and adverse event terms (Low risk).
ESPRIT	Sequence was randomly generated using computer software and stratified by hospital (Low risk)	Participants were allocated centrally by the Central Trial Coordination Centre (Low risk)	No blinding since this was an open study. However the outcome (stroke, MI, bleeding, vascular death) is unlikely to be influenced by the lack of blinding (Low risk)	Outcome cranial scans were audited by physicians who were blinded to treatment allocation (Low risk)	Primary outcome data available for 99.1% of randomised participants (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed events reported by unblinded site investigators (High risk).
FASTEST	Sequence was randomly	General practices were allocated in	Blinding was not possible, but the	All primary outcome events were	Primary outcome data available for	Protocol available and	Adjudicators reviewed events reported by

	generated using computer software (Low risk)	clusters centrally (Low risk)	outcome (stroke) was not likely to be influenced by a lack of blinding (Low risk)	adjudicated by a blinded neurologist (Low risk)	all participants who met inclusion (Low risk)	all outcomes reported in publications (Low risk)	unblinded site investigators (High risk).
HAEST	Sequence was randomly generated using computer software (SAS, version 6.10) and was blocked with block size four (Low risk)	Participants were allocated to sequential numbered packages containing either active drug and corresponding placebo (Low risk)	Participants and personnel were blinded through use of matching placebos (Low risk)	All primary outcome events were evaluated and classified by a blinded independent endpoint committee (Low risk)	Primary outcome data available for all participants in primary analysis (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed all events believed by blinded investigators to represent an endpoint (primary or secondary endpoint) (Medium risk).
ICSS	Sequence was randomly generated using computer software and stratified for a number of factors (Low risk)	Participants were allocated centrally and allocations were obtained by telephone or fax by staff not involved in other parts of the trial (Low risk)	Blinding was not possible, and the outcome (fatal or disabling stroke classified as mRS>3) could be influenced by a lack of blinding (High risk)	All primary outcome events were adjudicated by an independent endpoint committee that was unaware of treatment allocation (Low risk)	Primary outcome data available for all participants in primary analysis (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed vascular events and functional status reported by unblinded site investigators (High risk).
J-STARS	Sequence was randomly generated using computer software (Low risk)	Participants were allocated centrally via a web-based registration system (Low risk)	Blinding was not possible, but the outcome (stroke or TIA) was unlikely to be influenced by a lack of blinding (Low risk)	All primary outcome events were adjudicated by the central event evaluation committee blind to treatment allocation (Low risk)	Primary outcome data available for all participants in the intention to treat analysis (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed events reported by unblinded site investigators (High risk).
NASCET	Sequence was randomly generated using computer software and was stratified by trial centre (Low risk)	Participants were allocated centrally by the Data Management Centre (Low risk)	Blinding was not possible, but the outcome (stroke) is unlikely to be influenced by treatment allocation (Low risk)	All primary outcome events were adjudicated by blinded external adjudicators (Low risk)	Complete primary outcome data available for 99.7% of participants (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed CT scans and details of participant's cerebrovascular history for all participants (Low risk).

POINT	Sequence was randomly generated using computer software with stratification according to trial site (Low risk)	Participants were allocated centrally using interactive web-based system (Low risk)	Participants and personnel were blinded through use of matching placebo (Low risk)	All primary outcome events were adjudicated by an independent clinical-event committee unaware of group assignments (Low risk)	Primary analysis undertaken on all randomised participants. 98% of participants were followed up for at least 7 days (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed all suspected endpoints including those that did not meet strict definitions, identified by blinded site investigators (Low risk).
PROGRESS	Sequence was randomly generated using computer software (Low risk)	Participants were allocated centrally using a randomisation service accessed by telephone or facsimile (Low risk)	Participants and personnel were blinded through use of matching placebos (Low risk)	An endpoint adjudication committee unaware of treatment allocation reviewed all primary outcome events (Low risk)	Complete primary outcome data available for 99.9% of participants (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed events reported by blinded site investigators (Medium risk).
REVASCAT	Sequence was randomly generated using computer software using minimisation for age, stroke severity, therapeutic window and occlusion site (Low risk)	Participants were allocated centrally suing a web-based randomisation process (Low risk)	Participants and personnel were not blinded, as this was an open study. The outcome which was severity of disability could have been influenced by knowledge of randomisation (High risk)	All participants were adjudicated in a blinded manner (Low risk)	Complete primary outcome data was available for all participants (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	All participants who reached the primary endpoint were adjudicated (Low risk).
SOCRATES	Sequence was randomly generated using computer software and generated in blocks (Low risk)	Participants were allocated using an interactive telephone and web-based system (Low risk)	Participants and personnel were blinded through use of matching placebos (Low risk	All primary outcome events were adjudicated by an independent clinical-event adjudication committee who were unaware of treatment assignment (Low risk)	All participants included in primary analysis, with 99.2% of participants in the study at the end-of-treatment visit (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed all suspected endpoints including those that did not meet strict definitions, identified by blinded site investigators (Low risk).

SPS3	Sequence was randomly generated using computer software with a permuted block design with variable block size (Low risk)	Participants were allocated centrally by study coordinators using a web-based system (Low risk)	Participants and personnel were aware of treatment allocation (management of blood pressure) but the outcome (stroke) was unlikely to have been influenced by knowledge of this (Low risk)	All primary outcome events were adjudicated by a central adjudication committee that was unaware of treatment allocation (Low risk)	Primary outcome data available for all participants in primary analysis (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed events reported by blinded site investigators (Medium risk).
TARDIS	Sequence was randomly generated using computer software with minimisation for prognostic factors which included a random element for 5% of participants (Low risk)	Participants were allocated centrally using a web-based system (Low risk)	Blinding was not possible as this was an open trial. The outcome (incidence and severity of stroke) could have been influenced by knowledge of treatment allocation (High risk)	All primary outcome events were validated and categorised by expert adjudicators who were masked to treatment assignment (Low risk)	Primary outcome data was available for 99.2% of participants (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed events reported by blinded site investigators (Medium risk).
VITATOPS	Sequence was generated using random permuted blocks stratified by hospital (Low risk)	Participants were allocated using a central 24-hour telephone service or telephone website (Low risk)	Participants and personnel were blinded through use of matching placebo (Low risk)	All primary outcome events were audited by a masked adjudication committee (Low risk)	7462 (91%) participants were followed up until the trial ended. Primary analysis based on all participants (Low risk)	Protocol available and all outcomes reported in publications (Low risk)	Adjudicators reviewed events reported by blinded site investigators (Medium risk).

mRS refers to modified Rankin Scale, TIA refers to Transient Ischaemic Attack

3.3.5 Impact of adjudication on primary treatment effect

The results of the primary analysis for all included trials using both outcome assessment by site investigators and central adjudicators are displayed in Figure 6. For three of the comparisons (CABACS, FASTEST, NASCET:mild) the point estimates and confidence intervals were unchanged using both the centrally adjudicated outcome and site assessed outcome. For all remaining comparisons, the results using either outcome were similar and had consistent confidence intervals.

Thus, the RTE for each comparison in turn contained the null value, one, in its 95% confidence interval (Figure 7). The meta-analysis of these RTEs showed no evidence that adjudication altered estimates of treatment effect size when compared to assessment by site investigators (pooled RTE=1.02, 95% C.I: [0.95, 1.09]). There was no evidence of any interaction between the impact of adjudication on treatment effect estimates and blinding of site investigators to the trial allocation, number of participants randomised, number of trial sites, type of intervention, adjudication risk of bias or funding source (Figures 8-13). The sensitivity analysis was consistent with the random effects pooled meta-analysis result (Figure 14).

3.3.6 Impact of central adjudication on the number of events reported

There was evidence that central adjudication reduced the number of reported events (pooled OR=0.91, 95% C.I: [0.88, 0.95], Figure 14) when compared with the number of events reported by site investigators.

Figure 6 Analysis of 18 comparisons from 15 included trials, comparing the effect size for the primary outcome based on whether assessment was by central adjudicators or not

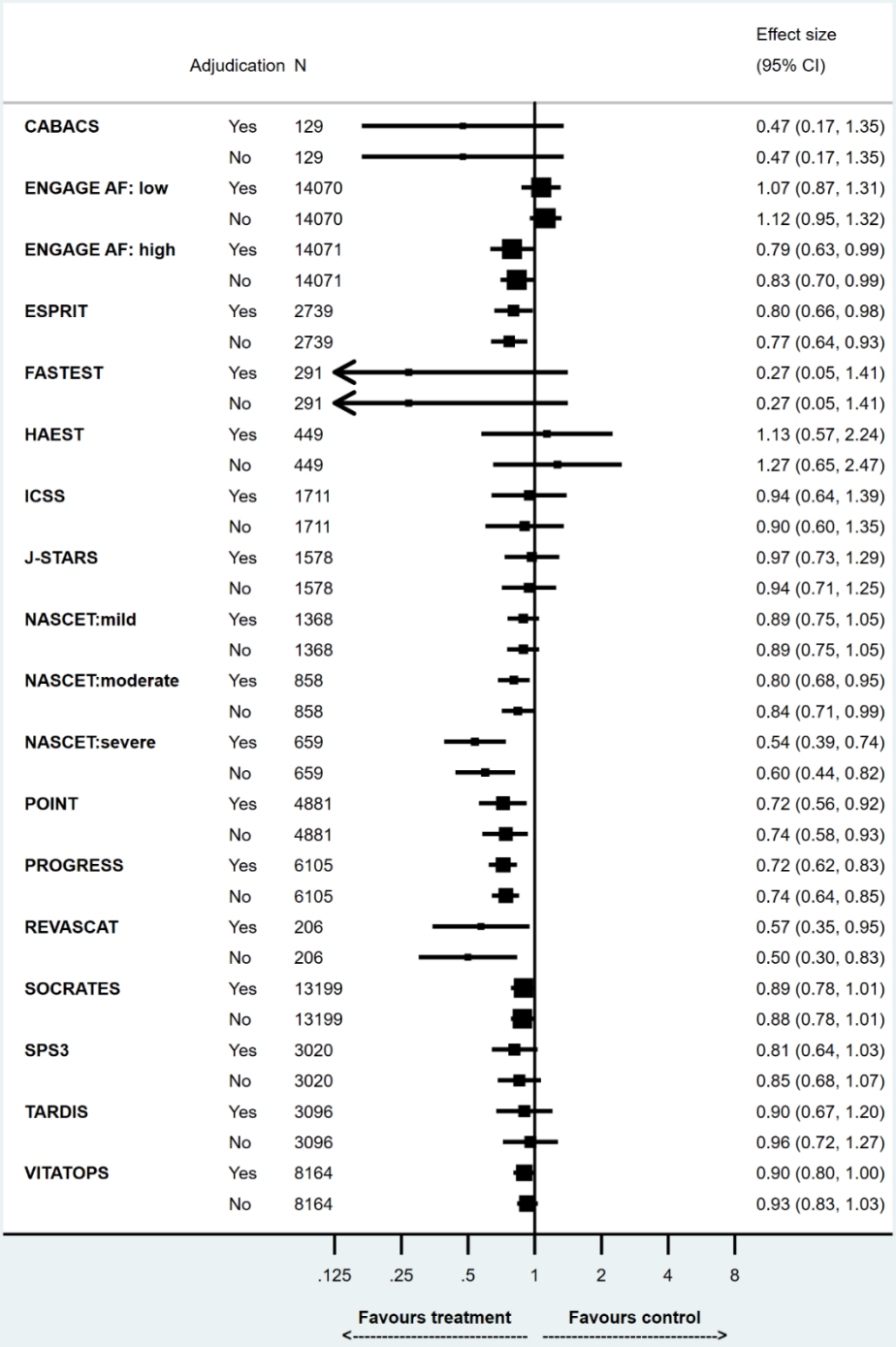
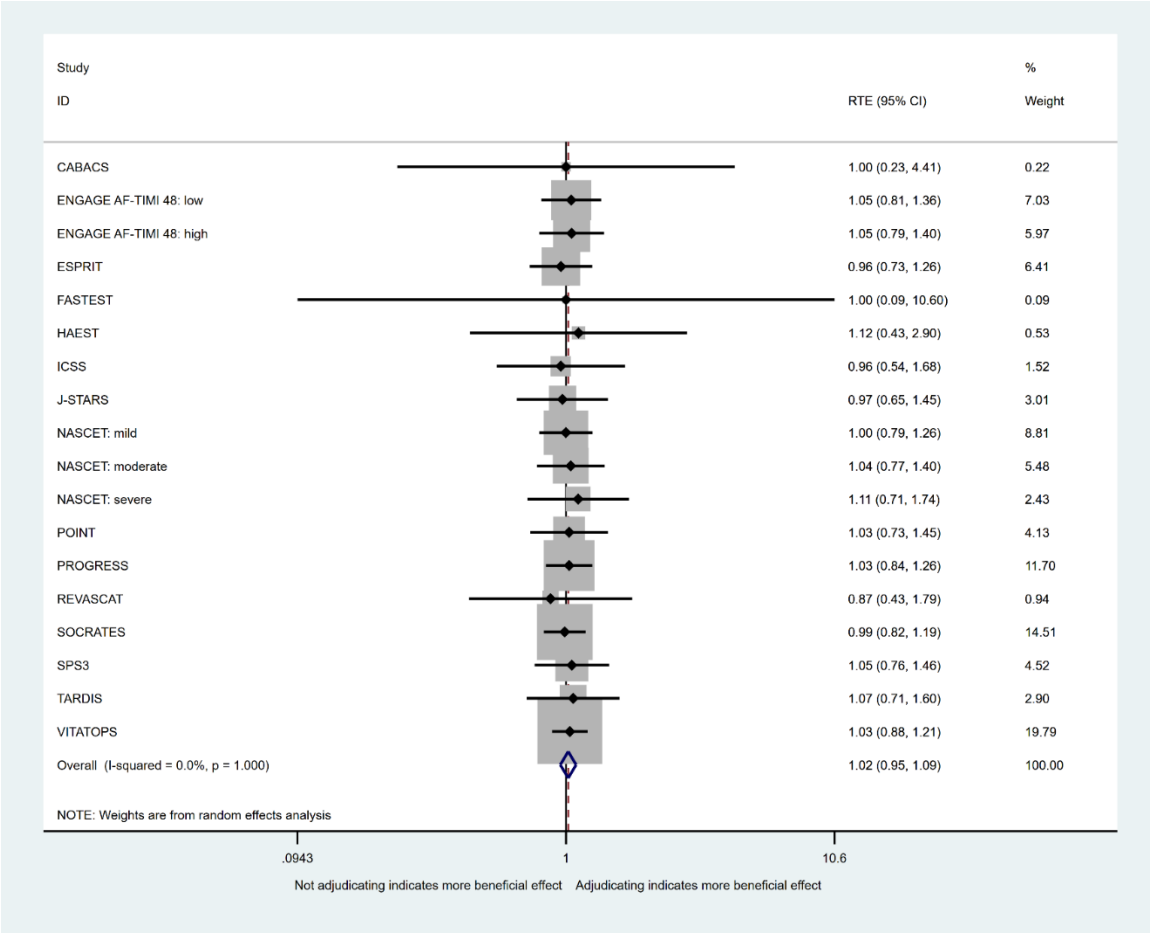
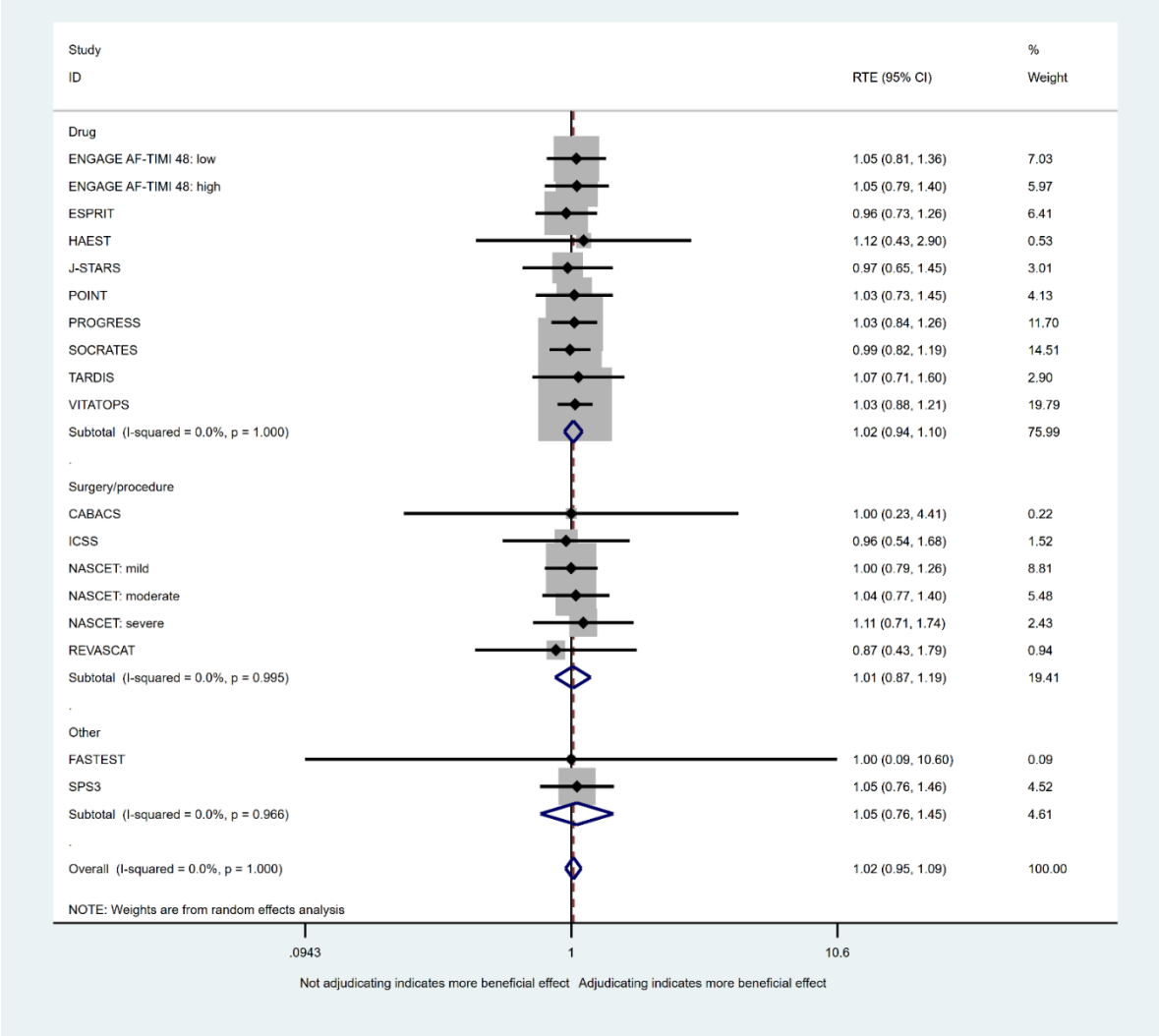


Figure 7 Meta-analysis of RTE in included studies, using a random effects model



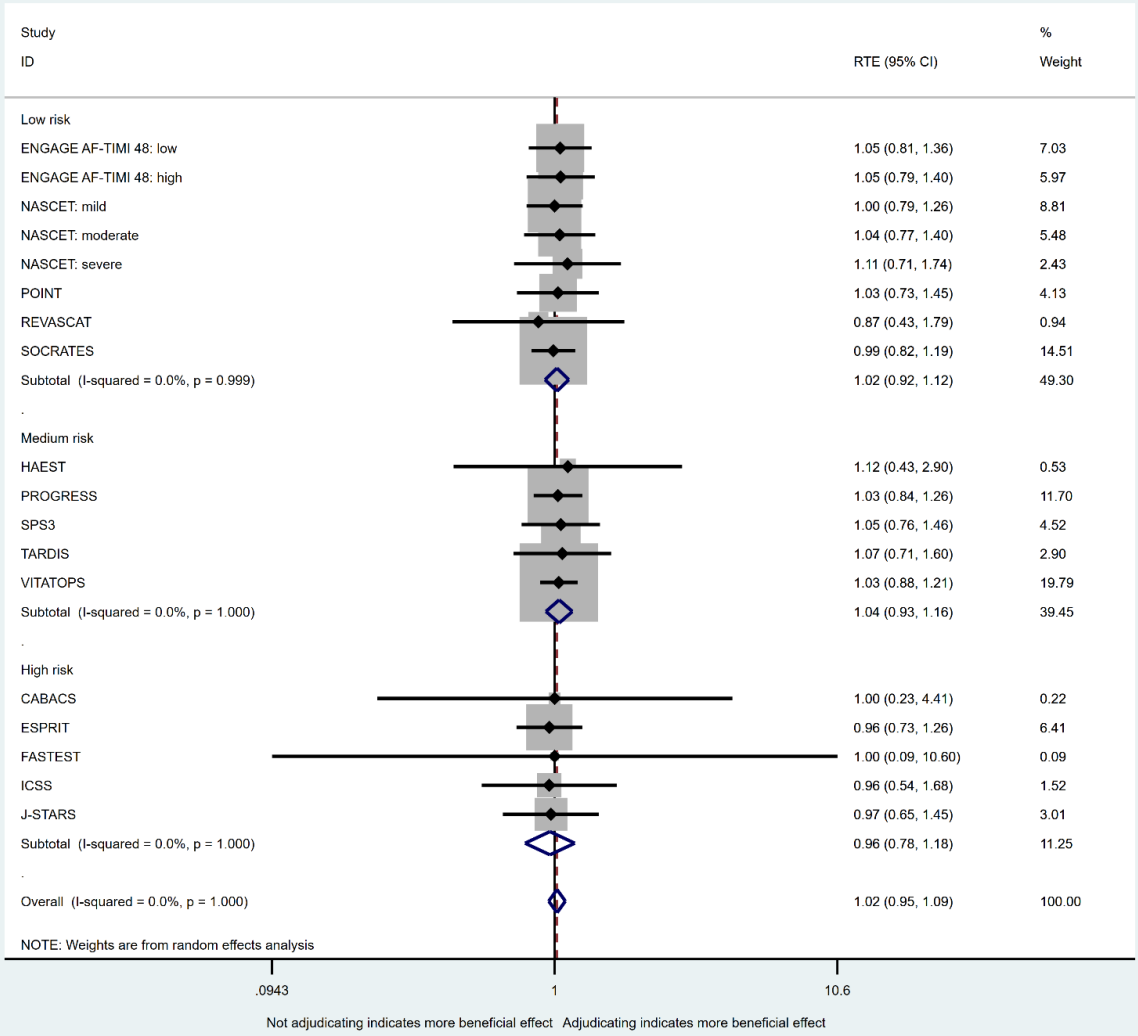
RTE refers to ratio of treatment effects

Figure 8 Meta-analysis of RTE by intervention type, using a random effects model



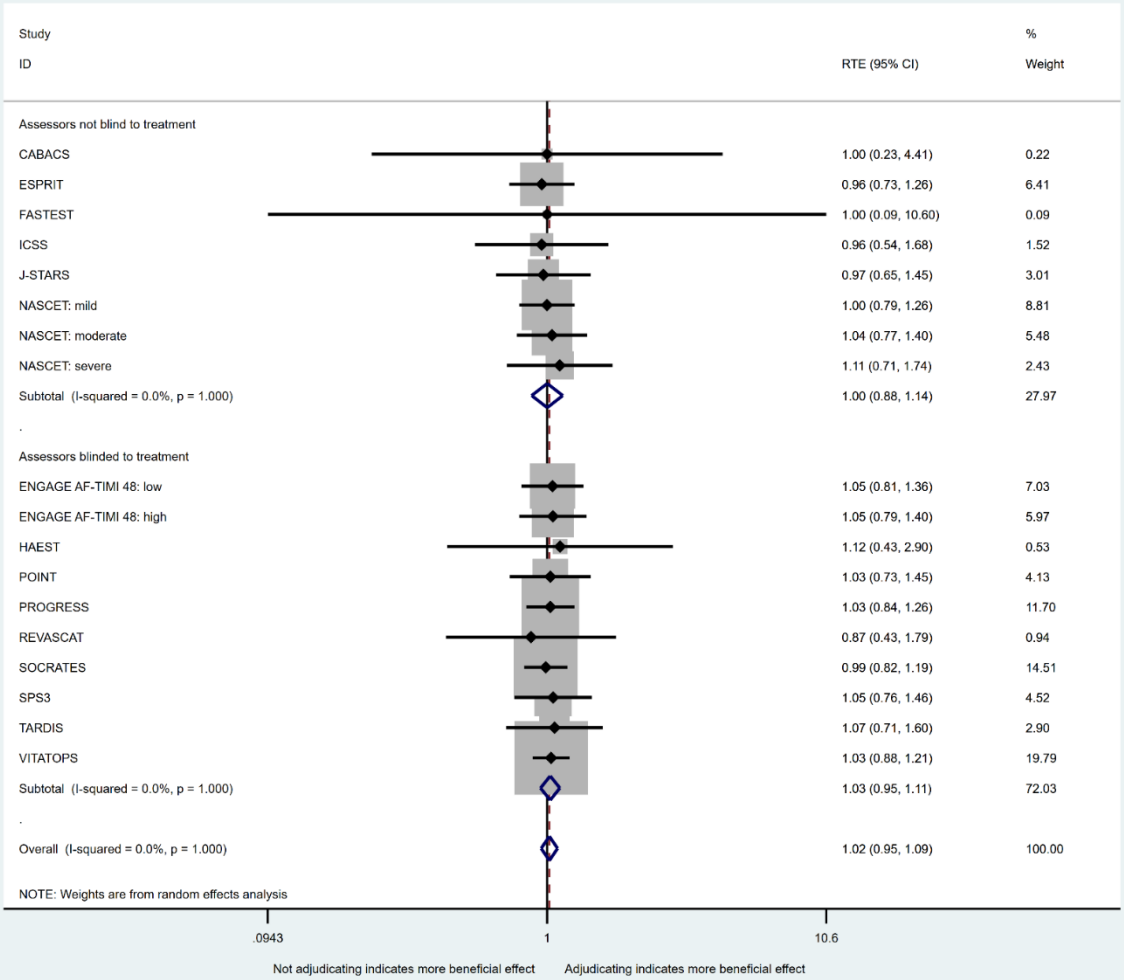
RTE refers to ratio of treatment effects

Figure 9 Meta-analysis of RTE by adjudication risk of bias, using a random effects model



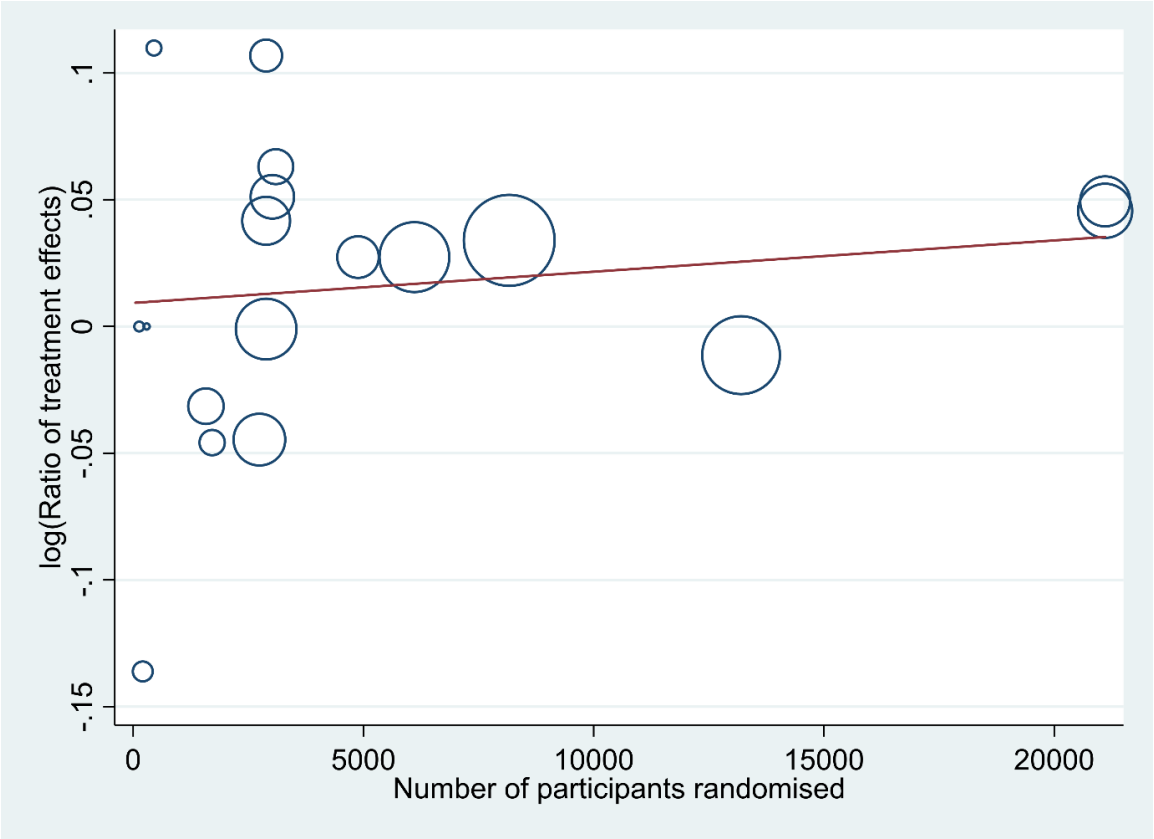
RTE refers to ratio of treatment effects

Figure 10 Meta-analysis of RTE by blinding status of site investigators, using a random effects model



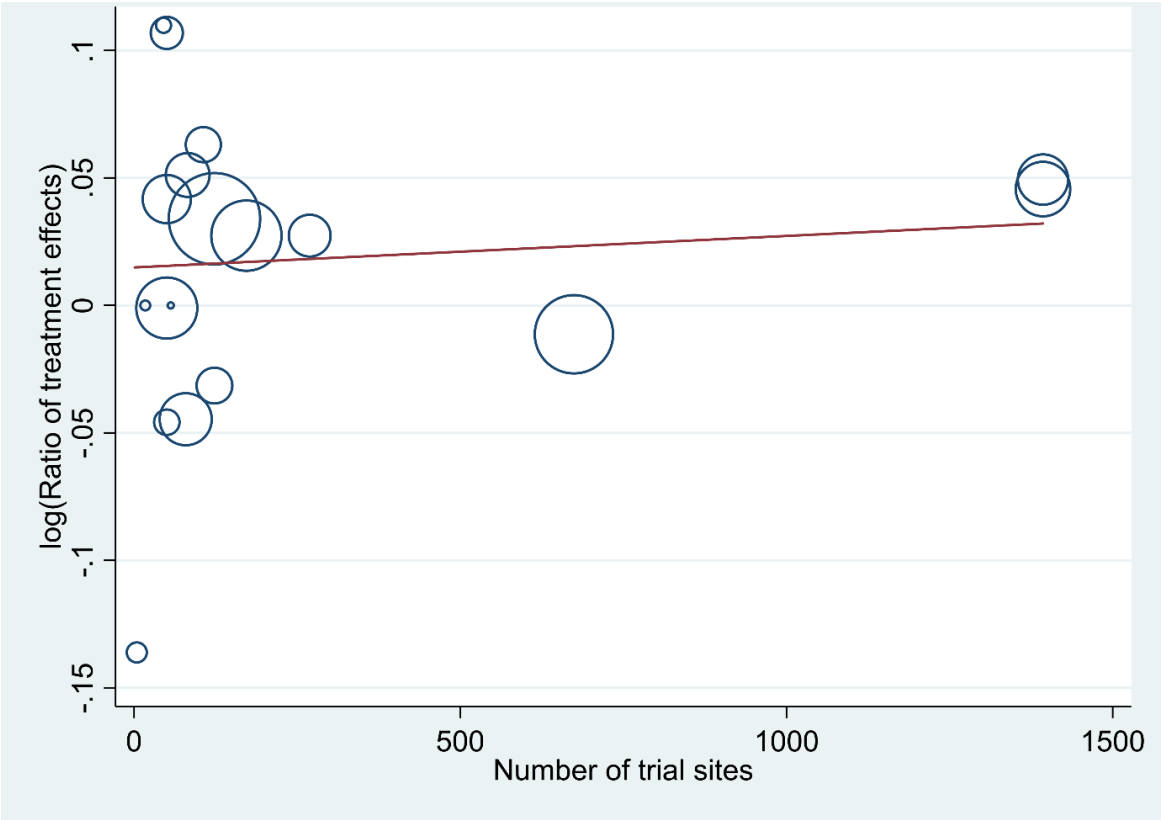
RTE refers to ratio of treatment effects

Figure 11 Meta-regression with number of participants randomised fitted as a covariate



Number of comparisons included = 18, estimate for covariate = 1.00 (95% C.I:[1.00, 1.00], p=0.83)
Size of circles correspond to weighting in meta-analysis

Figure 12 Meta-regression with number of trial sites fitted as a covariate



Number of comparisons included = 18, estimate for covariate = 1.00 (95% C.I:[1.00, 1.00], p=0.88)
Size of circles correspond to weighting in meta-analysis

Figure 13 Meta-analysis of RTE by funding source, using a random effects model

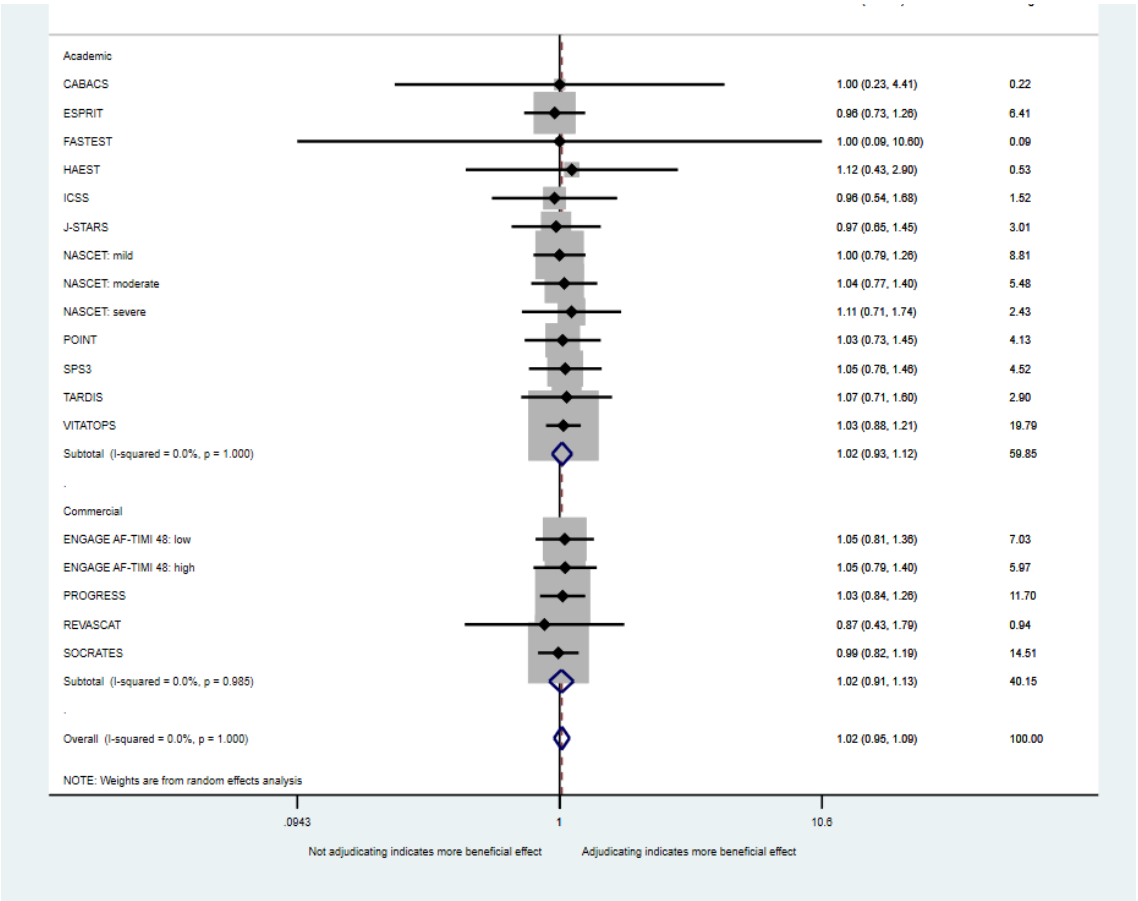
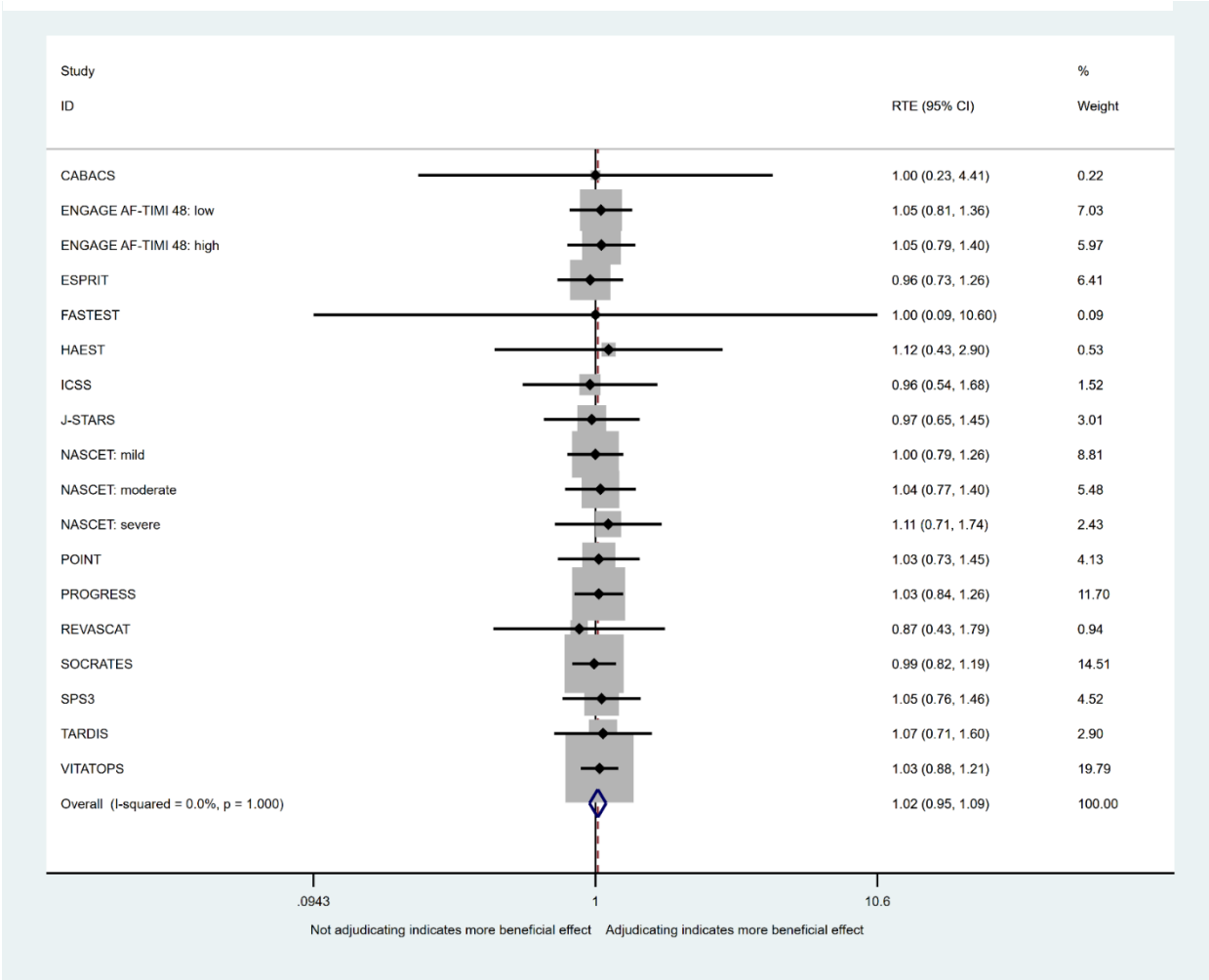
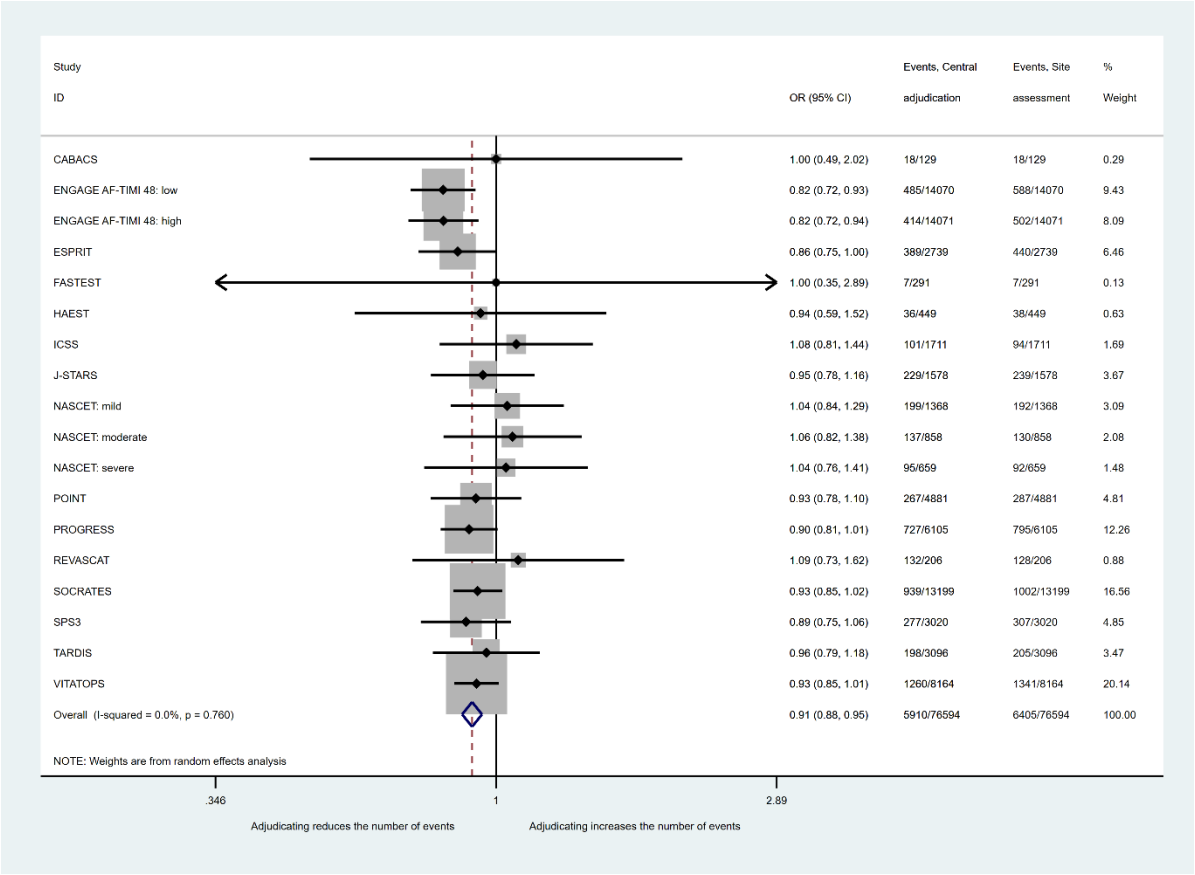


Figure 14 Meta-analysis of RTE, using a fixed effect model



RTE refers to ratio of treatment effects

Figure 15 Meta-analysis of change in number of reported events in included studies, using a random effects model



OR refers to odds ratio

3.4 Discussion

This chapter investigated whether central adjudication of the primary outcome in 15 stroke trials altered the primary trial efficacy result. In this meta-analysis of 15 stroke trials, including nearly 70,000 patients, there was no evidence to indicate that central adjudication of the primary outcome had any substantive impact on the primary trial result. Exploration of a number of pre-specified subgroups, including trial size, intervention type and blinding status of site investigators supported this finding. However, there was evidence that indicated that adjudication resulted in fewer events reported for the primary outcome.

This finding is consistent with two previous reviews. One included ten cardiovascular trials[44] and found no effect of adjudication versus site reported events (ratio of odds ratios, ROR=1.00, 95% C.I: [0.97, 1.02]). Another was a Cochrane Review[60] of subjective binary outcomes, across all clinical areas, that found adjudication in 47 trials to have no impact on treatment effect estimates (ROR=1.00, 95% C.I: [0.97, 1.04]). This review appears to bring further evidence that adjudicating primary outcomes in trials does not have a substantive effect on study conclusions. The Cochrane Review did suggest, however, that adjudication might be important when site investigators are not blinded to treatment allocation. This could be due to knowledge of the allocation influencing decisions about the primary outcome[30, 33, 34], but no evidence was found in this review. This could be because the primary outcomes in the 15 trials included (and thus commonly adjudicated in stroke trials) are predominantly objective. Therefore, if the primary

motivation of adjudication is to ensure validity of trial results, the process may be redundant for trials with objective outcomes.

In addition, it was found that site investigators over-reported the number of events for the primary outcome, which is consistent with a review of ten cardiovascular trials[44] that concluded that adjudication reduced the number of reported events. However, this reduction of events had no effect on the primary analysis in either this review or the previous review, indicating that site investigators over-report events in a similar proportion in both treatment arms. Whilst this non-differential misclassification by site investigators will not affect the primary trial analysis, it could affect the comparison of the rate of the primary outcome relative to other events (e.g. bleeding with a new anticoagulant). This could influence the risk-benefit calculation of bleeding versus prevention of the primary event, which is important for both clinicians and regulators. However, it is important to note that the process of adjudication in many trials does not enable central adjudication to add events as adjudicators only assess events reported by site investigators. Thus, the extent that sites over-report events may not actually be as high as found in this thesis. Additionally, there may be other reasons to adjudicate in randomised trials, which can have benefits that are more difficult to quantify. Central adjudication could identify poorly performing sites or even act as a policing effect that strengthens local assessment as investigators could perform more thorough assessment of outcomes that are to be adjudicated.

One limitation of the current review is that whilst a total of 89 trials were identified as being potentially eligible data was only obtained from 15 of these, even after members of the trial team were sent a number of reminders[176]. A large proportion of studies, that responses were received from, could not provide data since the unadjudicated data were not available, and it is possible that this review may have found different results had more trials provided data. However, the characteristics of the excluded studies were similar to those that were included, and the individual results from all 15 included trials agreed with the overall pooled estimate. Another limitation of the small sample size is that the included studies are high quality trials in respect to the usual components of risk of bias, and higher quality studies could have less need for adjudication than lower quality trials. Furthermore, the review largely included prevention stroke trials and studies with binary primary outcomes. A larger sample of trials with greater variation in quality, type and outcome could potentially have different findings. In the review protocol, it was stated that the impact of adjudication on the RTE would be investigated by the subjectivity of the outcome, but this analysis was not undertaken. This was because all the included outcomes are common in stroke trials and are sufficiently objective to change clinical practice.

In summary, no evidence was found that central adjudication of the primary outcome in stroke trials had any impact on estimated treatment effect size. However, central adjudication did control non-differential misclassification and limit over-reporting of events by site investigators.

If the primary purpose of central adjudication is to ensure validity of trial conclusions, then these results suggest that the potential advantages of central adjudication may not outweigh cost and time disadvantages in stroke studies.

The following chapter describes the process used to contact authors to obtain missing site-assessment data for use in this review. This includes a methodological randomised trial to evaluate method of author contact. Chapter 5 explores the findings of the current chapter in more detail, introducing differential and non-differential misclassification to investigate how much random and/or systematic error would be required before central adjudication would alter the treatment effect estimate for the primary outcome in stroke trials.

Chapter 4 – A nested randomised trial investigating short email with attachment versus long email without attachment when contacting authors to request unpublished data for a systematic review

Publications contributing to this chapter:

1. PJ Godolphin, PM Bath, AA Montgomery. Short email with attachment versus long email without attachment when contacting authors to request unpublished data for a systematic review: a nested randomised trial. BMJ Open 2019, 9:e025273. http://dx.doi.org/10.1136/bmjopen-2018-025273

2. PJ Godolphin, PM Bath, AA Montgomery. Randomised trial of different methods of author contact in order to receive data for a systematic review of randomised stroke trials. 2018. Poster presentation at 4th European Stroke Organisation Conference.

3. PJ Godolphin, PM Bath, AA Montgomery. Different methods of author contact in order to receive unpublished data for a Systematic Review. 2018. Oral presentation at Link ’18 Conference.

Abstract

Background: Systematic reviews often rely on the acquisition of unpublished analyses or data. A nested randomised trial was undertaken comparing two different approaches for contacting authors to request additional data for a systematic review.

Methods: A systematic review of central adjudication in stroke trials identified trials that needed to provide additional data before their inclusion. Research active authors were selected as contacts for each trial and were randomised to be sent either a short email with a protocol of the systematic review attached (“Short”), or a longer email that contained detailed information and without the protocol attached (“Long”). The unit of analysis was trial, accounting for clustering by author. The primary outcome was whether a response was received from authors, with time to response collected as a secondary outcome.

Results: 88 trials with 76 primary authors were identified in the systematic review, and of these, 36 authors were randomised to Short [trials=45], and 40 to Long [trials=43]. Responses were received for 69 trials. There was no evidence of a difference in response rate between trial arms (Short vs Long, odds ratio 1.10, 95% C.I: [0.36, 3.33]). There was no evidence of a difference in time to response between trial arms (Short vs Long, hazard ratio 0.91, 95% C.I: [0.55, 1.51]).

Conclusions: There was no evidence to suggest that email format had an impact on the number of responses received when acquiring data for a systematic review, or the time taken to receive these responses.

4.1 Introduction

The previous chapter described a systematic review and meta-analysis evaluating central adjudication of the primary outcome in randomised stroke trials. The review identified 88 potentially eligible studies that did not report the site-assessment of the primary outcome. 14 of these potentially eligible studies provided this missing data, and combined with an additional one that had already reported the data needed, resulted in 15 trials that were included in the systematic review and meta-analysis.

This chapter describes the process used to contact the authors of these 88 potentially eligible studies. Due to a lack of evidence about the best way to contact authors, as detailed in Chapter 1, authors were randomised to two different methods of email format in a methodological trial. The two methods of email format were a short email with the study protocol attached or a long email with no attachment. These methods were chosen to represent two common approaches to contacting academics/clinicians.

The aims of this chapter are to: (1) describe a nested randomised trial, which investigated two different approaches to contacting authors for data in a systematic review, with the primary aim of seeing which method elicits the most complete response; and (2) investigate the effect of allocated group on time to response, first response outcome and whether authors agree to eventually collaborate.

4.2 Methods

4.2.1 Rationale

In Chapter 3, a systematic review and meta-analysis was described that investigated whether central adjudication of the primary outcome in stroke trials had any impact on the main trial primary analysis[194]. As mentioned in Section 3.3.1, only one trial was identified that reported all the essential information that was needed to include that trial in the review. However, there were 88 potentially eligible trials which indicated that the primary outcome had undergone assessment by both site-assessors and central adjudicators, but did not report the primary outcome that was determined by the site assessors. Thus, without this information it was not possible to include these trials in the review.

In order to acquire this data, some member of the team for each trial needed to be contacted. As described in Chapter 1, the corresponding author would appear to be the most appropriate person to contact when attempting to acquire unpublished clinical trial data, but contacting these authors may be challenging due to the fact that they are often the senior authors and thus often already have a large number of restrictions on their time[76]. Therefore, if email contact is used, this must be clear, concise and easy to respond to. Some authors may prefer a shorter email, with additional information attached, whilst others may prefer a longer email with sufficient information they need to make a decision without reading additional attached information such as a study protocol.

The overall goal of contacting these authors is to receive appropriate data so that the studies can be included in the systematic review. Given the lack of evidence about the best way to contact authors, this systematic review would be a good opportunity to test out the two hypotheses mentioned above. However, testing whether initial author contact impacts on how much data received may not be a sensible outcome to test. This is due to the fact that you would anticipate that it would take a number of subsequent interactions before data would be sent, and the likelihood of receiving data could then be based largely on these subsequent interactions rather than the initial email. Therefore, as described in Chapter 1, an important outcome to assess is whether any reply is received from the initial contact, before considering whether the intervention affects if authors agree to collaborate and ultimately whether they are willing to share their clinical trial data with the systematic review team.

4.2.2 CONSORT

This methodological randomised trial was performed following the CONSORT (CONsolidated Standards of Reporting Trials) guideline[195]. The completed checklist is given in Appendix G.

4.2.3 Participants

The participants in this trial were authors of the potentially eligible studies for the systematic review described in Chapter 3. In brief, the studies included were those adjudged to be randomised trials of prevention or treatment of stroke, that had centrally adjudicated their primary outcome. MEDLINE, Embase, the Cochrane Central Register

of Controlled Trials (CENTRAL), Web of Science, PsycINFO and Google Scholar were searched for relevant articles. I restricted searches from Google Scholar to the first 300 articles[116], and only articles written in English were considered.

Two author contacts were chosen from each identified trial, with these being the participants selected for this study. The primary contact was ideally one of the corresponding authors. The chosen contact was checked to determine whether they were still an active researcher. To establish whether they were still research active, I carried out a search on PubMed to see when they had last published a journal article. If they had published a paper within the last 2 years, as lead or co-author, they were deemed research active. If the corresponding author appeared to have ceased research activity (no paper published in the last 2 years), then a secondary author who was still an active researcher (preferably first/second/last) was selected as the primary point of contact. To decide which author to choose, I read the author contribution statement and identified the author that appeared to have had the largest role in the trial. Additionally, a second research active author was chosen as a second contact for the trial. Ideally, this second contact also had a major role in the trial, determined by their position on the main trial publication and in the author contribution section of the manuscript, if completed.

Up to date email addresses were obtained from both contacts. This was achieved through searching the authors most recent publications for contact details, and confirming that these details were correct

(position/institution if email not available) on their institutions web page. For example, if the email address identified had contact details that indicated that the author was based at Nottingham Clinical Trials Unit, I went to Nottingham Clinical Trials Unit's web page to ensure that the author did currently hold that position there. All authors were identified and email addresses were obtained for each contact in turn before randomisation.

Upon emailing authors, I invited them to collaborate in the research programme described in this thesis, investigating adjudication in stroke trials. This was called the Adjudicating Outcomes in Stroke Trials (AOST) collaboration. In order to collaborate, they were asked to provide data in one of two formats: individual participant data or summary results (more details in Chapter 3). Authors that provided data were given the opportunity to critically review the findings of review and draft the manuscript. Those that accepted this offer were invited to join the systematic review writing committee.

4.2.4 Trial Design

This study was a parallel group, randomised trial of short email and attachment versus long email without attachement to elicit response from potential collaborators. The trial was nested within a systematic review and meta-analysis that was described in Chapter 3. It was not possible to blind participants or outcome assessors, as it was clear to contacted authors and the trial team which format of email had been sent. However, contacted authors were unaware that alternative formats of email were being compared in a nested randomised study

and therefore it was anticipated that their willingness to respond would be unbiased. In addition, the analysis plan was agreed before randomisation occurred, and was followed. Thus, it was not expected that the analysis methods carried out and decisions made in relation to the analysis would be affected by any predispositions the trial team had around the effectiveness of either intervention.

4.2.5 Randomisation

Randomisation occurred at an author level rather than at a trial level. This was to ensure that if any author was the contact for multiple trials that were to be included, then they only received one email, rather than multiple emails, which could be in different formats. Randomisation was stratified by year of publication (median split: <2011/≥2011), size of trial (median split: <1738/≥1738) and if the author had multiple trials which were included (Yes/No). Therefore, there were eight strata from which to randomly allocate participants. If authors had multiple trials to be included, then the within-author median size and median year were used when stratifying. Allocation was concealed and a researcher not involved in the trial randomised participants using a computer-generated random number list.

Participants were randomised in a 1:1 ratio to either of the two treatment groups: (1) Short, where they received a short email message that fits on one screen without the need to scroll, with the protocol attached (Intervention group); or (2) Long, where they received a longer email message that contained a fuller description of the systematic review and request for information, with the protocol

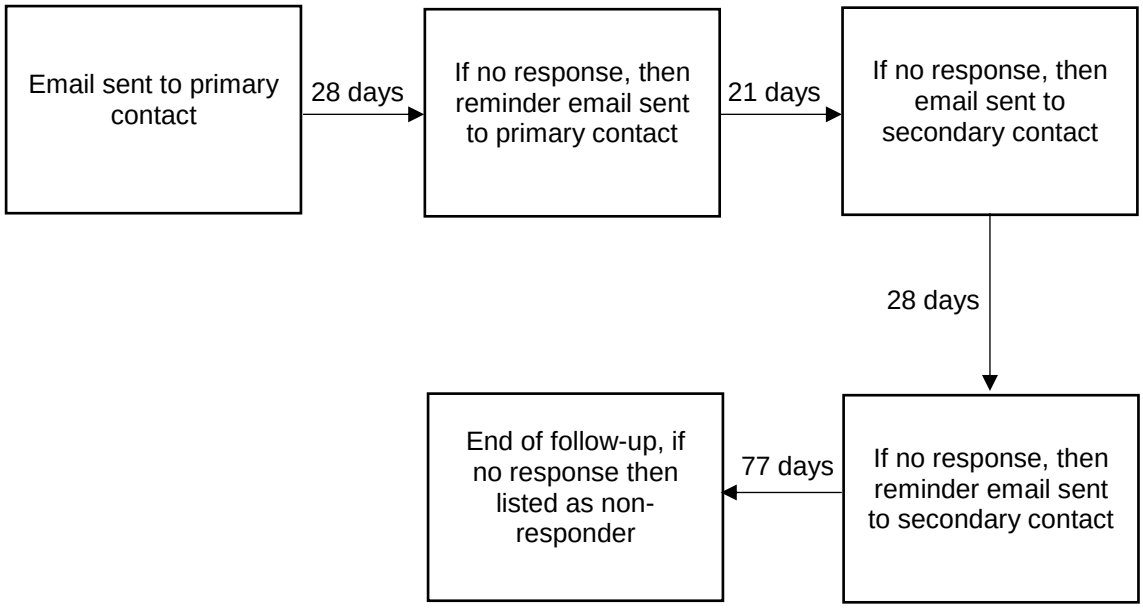
available on request (Comparator group). The protocol detailed the purpose of the systematic review, the data that was requested and the planned programme of research (see Appendices H & I for template emails and Appendix J for AOST protocol).

4.2.6 Plan for eliciting response

Each author received a maximum of two emails. The primary contact was first sent an email in the format of the randomly allocated intervention. If no response was received, then a second email was sent 28 days later, which included some additional brief reminder text, which was the same for both groups, added at the top of the original email (achieved using Reply to all in Microsoft Outlook).

If there was still no response from the primary contact author after a further 21 days, the second contact author was sent an email in an identical format, waiting 28 days before a reminder was sent. The second contact was not emailed if the first contact responded within the seven-week timeframe. Therefore, there was a maximum of four emails sent per trial. Participants were followed up for a further eleven weeks after the fourth and final email. After this stage, any authors that had not replied were recorded as non-respondents, regardless of contact thereafter (Figure 16). All primary contacts were emailed on the same day, with reminders sent (if needed) at the same time. Similarly, any secondary contacts that needed to be emailed, either initially or sent a reminder, had these done on the same day as every other secondary contact in this trial.

Figure 16 Plan to elicit response



4.2.7 Outcome Measures

The primary outcome for this nested randomised trial was whether a response was received from either of the two contacted authors in the trial timeframe. This was a binary outcome (Response received, Yes/No). Out of office automatic replies were not counted as a response, and no additional emails were sent upon receipt of an out of office reply.

Secondary outcomes were time to response, measured in days, and the number of emails needed to respond. Upon receipt of an initial response, this was categorised as either being Negative, Neutral or Positive. Positive responses were allocated if there was an interest in the collaboration and research questions, above simply asking for further information. Negative responses were assigned if participants gave the indication that they did not want to be involved/did not have time to take part. All other responses were coded as Neutral.

The eventual outcome upon the end of the trial timeframe was allocated as either (1) agree to collaborate, (2) do not agree to collaborate, (3) no decision reached or (4) no response. If there was an agreement in principle regardless of whether data had been received, then this was assigned as agree. If there was a clear indication that the author could not/did not want to help then this was allocated do not agree. No decision reached was given if it was still unclear whether the author would be willing to take part in the collaboration.

4.2.8 Statistical Analysis

The sample size for this nested randomised trial was fixed by the number of studies identified in the systematic review, and although allocation was by author, the unit of analysis is the individual trial. A sample size of 88 trials could detect a difference in response rate between groups of ≤30 percentage points (equivalent odds ratio 3.4) with 80% power and 5% two-sided alpha, ignoring any reduction in effective sample size due to the clustering effect of authors with multiple studies. The characteristics of each trial were summarised using appropriate descriptive statistics for each intervention arm. The primary approach to between-group comparative analyses was by intention-to-treat. Short was treated as the intervention group and Long treated as the comparator group for all between group analyses. The evaluation of response rate was performed using adjusted logistic regression models. The adjusted model included stratification factors; year of trial publication, trial sample size and multiple trials per author. The primary efficacy parameter comparing interventions was the odds ratio along

with the corresponding 95% confidence interval. Robust standard errors were used to account for correlation between multiple trials for the same author. A sensitivity analysis was carried out by further adjusting the primary analysis model for any characteristics of trials with marked imbalance between intervention groups.

Time to response was investigated using survival analysis, with response being the outcome of interest. Adjusted hazard ratios, adjusted for stratification factors, were reported alongside 95% confidence intervals. Additional secondary outcomes were summarised using mean, standard deviation, median, lower and upper quartiles, or frequency counts and percentages where appropriate. All analyses were performed in Stata version 15.0 or later.

4.3 Results

4.3.1 Characteristics of included studies

The study commenced on 11th July 2017. 88 trials were identified from a systematic search of the literature as potentially eligible studies to include in the systematic review and meta-analysis. This corresponded to 76 unique authors who were selected as the primary contact for these studies. 36 authors were randomised to Short (45 trials), and 40 to Long (43 trials) email formats. The majority of the trials had their main results paper published between 2006 and 2015, were either primary or secondary prevention stroke trials, were carried out in over 50 trial centres and had a median number of patients randomised above 1000 (Table 25). Trials randomised to each group had similar study design, intervention type and comparator, but studies allocated to Long tended to have more participants included and a larger number of trial centres. The number of intervention groups was similar for studies allocated to either Short or Long.

The majority of authors chosen as the primary contact were the first author of the main publication, or listed first in the writing committee if the article was published under a collaborative group. Conversely, the last author was the most common choice for the second contact, although there was greater variation in choice for this contact, with a position other than first, second or last also prevalent. For studies allocated to Long, the primary contact was more likely to be the

corresponding author compared to trials allocated to Short, although author position was similar for both groups (Table 26).

Table 25 Characteristics of included trials

	Short message and protocol (n=45)	Long message (n=43)
Number of authors randomised	36	40
Trials per author*		
1	30 (83%)	37 (93%)
2	4 (11%)	3 (8%)
3	1 (3%)	0 (0%)
4	1 (3%)	0 (0%)
Year of main trial publication		
1990-2000	5 (11%)	2 (5%)
2001-2005	7 (16%)	3 (7%)
2006-2010	11 (24%)	11 (26%)
2011-2015	19 (42%)	22 (51%)
2016-2017	3 (7%)	5 (12%)
Patients randomised		
Mean (sd)	3910.8 [5593.4]	3755.7 [5154.8]
Median [25th, 75th centile]	1224 [439, 5170]	1809 [500, 4576]
Min, Max	74, 20332	48, 21105
Type of trial		
Primary prevention	20 (44%)	16 (37%)
Secondary prevention	19 (42%)	21 (49%)
Acute stroke	6 (13%)	6 (14%)
Setting		
1 continent	30 (67%)	17 (40%)
>1 continent	14 (31%)	23 (53%)
Not found	1 (2%)	3 (7%)
No. of centres		
Mean (sd)	212.8 [286.7]	165.5 [293.4]
Median [25th, 75th centile]	67 [27, 260]	85 [32, 141]
Min, Max	1, 1034	4, 1393
Study design		
Parallel†	44 (98%)	41 (95%)
Factorial	1 (2%)	2 (5%)
Intervention		
Drug	30 (67%)	31 (72%)
Surgery/procedure	12 (27%)	11 (26%)
Device	2 (4%)	1 (2%)
Diet	1 (2%)	0 (0%)
Comparator		
Placebo	6 (13%)	10 (23%)
Standard care	28 (62%)	24 (56%)
Active treatment	8 (18%)	6 (14%)
Surgery/procedure	3 (7%)	3 (7%)
Number of intervention groups		
2	39 (87%)	38 (88%)
3	4 (9%)	3 (7%)
4	1 (2%)	1 (2%)
>4	1 (2%)	1 (2%)

Data are N(%) unless otherwise indicated.

*Unit of allocation was carried out at author level, therefore number of trials per author do not sum to total N

†Two parallel trials were cluster randomised (Short = 1; Long = 1)

Table 26 Author position in main results paper for primary and secondary contacts of included trials

	Short message and protocol (n=45)	Long message (n=43)
Primary contact		
First author	32 (71%)	30 (70%)
Second author	2 (4%)	3 (7%)
Last author	4 (9%)	5 (12%)
Other	7 (16%)	5 (12%)
Corresponding author	32 (71%)	37 (86%)
Secondary contact		
First author	5 (11%)	7 (16%)
Second author	8 (18%)	9 (21%)
Last author	18 (40%)	17 (40%)
Other	14 (31%)	10 (23%)
Corresponding author	4 (9%)	3 (7%)

Data are N(%).

4.3.2 Impact of email format on response rate

In total, responses were received for 69 trials within the trial timeframe, with authors of 19 studies (22%) not responding before the trial finished. The response rates for both trial arms were similar, and fitting a logistic regression model found no evidence of a difference in response rate between trial arms (Table 27). Adjusting the logistic regression model further for whether the primary contact was the corresponding author (due to the imbalance witnessed earlier) in a sensitivity analysis, supported the primary finding with a similar adjusted odds ratio and confidence interval (Table 28).

Table 27 Primary outcome: response from contacted authors

	Response	Adjusted Odds Ratio	95% C.I	p-value
Short message and protocol	36/45 (80%)	1.10	(0.36, 3.33)	0.87
Long message	33/43 (77%)			

Adjusted for year of publication, size of trial and if the author had multiple trials included. Total N included = 88. Robust standard errors were used in model fitting.

Table 28 Sensitivity analysis. Logistic regression model for primary outcome adjusted for variables with baseline imbalance

	Response	Adjusted Odds Ratio	95% C.I	p-value
Short message and protocol	36/45 (80%)	0.99	(0.33, 3.03)	0.99
Long message	33/43 (77%)			

Adjusted for year of publication, size of trial, if the author had multiple trials included and whether the primary contact was the corresponding author. Total N included = 88. Robust standard errors were used in model fitting.

4.3.3 Impact of email format on response time

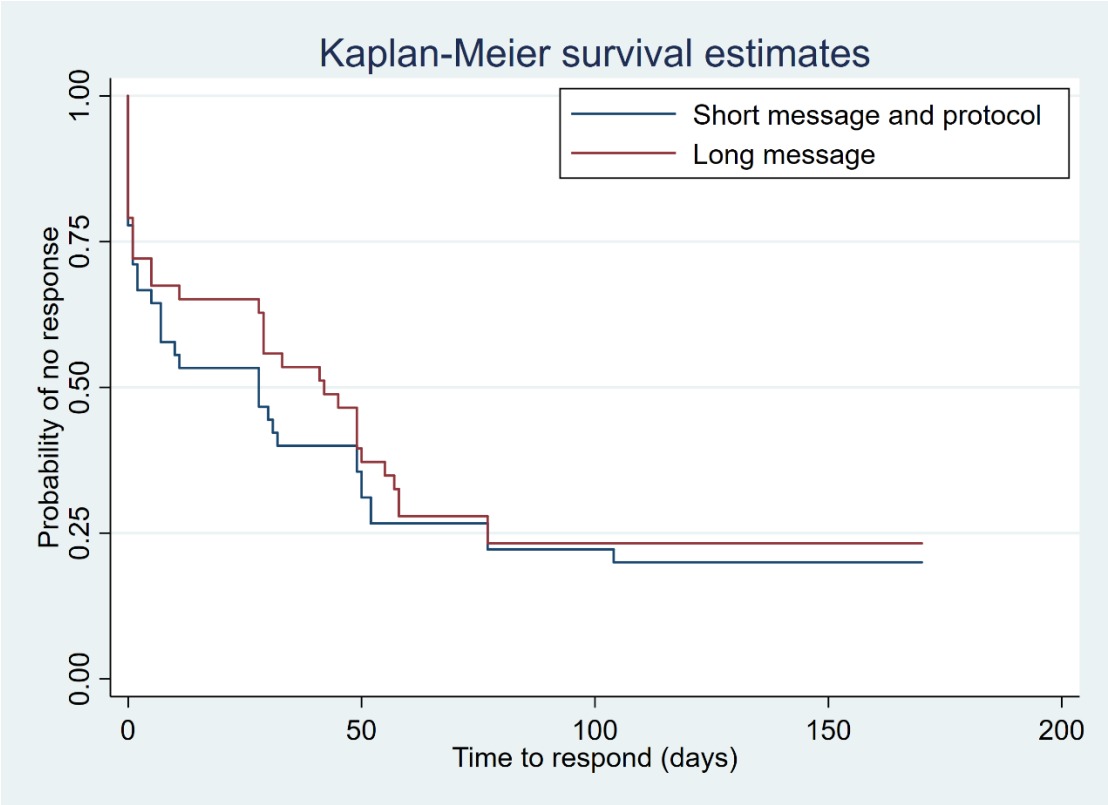
Average time to response was marginally quicker among those allocated to Short (median = 28 days) compared with Long (median = 42 days), with a greater proportion of authors in this trial arm responding to the first email they received (Table 29). However, there was insufficient evidence to determine a difference between trial arms on time to response (Table 29 & Figure 17). The steep drops seen in Figure 17 all occur just after an email is sent. In total, 27% of authors responded within a day and 38% replied within a week.

Table 29 Time to response (days)

	Time to Response				
	Mean [sd]	**Median [25th, 75th centile]**	**Adjusted Hazard Ratio**	**95% C.I**	**p-value**
Short message and protocol	51.6 [64.7]	28 [1, 77]	0.91	(0.55, 1.51)	0.72
Long message	60 [65.3]	42 [1, 77]			

Adjusted for year of publication, size of trial and if the author had multiple trials included. Total N included = 88.

Figure 17 Kaplan-Meier curve for time to response for both intervention groups



4.3.4 First response outcome and eventual collaboration

A similar number of emails were needed before a response was received for both trial arms. A short email elicited a more favourable first response, with few negative replies. However, this did not lead to a larger number of authors agreeing to collaborate, with many authors not following up on their positive intent. Conversely, those allocated the longer email were more likely to respond in a negative fashion, but this meant that the decision not to collaborate was established sooner, and there was a smaller proportion with no decision reached (Table 30).

Table 30 Secondary outcomes

	Short message and protocol (n=36)	Long message (n=40)
Number of emails before response received		
1	16 (44%)	15 (38%)
2	6 (17%)	10 (25%)
3	3 (8%)	3 (8%)
4	2 (6%)	2 (5%)
No response	9 (25%)	10 (25%)
Mean [sd]	1.7 [1.0]	1.7 [0.9]
Median [25th, 75th centile]	1 [1, 2]	1.5 [1, 2]
Min, Max	1, 4	1, 4
First response outcome		
Negative	2 (6%)	8 (20%)
Neutral	9 (25%)	14 (35%)
Positive	16 (44%)	8 (20%)
Eventual outcome		
Agree to collaborate	11 (31%)	10 (25%)
Do not agree to collaborate	3 (8%)	11 (28%)
No decision reached	13 (36%)	9 (23%)
No response	9 (25%)	10 (25%)

Data are N(%) unless otherwise indicated. Data are presented on author level, rather than trial level.

4.4 Discussion

This chapter has described a nested randomised trial of the effect of email format on author response when attempting to acquire data for a systematic review. There was no evidence to suggest that either email format was preferable to the other. Authors responded in a similar timeframe, and needed a similar number of reminders before a response was received. There was some indication that authors that received a shorter email were more likely to respond positively at first, but there was no evidence that email format changed the likelihood of collaboration.

Approximately three-quarters of authors in the study responded regardless of email format. A study which attempted to contact authors to obtain data for a diagnostic accuracy review for hepatic fibrosis found a similar response rate, with 68% of authors responding to requests for data[196]. For both email formats, text was included that invited authors who provided data into a collaboration to assist (at their discretion) with interpretation and write-up of results, and offered co-authorship as part of a collaborative group. It was thought that offering greater opportunity for collaboration to authors rather than simply providing data or further results might produce a greater and more complete response. Offering collaboration may have influenced overall response rate, and thus the results of this study may not be generalisable to others where collaboration is not offered. For example, a study which aimed to establish whether corresponding authors accepted responsibility of correspondence, where collaboration was not offered, found a far

lower response rate (190/446, 43%)[197]. A qualitative interview study which investigated strategies to access unpublished clinical trial data mentioned a lack of collaboration as a main barrier to data sharing[198]. Thus, offering the opportunity for authors to collaborate and being upfront about giving collaborators the chance to input in both interpreting results and the write-up of any papers could result in systematic reviewers receiving both a more positive and a more thorough response.

Whilst short emails elicited a marginally quicker response on average, there was insufficient evidence to conclude benefit of this email style on time to response. Selph et al.[196] found that responses were received from 23% of authors after one request, with 55% responding after just two requests (one reminder). This is similar to this trial, where I found that 41% of authors had responded after one email and two emails (one reminder) elicited 62% response rate. A third email provided 70% response rate, with little further improvement following a fourth email. The number of emails sent and the number sent to each contact (if there are multiple like in this study) are additional areas where similar research could help to introduce a more rigorous evidence base.

There was some indication that authors allocated to a short email format were more likely to respond positively, with only two negative first responses received from these authors, and double the number of positive responses. However, the number of authors eventually agreeing to collaborate was similar for both groups. This suggests that whilst the long email format resulted in more negative initial responses,

authors who did respond were aware of the implications of this study and could make an informed decision. Thus, the majority of positive first responses to those allocated to the long email format turned into collaboration. It could be hypothesised that whilst those allocated to the short email format had a detailed protocol attached, authors may not have opened attachments and therefore were agreeing to be involved in the study without fully realised the commitment. This led to a far larger amount of time and resource dealing with these authors, who eventually declined to collaborate, or no decision was reached in the trial timeframe.

A limitation of this study was the constrained sample size, enforced by the systematic review within which this study was nested. Thus, power to detect a small difference in response rate was low. However, this study does provide a framework to carry out a larger trial of this kind. A further limitation of this trial was that blinding was not possible, because the investigator was responsible for sending all emails and collating the data. It was impossible to mask which intervention participants were assigned to, due to the email thread being used to reply to messages and send reminders. Therefore, there is a possibility that some outcomes are at a risk of bias. Furthermore, the eventual outcome was not expected to be greatly influenced by the intervention, and was more reliant on subsequent communication such as emails or phone contact. However, the protocol was finalised in advance of randomisation, and the trial team followed this protocol, therefore the risk of bias from the trial team was low. In addition, study authors, although aware of email

format they received, were not aware of the alternative email format nor were they aware of the nested randomised study, therefore the risk of bias from participants was low.

It could be argued that this study did not investigate what could be perceived as the key question, which was whether data was eventually shared. This was due to the belief that whether authors provided data was dependent on other factors, and not the original email length and style. This study investigated an intervention which allows researchers to open up a channel of communication with the data custodian, and there is the potential to use a different intervention to evaluate how best to acquire data (for example financial incentives[72]).

In summary, this nested randomised trial found no evidence that email length had an impact on author response, time to response, or agreement to collaborate, when attempting to acquire data for the systematic review described in Chapter 3.

The next chapter reverts back to central adjudication in randomised trials and builds on the results of the systematic review and meta-analysis described in Chapter 3. Chapter 5 details a simulation study that uses data from five of the fifteen trials included in the review and explores how much bias (introduced through differential misclassification) needs to be introduced before central adjudication would give a different results to the one with only site-assessment of the primary outcome. Additionally, hypothetical trials are simulated that evaluate whether central adjudication for a study with adequate blinding

is useful, and how much error is needed before central adjudication changes trial conclusions.

Chapter 5 – Outcome assessment by central adjudicators in randomised stroke trials: simulation of differential and non-differential misclassification

Publications contributing to this chapter:

1. PJ Godolphin, PM Bath, C Partlett, E Berge, MM Brown, M Eliasziw, PM Sandset, J Serena, AA Montgomery. Central adjudication of outcomes is most valuable in unblinded trials: a simulation study. [Under Review].

2. PJ Godolphin, PM Bath, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in randomised stroke trials. 2019. Poster presentation at 40th International Society for Clinical Biostatisticians Conference.

3. PJ Godolphin, PM Bath, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators in randomised stroke trials: simulation of differential and non-differential misclassification. 2019. Poster presentation at 5th International Clinical Trials Methodology Conference.

Abstract

Background: Adjudication of the primary outcome is thought to control misclassification. In this chapter, the amount of misclassification needed before adjudication changed trial results is investigated.

Methods: Data from five randomised stroke trials were included. Differential misclassification was introduced for each study until the estimated treatment effect was altered. This was simulated 1000 times. The between-simulation mean proportion of participants that needed to be differentially misclassified to alter the treatment effect was calculated. In addition, hypothetical trials were simulated with a binary outcome and varying sample size (1000-10000), overall event rate (10-50%), and treatment effect (0.67-0.90). Non-differential misclassification was simulated until the treatment effect was non-significant at 5%.

Results: For the five stroke trials, unweighted kappa was reduced from 0.89-0.97 to 0.65-0.85 before the treatment effect was altered. This corresponded to 2.1%-6% of participants misclassified differentially for trials with a binary outcome. For the hypothetical trials, those with a larger sample size, stronger treatment effect and overall event rate closer to 50% needed a higher proportion of events non-differentially misclassified before the treatment effect became non-significant.

Conclusions: Adjudication is important to control for differential misclassification in a clinical trial. However, for large blinded studies, extensive random noise is required before adjudication changes conclusions.

5.1 Introduction

The previous chapter explored methods for contacting authors, which may be particularly useful for researchers undertaking a systematic review involving adjudication. Chapter 3 showed that central adjudication does not alter the estimated treatment effect in stroke trials. This chapter investigates how much misclassification would be needed in the site investigators assessment of the primary outcome before central adjudication would alter trial conclusions.

To investigate when central adjudication would change a trial's results it is useful to explore how much differential misclassification by site investigators is necessary to alter the estimated treatment effect, i.e. the confidence interval of the RTE excludes the null value, one. However, this investigation ignores the statistical significance of the treatment estimate pre- and post-misclassification and only identifies when the treatment effect estimate differs significantly to the estimate obtained after central adjudication (RTE≠1). Therefore, for completeness, situations can be considered where the RTE remains at one after misclassification (here non-differential), and the significance of the treatment estimate differs pre- and post-misclassification.

Thus, the aims of this chapter are to: (1) evaluate how much differential misclassification is needed to alter the estimated treatment effect; and (2) explore how much non-differential misclassification will cause a significant treatment effect to become non-significant at 5% level.

5.2 Methods

5.2.1 Differential misclassification using real trial data

For studies with adequate blinding, central adjudication should control for non-differential misclassification by reducing random 'noise' around the main estimate of interest. However, increasing this 'noise' in a simulation will not meaningfully shift the estimate of interest, because the amount of misclassification in a blinded trial will be equal in both treatment arms. Therefore, to explore the situation where central adjudication does alter the treatment effect estimate, I introduced differential misclassification. Previous studies have shown that site investigators often exaggerate treatment effect estimates[30, 33, 34], so differential misclassification of outcomes assessed by site investigators was introduced to make the treatment effect based on outcome data reported by local site assessors more beneficial. The starting point for misclassification was the centrally adjudicated data, as this is the gold standard, and outcomes were misclassified to increasing extent. This misclassification differs for binary and ordinal variables, as explained below.

5.2.1.1 Data collection

The systematic review of central adjudication in stroke trials described in Chapter 3 included 15 trials totalling 69,650 participants[194]. All included trials were asked to provide either summary results or individual patient data (IPD). Only those trials that provided IPD were considered for the inclusion in the simulation work presented here, as

differential misclassification is introduced at a patient level. Five studies were selected for this simulation study as they covered a variety of outcomes, number of participants randomised, and treatment effectiveness, as well as providing suitable IPD[178, 179, 183, 185, 186, 190].

The five studies selected corresponded to seven unique populations as one study, NASCET, carried out separate analyses for patients with mild-, moderate- and severe-grade carotid artery stenosis (denoted as NASCET:mild, NASCET:moderate and NASCET:severe respectively). Throughout the remainder of this chapter, for ease of reporting, these will be referred to as individual trials, as they were analysed separately and each contain distinct populations. Therefore, in this simulation study there were seven trials included (HAEST, ICSS, REVASCAT, TARDIS, and the three aforementioned NASCET subpopulations).

5.2.1.2 Misclassification for binary outcomes

For binary outcomes, differential misclassification was introduced by increasing the proportion of participants who (a) were in the control arm and had an event, and, (b) were in the treatment arm and did not have an event. For each trial, a random number was generated that followed the standard uniform distribution (achieved using runiform in Stata, starting seed 2206 for each study). Following this, a proportion of participants were differentially misclassified. Participants were chosen randomly using the random number generated (e.g. if 10% of participants were to be misclassified, then any participant that had a random number ≤0.1 was misclassified, if appropriate). Only participants in the control arm without the event and participants in the

treatment arm with the event were misclassified, as the objective was to introduce differential misclassification.

For trials where the analysis involved time-to-event data, the median difference in time between any events misclassified in the trial was calculated. For any participants misclassified to have an event in the present study, the event time for these participants was reduced by the median calculated previously. This time was constrained at the lowest time of an event witnessed in the trial.

5.2.1.3 Misclassification for ordinal outcomes

For ordinal outcomes, a similar approach was taken, with a random number generated in the same fashion. However, the misclassification here varies due to the ordinal nature of the outcome. In both trials where the outcome was analysed in an ordinal fashion, participants could be allocated one of six categories (this was due to REVASCAT having the mRS levels for 5 and 6 combined, and TARDIS having an ordinal primary outcome with six categories originally). Using the real trial data, I calculated the probability of any participant being misclassified (site investigator vs central adjudicator) by one, two, three, four and five categories.

To simulate increased differential misclassification, selected participants in the control arm had their outcome value increased (worse outcome) and those in the treatment group had their outcome value decreased (better outcome). As the proportion of participants misclassified in the simulation increased, the number of participants

misclassified by one category, two categories and so on, increased proportionally using the probabilities calculated. Outcomes were constrained by the minimum (0) and maximum (5) values.

5.2.1.4 The proportion of misclassification necessary to alter the estimated treatment effect

Increasing differential misclassification was introduced in 0.1% increments of participants, following which the trial's primary outcome was reanalysed. The treatment effect was then compared with that calculated based on central adjudicated data (remains constant for each trial) using the ratio of treatment effects (RTE). An RTE < 1 indicates that the misclassified data produces a more beneficial treatment effect. For each 0.1% increment in the percentage of participants misclassified, 1000 simulations were run, from which the mean RTE and 95% confidence interval were calculated. An upper bound of the 95% confidence interval less than 1 indicates there is sufficient misclassification in the data to produce a treatment effect that is different to the study result based on central adjudicated outcome data.

5.2.1.5 Statistical analysis

Percent agreement and unweighted kappa were calculated between central adjudicators and site investigators for the primary outcome of each trial before misclassification. For trials with ordinal outcomes, weighted kappa was also determined, where the weighting used linear weights (linear weights were used following consultation with trialists and stroke researchers). Each trial was analysed as per the analysis

specified in their main results paper, except for the three NASCET trials, where a univariate Cox proportional hazards model was fitted for each trial.

After simulation, the within-simulation mean and standard deviation of the treatment effect after misclassification, number of participants misclassified, crude percent agreement and unweighted (and weighted if appropriate) kappa were determined for each trial. One of the 1000 simulations was selected to give an example primary outcome split by treatment group, and an example of the agreement between central adjudicators and the misclassified site investigators for the primary outcome. All analyses, including those described in the following sections, were undertaken using Stata version 15.1 or later.

5.2.2 Non-differential misclassification using hypothetical trial data

For studies with adequate blinding, any misclassification of an outcome is expected to be equal between treatment and control arms, that is, non-differential. For these studies, the RTE will be close to one even with introduction of a large amount of non-differential misclassification. However, this could still impact on trial conclusions by introducing greater random error resulting in wider 95% confidence intervals around the estimated treatment effect. Thus, the amount of non-differential misclassification required to cause a loss of precision such that the 95% confidence interval for a real treatment effect no longer excludes the null can be estimated. Furthermore, it would be interesting to explore

whether trials with certain characteristics are more susceptible to treatment effects being missed due to random error.

5.2.2.1 Data generation

Data was generated using Stata to represent a simple parallel group trial with a binary primary outcome. This was chosen as 13/15 stroke trials included in the systematic review described in Chapter 3 had a binary outcome. I estimated the treatment effect using a relative risk and significance level was set at 5%. The aim of this part of the simulation study was to establish how much non-differential misclassification was required for a previously significant treatment effect to become non-significant. In order to do this, I chose to simulate the entire dataset to enable greater manipulation of specific trial characteristics that were varied.

5.2.2.2 Characteristics to vary

Three different treatment effects were chosen: relative risks of 0.67 (for example, events in a ratio of 3:2 between control and treatment groups respectively), 0.82 (ratio of 11:9) and 0.90 (ratio of 21:19) to represent strong, moderate and modest treatment effects respectively. In stroke trials (and many other clinical areas) overall event rate is low, so situations were explored where the overall event rate was ≤50%.The overall event rate was increased in 10% intervals, from 10% to 50%. Finally, the overall trial sample size was simulated to be either 1000, 3000, 5000 or 10000, to represent a number of plausible scenarios in a large number of trial settings. Thus, by varying sample size, overall

event rate and treatment effect there were 60 distinct scenarios. This is summarised is Table 31.

Table 31 Summary of parameters used in the simulation of non-differential misclassification

Description	Values
Treatment effect	0.67, 0.82, 0.90
Overall event rate	10%, 20%, 30%, 40%, 50%
Sample size	1000, 3000, 5000, 10000

5.2.2.3 Misclassifying events

For each scenario, events were misclassified proportionately in each arm in order to preserve the relative risk and thus keep the RTE equal to one. For example, to preserve a treatment effect of relative risk = 0.67 three events were as non-events in the treatment group for every two events misclassified in the control group. The amount of misclassification required for the relative risk 95% confidence interval to include the null value of one was expressed as a percentage of the total number of events in the original dataset.

5.3. Results:

5.3.1 Differential misclassification using real trial data

For five of the trials, the primary outcome was binary, whereas for the remaining two trials the primary outcome was analysed on an ordinal scale (Table 32). The number of participants randomised varied between 206 (REVASCAT) and 3096 (TARDIS). Agreement was high between central adjudicators and site investigators, with crude agreement ranging from 93.2% to 99.6% and kappa ranging from 0.89 to 0.97 (Tables 33-37).

For trials with a binary outcome, between 2.1% and 6% of participants needed to be differentially misclassified to alter the estimated treatment effect obtained through central adjudication, with the amount of misclassification inversely associated with study size (Table 38). In the two trials with ordinal primary outcomes, there was substantial variation in the proportion of participants that needed to be misclassified (1.9% and 27.8%). However, these studies did represent the trials with the largest and smallest number of participants respectively.

Following misclassification, crude agreement remained high for all but one of the trials, but kappa was considerably reduced with mean between-simulation unweighted kappa values ranging from 0.65 to 0.85 (Tables 38-43). Weighted kappa was higher for the two ordinal trials, but still resulted in a drop from 0.96 to 0.84 and 0.91 to 0.87.

Table 32 Summary of trials included in simulation study

Trial name	Population	Intervention	Comparator	Primary outcome	Were site investigators blind to treatment arm?	Adjudication information
HAEST	Patients with acute ischaemic stroke and atrial fibrillation (n=449).	Dalteparin (n=224)	Aspirin (n=225)	Recurrent ischaemic Stroke. Analysed using logistic regression model.	Yes	Two clinicians assessed medical notes (including reports from cranial scans) and original case report forms with diagnosis concealed.
ICSS	Patients with symptomatic carotid stenosis (n=1713)	Stenting (n=855)	Carotid endarterectomy (n=858)	Fatal or disabling stroke. Analysed using cox proportional hazards model.	No	Two clinicians assessed outcome without knowledge of site assessment
NASCET	Patients with non-disabling stroke and carotid stenosis of 30-99% in the internal carotid artery. There were three populations mild (<50%, n=1368), moderate (50-69%, n=858) and severe (70-99%, n=659) stenosis.	Carotid endarterectomy. In addition, patients received medical care, including antiplatelet therapy. Mild (n=678), moderate (n=430), severe (n=328).	Medical care, including antiplatelet therapy. Mild (n=690), moderate (n=428), severe (n=331)	Fatal or non-fatal ipsilateral stroke. Analysed using Mantel–Haenszel chi-square test. To obtain an estimate, the NASCET trials were analysed using univariate cox-proportional hazard model.	No	Neurologists and surgeons assessed original case report forms and cranial scans without knowledge of site assessment.
REVASCAT	Patients with acute ischaemic stroke who could be treated within 8 hours (n=206)	Medical therapy (including alteplase if eligible) and thrombectomy (n=103)	Medical therapy (including alteplase if eligible) (n=103)	Functional outcome at 90 days (mRS). Patients who scored 5 or 6 were grouped in a single category. Analysed using ordinal logistic regression model (6 point scale)	Yes	Neurologists assessed audio-tape or video recording of patient evaluation of the primary outcome.

TARDIS	Patients with acute ischaemic stroke or TIA (n=3096).	Aspirin, clopidogrel and dipyridamole (n=1556)	Aspirin and dipyridamole, or clopidogrel alone (n=1540)	Functional outcome and recurrent stroke and TIA. Analysed using ordinal logistic regression (6 point scale)	Yes	Clinicians assessed medical notes, original case report forms and cranial scans, if requested.

mRS refers to modified Rankin Scale, TIA refers to transient ischaemic attack

Table 33 Agreement between central adjudicators and site investigators using original trial data

Trial	Central adjudicator data			Site investigator data			Agreement between central adjudicators and site investigators			Crude agreement	Kappa
HAEST		Treated	Control		Treated	Control	SI \ CA	No event	event	99.6%	0.97*
	No event	205	208	No event	203	208	No event	411	0		
	Event	19	17	Event	21	17	event	2	36		
ICSS		Treated	Control		Treated	Control	SI \ CA	No event	event	98.8%	0.89*
	No event	808	801	No event	812	802	No event	1600	14		
	Event	49	52	Event	44	50	event	7	87		
NASCET: mild		Treated	Control		Treated	Control	SI \ CA	No event	event	98.9%	0.96*
	No event	589	580	No event	592	584	No event	1165	11		
	Event	89	110	Event	86	106	event	4	188		
NASCET: moderate		Treated	Control		Treated	Control	SI \ CA	No event	event	99.2%	0.97*
	No event	373	348	No event	374	354	No event	721	7		
	Event	57	80	Event	56	74	event	0	130		
NASCET: severe		Treated	Control		Treated	Control	SI \ CA	No event	event	99.2%	0.97*
	No event	300	264	No event	299	268	No event	563	4		
	Event	28	67	Event	29	63	event	1	91		
REVASCAT	see Table 34						see Table 35			93.2%	0.91* 0.96†
TARDIS	see Table 36						see Table 37			98.8%	0.91* 0.91†

SI refers to Site investigators; CA refers to Central adjudicators
*Unweighted kappa
†Weighted kappa using linear weights

Table 34 Functional outcome by treatment group using central adjudicator and site investigator data in REVASCAT

	Central adjudicator data		Site investigator data	
Outcome	**Treated (n=103)**	**Control (n=103)**	**Treated (n=103)**	**Control (n=103)**
mRS 0	7	6	6	3
mRS 1	18	7	23	10
mRS 2	20	16	20	16
mRS 3	19	20	15	20
mRS 4	8	17	7	15
mRS 5/6	31	37	32	39

mRS refers to modified Rankin Scale

Table 35 Agreement on functional outcome between central adjudicators and site investigators in REVASCAT

	Central Adjudicators						
Site Investigators	mRS 0	mRS 1	mRS 2	mRS 3	mRS 4	mRS 5/6	**Total**
mRS 0	9	0	0	0	0	0	9
mRS 1	4	25	3	1	0	0	33
mRS 2	0	0	33	3	0	0	36
mRS 3	0	0	0	35	0	0	35
mRS 4	0	0	0	0	22	0	22
mRS 5/6	0	0	0	0	3	68	71
Total	13	25	36	39	25	68	206
Disagreements (%)	4 (31%)	0 (-)	3 (8%)	4 (10%)	3 (12%)	0 (-)	14 (7%)

Crude agreement = 192/206 = 93%, unweighted kappa = 0.91, weighted kappa using linear weights = 0.96. mRS refers to modified Rankin Scale

Table 36 Functional outcome by treatment group using central adjudicator and site investigator data in TARDIS

	Central adjudicator data		Site investigator data	
Outcome	**Treated (n=1556)**	**Control (n=1540)**	**Treated (n=1556)**	**Control (n=1540)**
No recurrent event	1463	1435	1457	1434
TIA	32	48	34	53
Stroke: mRS 0/1	15	18	16	16
Stroke: mRS 2/3	22	23	23	22
Stroke: mRS 4/5	11	9	13	10
Fatal stroke: mRS 6	13	7	13	5

mRS refers to modified Rankin Scale

Table 37 Agreement on incidence and severity of stroke between central adjudicators and site investigators in TARDIS

	Central Adjudicators						
Site Investigators	No recurrent event	TIA	Stroke: mRS 0/1	Stroke: mRS 2/3	Stroke: mRS 4/5	Fatal stroke: mRS 6	**Total**
No recurrent event	2881	1	3	4	0	2	2891
TIA	5	77	4	1	0	0	87
Stroke: mRS 0/1	4	2	26	0	0	0	32
Stroke: mRS 2/3	5	0	0	40	0	0	45
Stroke: mRS 4/5	3	0	0	0	20	0	23
Fatal stroke: mRS 6	0	0	0	0	0	18	18
Total	2898	80	33	45	20	20	3096
Disagreements (%)	17 (1%)	3 (4%)	7 (21%)	5 (11%)	0 (-)	2 (10%)	34 (1%)

Crude agreement = 3062/3096 = 98.9%, unweighted kappa = 0.91, weighted kappa using linear weights = 0.91. TIA refers to Transient Ischaemic Attack, mRS refers to modified Rankin Scale

Table 38 Number and proportion of participants required to be differentially misclassified to alter estimated treatment effect

Trial (N)	Treatment effect before misclassification (95% CI)	Mean treatment effect after misclassification (SD)	Mean number of participants misclassified (SD)	Mean percentage of participants misclassified (SD)	RTE (95% CI)
HAEST (n=449)	1.13 (0.57, 2.24)	0.45 (0.07)	20.4 (4.3)	4.5% (1.0%)	0.40 (0.16, 0.99)
ICSS (n=1710)	0.94 (0.64, 1.39)	0.54 (0.04)	35.9 (5.9)	2.1% (0.3%)	0.59 (0.34, 0.99)
NASCET: mild (n=1368)	0.83 (0.72, 0.95)	0.55 (0.03)	55.0 (7.0)	4.0% (0.5%)	0.68 (0.46, 0.99)
NASCET: moderate (n=858)	0.71 (0.60, 0.84)	0.43 (0.03)	51.4 (6.7)	6.0% (0.8%)	0.63 (0.39, 0.99)
NASCET: severe (n=659)	0.36 (0.28, 0.43)	0.19 (0.02)	39.3 (5.9)	6.0% (0.9%)	0.53 (0.28, 0.99)
REVASCAT (n=206)	0.57 (0.35, 0.95)	0.28 (0.02)	57.2 (5.8)	27.8% (2.8%)	0.48 (0.23, 0.99)
TARDIS (n=3096)	0.90 (0.67, 1.20)	0.59 (0.03)	60.3 (7.3)	1.9% (0.2%)	0.66 (0.44, 0.99)

Data from 1000 simulations (starting seed 2206). Treatment effects lower than one indicates treatment is beneficial. SD refers to standard deviation. RTE refers to ratio of treatment effects.

Table 39 Agreement between central adjudicators and site investigators on primary outcome after differential misclassification

Trial	Central adjudicator data			Example misclassified site investigator data			Example agreement between central adjudicators and misclassified site investigators			Mean crude agreement (SD)	Mean kappa (SD)
HAEST		Treated	Control		Treated	Control	SI \ CA	No event	event	95.4% (1.00%)	0.75 (0.05)*
	No event	205	208	No event	206	189	No event	394	1		
	Event	19	17	Event	18	36	event	19	35		
ICSS		Treated	Control		Treated	Control	SI \ CA	No event	event	97.9% (0.34%)	0.84 (0.02)*
	No event	808	801	No event	811	769	No event	1577	3		
	Event	49	52	Event	46	84	event	32	98		
NASCET: mild		Treated	Control		Treated	Control	SI \ CA	No event	event	96.0% (0.51%)	0.85 (0.02)*
	No event	589	580	No event	594	534	No event	1123	5		
	Event	89	110	Event	84	156	event	46	194		
NASCET: moderate		Treated	Control		Treated	Control	SI \ CA	No event	event	94.0% (0.78%)	0.80 (0.02)*
	No event	373	348	No event	383	320	No event	683	10		
	Event	57	80	Event	47	128	event	38	127		
NASCET: severe		Treated	Control		Treated	Control	SI \ CA	No event	event	94.0% (0.88%)	0.79 (0.03)*
	No event	300	264	No event	303	231	No event	531	3		
	Event	28	67	Event	25	100	event	33	92		
REVASCAT	see Supplementary Table 40						see Supplementary Table 41			72.2% (2.80%)	0.65 (0.03)* 0.84 (0.02)†
TARDIS	see Supplementary Table 42						see Supplementary Table 43			98.1% (0.24%)	0.85 (0.02)* 0.87 (0.02)†

Crude agreement and kappa are from 1000 simulations (starting seed 2206). Example site investigator data and example agreement are taken from one of the 1000 simulations. SI refers to Site investigators; CA refers to Central adjudicators

*Unweighted kappa

†Weighted kappa using linear weights

Table 40 Functional outcome by treatment group using central adjudicator and misclassified site investigator data in REVASCAT

	Central adjudicator data		Example misclassified site investigator data	
Outcome	**Treated (n=103)**	**Control (n=103)**	**Treated (n=103)**	**Control (n=103)**
mRS 0	7	6	14	3
mRS 1	18	7	15	7
mRS 2	20	16	21	11
mRS 3	19	20	19	20
mRS 4	8	17	14	19
mRS 5/6	31	37	20	43

Misclassified site investigator data is from one of 1000 simulations (starting seed 2206). mRS refers to modified Rankin Scale

Table 41 Agreement on functional outcome between central adjudicators and example misclassified site investigators in REVASCAT (data from one of 1000 simulations)

	Central Adjudicators						
Site Investigators	mRS 0	mRS 1	mRS 2	mRS 3	mRS 4	mRS 5/6	**Total**
mRS 0	10	7	0	0	0	0	17
mRS 1	3	15	4	1	0	0	22
mRS 2	0	3	24	5	0	0	32
mRS 3	0	0	8	26	4	1	39
mRS 4	0	0	0	8	15	10	33
mRS 5/6	0	0	0	0	6	57	63
Total	13	25	36	39	25	68	206
Disagreements (%)	3 (23%)	10 (40%)	12 (33%)	14 (36%)	10 (40%)	11 (16%)	60 (29%)

Mean crude agreement (SD) = 72.2% (2.80%), mean unweighted kappa (SD) = 0.65 (0.03), mean weighted kappa using linear weights (SD) = 0.84 (0.02). Mean crude agreement, unweighted and weighted kappa are from 1000 simulations (starting seed 2206). mRS refers to modified Rankin Scale

Table 42 Functional outcome by treatment group using central adjudicator and misclassified site investigator data in TARDIS

	Central adjudicator data		Example misclassified site investigator data	
Outcome	**Treated (n=1556)**	**Control (n=1540)**	**Treated (n=1556)**	**Control (n=1540)**
No recurrent event	1463	1435	1472	1403
TIA	32	48	31	58
Stroke: mRS 0/1	15	18	16	26
Stroke: mRS 2/3	22	23	17	31
Stroke: mRS 4/5	11	9	10	12
Fatal stroke: mRS 6	13	7	10	10

Misclassified site investigator data is from one of 1000 simulations (starting seed 2206). mRS refers to modified Rankin Scale

Table 43 Example agreement on incidence and severity of stroke between central adjudicators and misclassified site investigators in TARDIS (data from one of 1000 simulations)

	Central Adjudicators						
Misclassified site Investigators	No recurrent event	TIA	Stroke: mRS 0/1	Stroke: mRS 2/3	Stroke: mRS 4/5	Fatal stroke: mRS 6	**Total**
No recurrent event	2866	4	0	4	0	1	2875
TIA	10	76	2	0	0	1	89
Stroke: mRS 0/1	8	0	31	2	1	0	42
Stroke: mRS 2/3	8	0	0	39	1	0	48
Stroke: mRS 4/5	3	0	0	0	18	1	22
Fatal stroke: mRS 6	3	0	0	0	0	17	20
Total	2898	80	33	45	20	20	3096
Disagreements (%)	32 (1%)	4 (5%)	2 (6%)	6 (13%)	2 (10%)	3 (15%)	49 (2%)

Mean crude agreement (SD) = 98.1% (0.24%), mean unweighted kappa (SD) = 0.85 (0.02), mean weighted kappa using linear weights (SD) = 0.87 (0.02). Mean crude agreement, unweighted and weighted kappa are from 1000 simulations (starting seed 2206). TIA refers to Transient Ischaemic Attack, mRS refers to modified Rankin Scale

5.3.2 Non-differential misclassification using hypothetical trial data

For 14 of the scenarios, the initial risk ratio was not significant at 5% before misclassification, so these cases are not given (displayed as NA in Table 44 and not displayed in any of Figures 18-20). As expected, more events were required to be misclassified to change a significant treatment effect to non-significant at the 5% level when the original treatment effect was strongest (Figures 18-20, Table 44).

Greater sample size and higher overall event rate both required a larger proportion of events to be misclassified before significant treatment effects become non-significant (Figures 18-20). For example, in a hypothetical blinded trial with 5000 participants, overall event rate of 50% and a strong treatment effect, 2405 (96.2%) of the events would need to be misclassified non-differentially before a significant treatment would become non-significant.

The lowest amount of events that needed to be non-differentially misclassified before a significant treatment effect would have become non-significant was the scenario of a weak treatment effect (risk ratio = 0.90), small sample size (n = 1000) and moderate overall event rate (rate = 0.30%). For this scenario, nearly 10% of the events witnessed in the trial would have had to have been misclassified, resulting in crude agreement of 97.1% between the central adjudicators and site investigators. The lowest crude percent agreement for a binary outcome from the trials involved in the simulation study was 98.8%.

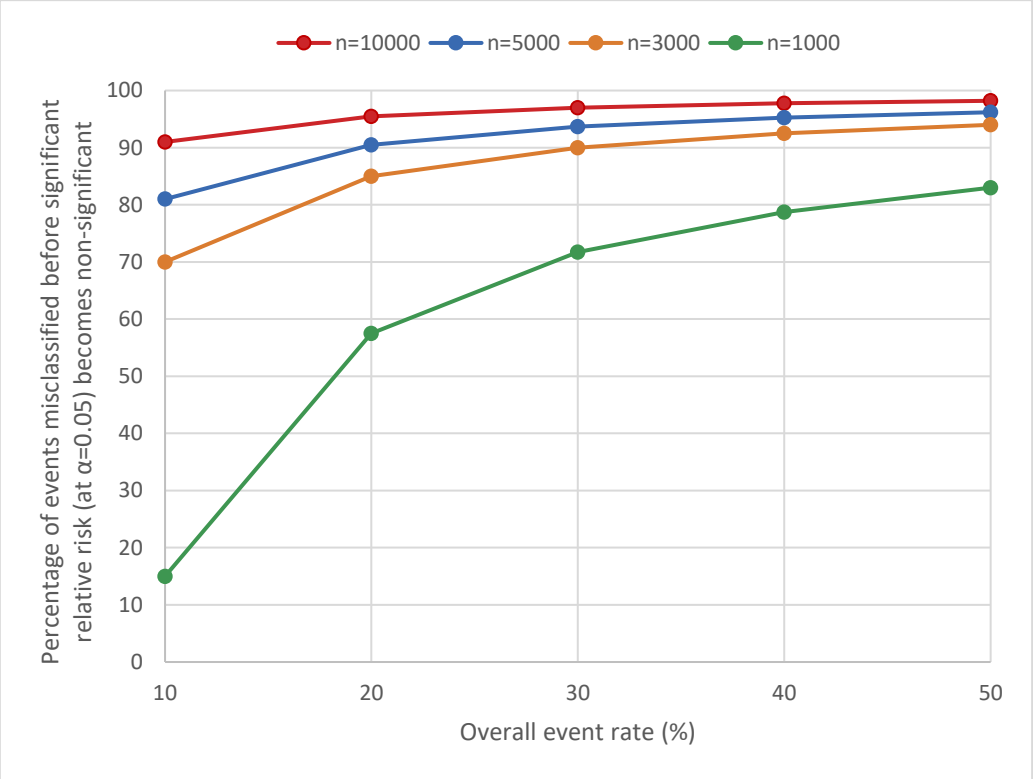
Thus, even for this scenario, the extent of disagreement required for a relatively objective outcome (as stroke often is) is not regularly seen in randomised trials.

Table 44 Number and proportion of non-differentially misclassified events required such that treatment effect is no longer significant at 5% level

		Overall event rate				
Proportion of events: Treatment vs Control		**10%**	**20%**	**30%**	**40%**	**50%**
N=10000	**60% vs 40%**	910/1000 (91%)	1910/2000 (95.5%)	2910/3000 (97%)	3910/4000 (97.75%)	4910/5000 (98.2%)
	55% vs 45%	629/1000 (62.9%)	1629/2000 (81.45%)	2629/3000 (87.63%)	3629/4000 (90.73%)	4629/5000 (92.58%)
	52.5% vs 47.5%	NA	680/2000 (34%)	1680/3000 (56%)	2680/4000 (67%)	3690/5000 (73.8%)
N=5000	**60% vs 40%**	405/500 (81%)	905/1000 (90.5%)	1405/1500 (93.63%)	1905/2000 (95.25%)	2405/2500 (96.2%)
	55% vs 45%	149/500 (29.8%)	649/1000 (64.9%)	1149/1500 (76.6%)	1649/2000 (82.45%)	2149/2500 (85.96%)
	52.5% vs 47.5%	NA	NA	329/1500 (21.93%)	829/2000 (41.45%)	1329/2500 (53.16%)
N=3000	**60% vs 40%**	210/300 (70%)	510/600 (85%)	810/900 (90%)	1110/1200 (92.5%)	1410/1500 (94%)
	55% vs 45%	NA	260/600 (43%)	560/900 (62%)	860/1200 (71.63%)	1160/1500 (77.33%)
	52.5% vs 47.5%	NA	NA	NA	200/1200 (17%)	499/1200 (41.6%)
N=1000	**60% vs 40%**	15/100 (15%)	115/200 (57.5%)	215/300 (71.63%)	315/400 (78.75%)	415/500 (83%)
	55% vs 45%	NA	NA	29/300 (9.67%)	129/400 (32.25%)	229/500 (45.8%)
	52.5% vs 47.5%	NA	NA	NA	NA	NA

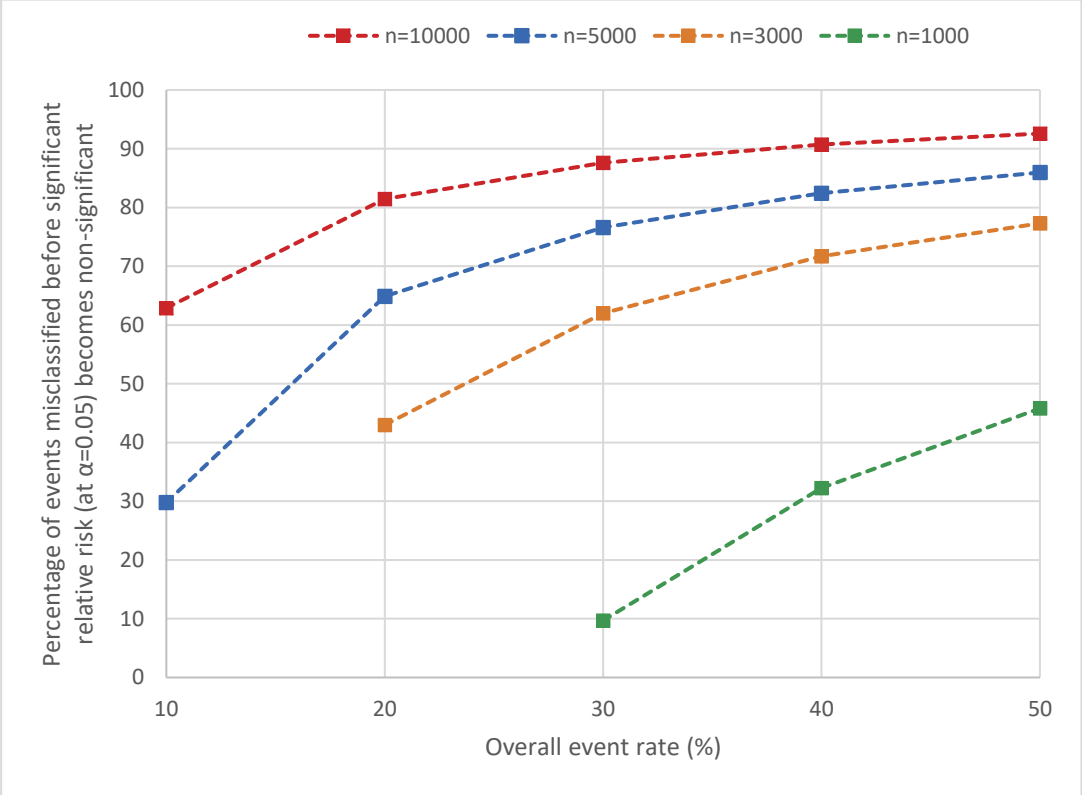
Data are number of events misclassified/total number of events (%)
Significant treatment effect is set at α=0.05 and is from a risk ratio
Proportion of events 60% vs 40% corresponds to a treatment effect of 0.67, a proportion of events 55% to 45% corresponds to a treatment effect of 0.82, and a proportion of events 52.5% to 47.5% corresponds to a treatment effect of 0.90. A treatment effect less than one indicates treatment is beneficial.
NA refers to scenarios where the initial treatment effect before misclassification was non-significant ($p>0.05$)

Figure 18 Amount of non-differential misclassification required such that treatment effect (relative risk=0.67) is no longer significant at 5% level for various sample sizes and overall event rates



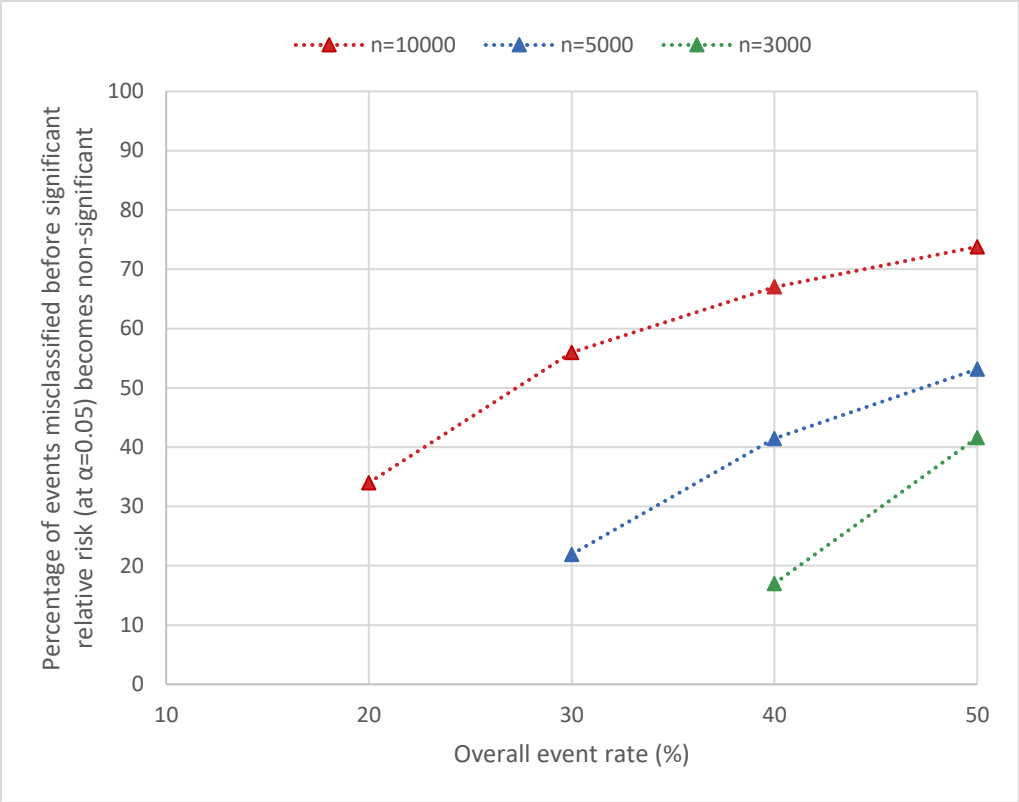
n refers to hypothetical trial sample size

Figure 19 Amount of non-differential misclassification required such that treatment effect (relative risk=0.82) is no longer significant at 5% level for various sample sizes and overall event rates



Missing scenarios are due to the initial treatment effect before misclassification being non-significant (p>0.05). n refers to hypothetical trial sample size

Figure 20 Amount of non-differential misclassification required such that treatment effect (relative risk=0.90) is no longer significant at 5% level for various sample sizes and overall event rates



Missing scenarios are due to the initial treatment effect before misclassification being non-significant (p>0.05). n refers to hypothetical trial sample size

5.4. Discussion

In this simulation study based on seven distinct stroke trial populations it has been shown that only a small amount of differential misclassification was needed before central adjudication would have altered the estimated treatment effect. Larger trials appeared to be most vulnerable to this bias, in part due to their larger sample size being able to detect a smaller difference in treatment effect. However, for blinded studies where differential misclassification should not occur, an implausible amount of random error is required to alter trial conclusions.

Whilst ordinal outcomes could be misclassified by more than one level (i.e. mRS of 1 to 3), it can be argued that even this would be less severe than misclassification of a binary event (0 to 1 or vice versa). Therefore, the results from binary and ordinal outcomes should not be compared. Overall, I found that a relatively small amount of differential misclassification was needed to alter the estimated treatment effect. This suggests that central adjudication is important to control for differential misclassification in randomised trials. However, three of the five trials included had blinded outcome assessment, so the plausibility of this amount of differential misclassification happening in practice to these studies is far less than the unblinded trials. In the systematic review described in Chapter 3[194] there was no indication of any detection bias through differential misclassification, so even the small proportion needed before the treatment effect changes may be a rare occurrence in trials. One reason for this finding in this review could be

due to 9 (60%) of the included studies having the site investigators blind to treatment allocation and the majority of the studies had stroke as their primary outcome, which is well defined and accurately measured[199]. In the review, there was not a significant interaction between blinding status and RTE, but this may have been due to the reviews small sample size. A Cochrane review[60] that included 47 trials which adjudicated subjective binary events did find an interaction between blinding status of the site investigators and the ratio of odds ratios (RORs), with the suggestion that unblinded site investigators exaggerate treatment effect estimates (two trials, ROR=0.76, 95% C.I: [0.46, 1.12]). Furthermore, unblinded site assessors have been shown to exaggerate treatment effect estimates in multiple studies by Hróbjartsson[30, 33, 34]. Thus, differential misclassification is a real possibility in medical research, and adjudication can control for this.

However, for studies with adequate blinding, central adjudication would not be expected to control for differential misclassification, and instead only reduce random noise around the effect of interest. The proportion of events needed to be misclassified before a significant treatment effect becomes non-significant increases with trial size, overall event rate and strength of treatment effect. This can also vary with method of adjudication[56], but this is not something explored in this study. This chapter has shown for a trial with a binary outcome that a large amount of non-differential misclassification is necessary before even a modest treatment effect is missed. For the five stroke trials included in the first part of this study, the largest agreement for a trial with a binary outcome

was 98.8%. Far higher disagreement would have been needed before central adjudication ensures that a moderate and statistically significant treatment effect does not become non-significant through random error. In a previous simulation study that explored central adjudication of stroke type in a large stroke trial with blinded outcome assessment[52], the agreement between the adjudicators and site investigators was 98% and kappa had to reduce from 0.92 to 0.46 before a true subgroup effect by stroke type was missed. This amount of random error is not plausible for many trial settings, especially if the event rate in that population is near to 50%. Other studies investigating adjudication in stroke trials found agreement between adjudicators and site investigators of 91% for all stroke[98], and 90% for stroke[51]. Thus, for large blinded trials, central adjudication is an unnecessary expenditure to control for non-differential misclassification.

One limitation of the study in this chapter is that only one clinical area has been explored, and the high level of agreement found across the included studies may be lower in other clinical areas or for different outcomes. However, within this single clinical area a variety of stroke trials were chosen that represent acute stroke, primary and secondary prevention studies as well as including the majority of common primary outcomes in these studies. Another limitation is that non-differential misclassification was only explored for binary outcomes. The justification for this is that the majority of stroke trials included in this review[194] had a binary primary outcome. Furthermore, it is possible that trials with ordinal outcomes would need greater misclassification

than those with binary outcomes, due to the ordinal scale the outcome is measured on. This simulation study only computed 1000 simulations for each study, due to the complex nature of the simulation program and multiple actions for each run of the simulation. The precision of the simulation estimates may have been improved had a larger number of simulations been run (e.g. 5000), but this was not possible due to the complex nature of the simulation program.

To conclude, this chapter found that central adjudication has the potential to be important for stroke trials with inadequate blinding through its control of differential misclassification. However, for studies that have blinded outcome assessment, central adjudication is an unneeded expense, both in cost and time, for randomised trials.

The next chapter provides an estimation of the cost of the central adjudication process in stroke trials. Data required to estimate the cost of central adjudication was obtained from nine of the fifteen trials included in the systematic review, and this information is used alongside a measure of benefit of the adjudication process, number of events corrected by adjudication, to establish cost-benefit of central adjudication.

Chapter 6 – Cost-benefit of central adjudication in randomised stroke trials

Publications contributing to this chapter:

1. PJ Godolphin, PM Bath, A Algra, E Berge, J Chalmers, M Eliasziw, GJ Hankey, N Hosomi, A Ranta, C Weimar, L Woodhouse, AA Montgomery. Cost-benefit of outcome adjudication in nine randomised stroke trials. [Under Review].

2. PJ Godolphin, PM Bath, AA Montgomery on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration. Outcome assessment by central adjudicators versus site investigators in randomised stroke trials. 2019. Poster presentation at 40th International Society for Clinical Biostatisticians Conference.

Abstract

Background: Central adjudication of outcomes is common for randomised stroke trials. However, adjudication is potentially a time-consuming and expensive process. Few studies have estimated the cost of central adjudication in randomised trials. In this chapter, the cost of the adjudication process in stroke trials is estimated.

Methods: This chapter estimated the cost of adjudicating the primary outcome in nine randomised stroke trials. The costs included adjudicators' time, direct payments to adjudicators, and co-ordinating centre costs (e.g. uploading cranial scans and general set-up costs). All costs were taken from a societal perspective. The number of events corrected after adjudication was the measure of benefit in this study. The cost per correct event was calculated for each trial and in total.

Results: The primary outcome in all nine trials was either stroke or a composite that included stroke. In total, the adjudication process associated with this primary outcome cost in excess of £100,000 for a third of the trials (3/9). Mean cost per event corrected by adjudication was £2295.10 (standard deviation: £1482.42).

Conclusions: Central adjudication is a time-consuming and potentially costly process. These costs need to be considered when designing a trial and should be evaluated alongside the potential benefits adjudication brings to determine whether they outweigh this expense.

6.1 Introduction

The previous chapters have demonstrated that central adjudication does not alter estimated treatment effects in practice, but that it does have the potential to be a beneficial process when blinding for outcome assessment is not possible. However, the cost of adjudication must be taken into account with any benefits that adjudication brings, including those benefits that are hard to measure (e.g. providing reassurance). This chapter estimates the cost of the adjudication process in nine of the stroke trials included in the systematic review detailed in Chapter 3.

The cost of adjudication can be measured in many ways. The cost could be interpreted simply as the direct cost paid to the adjudicator, which is often a menial amount (£10-15 per adjudication). However, this does not take into account the actual time it takes the adjudicator to carry out their assessment. Furthermore, preparing data for the adjudicators, blinding and obtaining source information for them, and general set-up costs are other expenses with the adjudication process that are borne by the co-ordinating centre (often a clinical trials unit). Thus, the cost of the adjudication process can be made up from (1) direct costs paid to the adjudicators; (2) time costs of the adjudicators; and (3) co-ordinating centre costs.

The aims of this chapter are to (1) provide an estimate of the cost of adjudicating the primary outcome in randomised stroke trials; and (2) present this cost in terms of the number of events corrected by adjudication.

6.2 Methods

6.2.1 Data Collection

Authors of all fifteen stroke trials included in the systematic review described in Chapter 3[194] were invited to provide data on the cost of outcome adjudication in an online questionnaire designed on Google Forms (see Appendix K). The data collected included the time that was required to adjudicate events, an approximation of the adjudicators' annual salary in today's costs and information on how disagreements were handled (if any).

6.2.2 Benefit of Adjudication

In this study, the measure of benefit of the adjudication process was taken as the number of events corrected after adjudication. This outcome is not influenced by whether the trial results are altered or whether the ratio of treatment effects deviates from one, but simply collects gives an idea of the extent that adjudicators change the trial outcome that was determined by site investigators. This outcome was collected in the previous systematic review described in Chapter 3[194].

6.2.3 Cost of Adjudication

All costs were taken from a societal perspective and were reported in various currencies (Australian dollars, Euros, Pounds Sterling and New Zealand Dollars). If costs were not entered as pounds sterling, then all other currencies were converted into pounds sterling on 16th May 2019.

The hourly salary for all adjudicators was calculated assuming that adjudicators worked 1680 hours a year (40 hours a week for 42 weeks

of the year). For two trials, the annual salary of the adjudicators was unknown. Thus, for these two studies (NASCET & J-STARS), hourly salary was estimated as the approximate hourly salary for a new consultant in the UK (£47.62 an hour). The total time taken by adjudicators was determined using the amount of time it took for each adjudication, the number of events adjudicated and whether events were adjudicated multiple times. In addition to staff salary costs assumed to be borne by employers rather than study funders, it was also assumed that adjudicators were directly paid £10 per event from study funds. This is not paid in all studies; however, it enables a comparison between studies and where paid, is typical of the amount in UK studies. The costs involved in resolving disagreements was based on staff time only and did not include additional £10 payments to the adjudicators.

For each study, a fixed set-up cost was assumed that took into account creating a website for data sharing (brain scans, video footage etc.) and to enable adjudicators to submit their final assessments. This was estimated to require 40 hours of work for a database programmer (approximate cost £650). In addition, it was assumed that each adjudicated event required four hours of administrative time at the co-ordinating centre (approximate cost £40 per adjudicated event).

6.2.4 Statistical Analysis

Continuous variables were summarised with mean and standard deviation, or median and interquartile range. Categorical variables were

described with frequency counts and percentages. The cost per corrected event was determined for each trial and for all trials.

6.3 Results

Authors from nine studies, comprising 25,436 participants, completed the questionnaire and these nine trials are included in this study[178-183, 185, 188, 191, 192] (Table 45). All outcomes adjudicated were either stroke or a composite including stroke. The time taken per adjudication ranged from five minutes to two hours. However, the trial with the lowest time had three adjudicators assess each event (Table 46). All other trials assessed each event either once or twice. The total time spent adjudicating ranged from 5.7 hours to 1744 hours. The trial with highest adjudicator salary cost was around twice that of the lowest.

Three of the trials (33%) had no disagreements (Table 47). For trials that had disagreements between adjudicators, the cost and time associated with this was far less than that associated with the original adjudication (Table 48). The coordinating centre cost was greater than the cumulative adjudicator and disagreement cost for three of the nine trials, and was within 10% of this cumulative cost for another four trials.

Two studies had no events corrected after adjudication, although the total cost of the adjudication process for both of these was less than £4,000. Three trials had a total estimated cost of the adjudication process in excess of £100,000. On average, across the seven trials with corrected events, adjudication cost £2,295.10 per event corrected (standard deviation: £1482.42).

Table 45 Characteristics of included trials

	Included (n=9)
Year of main trial publication	
1990-2000	2 (22%)
2001-2005	1 (11%)
2006-2010	2 (22%)
2011-2015	2 (22%)
2016-2018	2 (22%)
Study design	
Parallel	8 (89%)
Factorial	1 (11%)
Type of trial	
Primary prevention	1 (11%)
Secondary prevention	6 (67%)
Acute stroke	2 (22%)
Participants randomised	
Mean (SD)	2826 (2744)
Median [25th, 75th centile]	2739 [449, 3096]
Min, Max	129, 8164
No. of sites	
Mean (SD)	86 (49)
Median [25th, 75th centile]	79 [50, 123]
Min, Max	17, 172
Intervention	
Drug	6 (67%)
Surgery/procedure	2 (22%)
Other	1 (11%)
Comparator	
Placebo	2 (22%)
Standard care	6 (67%)
Surgery/procedure	1 (11%)
Primary outcome	
Stroke	6 (67%)
Composite including stroke	3 (33%)
Blinding status of site investigators	
Blind to treatment allocation	4 (44%)
Not blind to treatment allocation	5 (56%)
Method used to select cases to adjudicate	
Only cases identified by unblinded site investigators are adjudicated	4 (44%)
Only cases identified by blinded site investigators are adjudicated	4 (44%)
All participants, or all suspected cases identified by blinded site investigators are adjudicated	1 (11%)

Table 46 Time taken for adjudication and dealing with disagreements

Trial name	Events adjudicated	Number of adjudications per event	Time per adjudication (minutes)	Total time for adjudication (hours)	Adjudicator hourly salary	Cost of adjudication*
CABACS	34	Once	10	5.7	£67.62	£723.18
ESPRIT	440	Three	5	110	£62.40	£20064
FASTEST	24	Once	45	18	£70.16	£1502.88
HAEST	93	Once	15	23.25	£47.62	£2037.17
J-STARS	239	Once	30	119.5	£47.62	£8080.59
NASCET*	436	Twice	120	1744	£47.62	£91769.28
PROGRESS	992	Once	15	248	£48.22	£21878.56
TARDIS	1656	Twice	10	552	£49.40	£60388.80
VITATOPS	1260	Once	20	420	£96.40	£53088

Calculated as: Cost of adjudication = (Total time for adjudication)*(Adjudicator hourly salary) + [(Events adjudicated)*Number of adjudication per event)*(£10)]
*NASCET was the only trial where adjudicators could either increase or reduce the overall number of cases

Table 47 Cost of dealing with disagreements between adjudicators

Trial name	Number of disagreements	Number of adjudicators involved in resolving each disagreement	Time taken per adjudicator per disagreement (minutes)	Total time for disagreements (hours)	Hourly salary of adjudicators dealing with disagreements	Cost to resolve disagreements
CABACS	0	NA	NA	NA	NA	£0
ESPRIT	22	1	5	1.83	£62.40	£114.40
FASTEST	6	2	45	9	£70.16	£631.44
HAEST	0	NA	NA	NA	NA	£0
J-STARS	10	5	60	50	£47.62	£2381
NASCET*	27	1	60	27	£47.62	£1285.74
PROGRESS	0	NA	NA	NA	NA	£0
TARDIS	169	3	20	169	£49.40	£8348.60
VITATOPS	13	3	20	13	£96.40	£1253.20

*NASCET was the only trial where adjudicators could either increase or reduce the overall number of cases

Table 48 Cost of adjudication per corrected event by included trial

Trial name	Coordinating centre cost*	Adjudication cost	Disagreement cost	Total cost	Number of events corrected after adjudication	Cost per corrected event
CABACS	£2010	£723.18	£0	£2733.18	0	NA
ESPRIT	£18250	£20064	£114.40	£38428.40	51	£753.50
FASTEST	£1610	£1502.88	£631.44	£3744.32	0	NA
HAEST	£4370	£2037.17	£0	£6407.17	2	£3203.58
J-STARS	£10210	£8080.59	£2381	£20671.59	10	£2067.16
NASCET††	£18090	£91769.28	£1285.74	£111145.02	27	£4116.48
PROGRESS	£40330	£21878.56	£0	£62208.56	98	£634.78
TARDIS	£66890	£60388.80	£8348.60	£135627.40	34	£3989.04
VITATOPS	£51050	£53088	£1253.20	£105391.20	81	£1301.13
Mean (SD)†						£2295.10 (£1482.42)
Median†						£2067.16
Min†						£634.78
Max†						£4116.48

*Coordinating centre cost includes £650 set-up cost and £40 cost per event

†Summary statistics do not include CABACS or FASTEST

††NASCET was the only trial where adjudicators could either increase or reduce the overall number of cases

6.4 Discussion

In this retrospective analysis described in this chapter, the cost of the outcome adjudication process has been estimated for nine randomised trials. On average, this study showed that adjudication cost approximately £2,300 per corrected event, with the total cost of adjudication exceeding £100,000 for three trials. Another study estimated the co-ordinating centre cost for the adjudication process in a thromboprophylaxis trial to be approximately £55 per adjudicated event[59] compared to the assumption used here of £40. Thus, it is possible that the total cost of the adjudication process has been marginally underestimated in this cost-benefit analysis.

There are several limitations of the work presented in this chapter. Firstly, the data described in this chapter was collected in a retrospective fashion and so there is the possibility that they contain inaccuracies. Collecting cost data prospectively would have been preferable, albeit impossible for this set of trials as many of these were completed well before the start of this research programme. Secondly, it would have been interesting to compare the cost-benefit of trials that could increase as well as decrease the number of cases (e.g. where all participants were adjudicated). However, only one trial, NASCET, completed cost data and used an adjudication methods that permitted this. Thus, this comparison was not performed due to a lack of studies.

In summary, central adjudication has been shown to control for differential misclassification and is important for trials where blinding is

inadequate or impossible. However, as shown in this chapter, it can be costly, and these costs need to be estimated when designing a trial to evaluate whether the benefits of central adjudication outweigh the potential expense.

The following chapter describes central adjudication of safety outcomes, rather than the primary outcome, in a single stroke trial. The case study chosen was the Efficacy of Nitric Oxide in Stroke Trial, in which serious adverse events were adjudicated. The purpose of Chapter 7 is to explore whether the results seen in the preceding chapters are similar for adjudication of safety outcomes as well.

Chapter 7 – Central adjudication of Serious Adverse Events in the Efficacy of Nitric Oxide in Stroke (ENOS) Trial

Publications contributing to this chapter:

1. PJ Godolphin, AA Montgomery, LJ Woodhouse, D Bereczki, E Berge, R Collins, E Diez-Tejedor, J Gommans, KR Lees, S Ozturk, S Phillips, S Pocock, K Prasad, S Szatmari, Y Wang, PM Bath, N Sprigg. Central adjudication of serious adverse events did not affect trial's safety results: Data from the Efficacy of Nitric Oxide in Stroke (ENOS) trial. PLoS ONE 2018, 13(11):e0208142. https://doi.org/10.1371/journal.pone.0208142

2. PJ Godolphin, L Woodhouse, D Bereczki, AA Montgomery, PM Bath, N Sprigg. Assessing the impact of independent adjudication of serious adverse events on the safety results of the efficacy of nitric oxide in stroke (ENOS) trial. Trials 2017, 18(Suppl 1):P166. Poster presentation at 4th International Clinical Trials Methodology Conference.

Abstract

Background: Central adjudication of serious adverse events (SAEs) is common for open-label trials where outcome assessment is at risk of bias. This chapter explores the effect of central adjudication of SAEs on the safety results of the Efficacy of Nitric Oxide in Stroke (ENOS) Trial.

Methods: ENOS assigned patients with acute stroke to receive either transdermal glyceryl trinitrate (GTN) or no GTN and to Stop or Continue previous antihypertensive treatment. SAEs were reported by site investigators who were not blind to treatment allocation. Central adjudicators, blind to treatment allocation, reviewed the investigators reports and confirmed or re-categorised the classification of event, likely causality, diagnosis and expectedness of event.

Results: Of 4011 patients enrolled in ENOS, 1473 SAEs were reported by site investigators; this was reduced to 1444 after the review by adjudicators, with 29 re-classified as not an SAE. There was fair agreement between investigators and adjudicators on likely causality, with 808 agreements and 644 disagreements (56% crude agreement, weighted kappa, $\kappa = 0.31$). Repeating the main trial safety analysis with investigator reported events showed that adjudication had no effect on the main trial safety conclusions.

Conclusions: In a large trial, with many SAEs reported, central adjudication of these events did not affect trial conclusions. This suggests that adjudication of SAEs may not be necessary in a clinical trial where the intervention already has a well-established safety profile.

7.1 Introduction

The previous chapters have explored the effect of central adjudication of the primary efficacy outcome on the primary results in randomised stroke trials. As with efficacy outcomes, central adjudicators also commonly assess safety outcomes in randomised trials. This chapter will apply similar methodology utilised in previous chapters to central adjudication of serious adverse events (SAEs) in a single case study.

The case study used for this investigation is the Efficacy of Nitric Oxide in Stroke (ENOS) trial, a large acute stroke trial. In ENOS, central adjudicators assessed SAEs and reported on a number of different domains, including diagnosis of event, classification of event and whether the event was likely to be related to the treatment received. This allows an investigation into the agreement between adjudicators and site investigators for all these domains, as well as permitting an evaluation of the effect of adjudication on the safety results of the trial. This evaluation follows a similar process to that used in Chapter 3, through a re-analysis of the trial using investigator reported events and comparing these results to those obtained using adjudicator reported events with a ratio of odds ratios (RORs).

The aims of this chapter are to (1) describe the level of agreement between central adjudicators and site investigators; (2) assess the effect of central adjudication of SAEs on the primary safety analysis of the trial.

7.2 Methods

7.2.1 Efficacy of Nitric Oxide in Stroke (ENOS) Trial

7.2.1.1 Summary of ENOS Methods

The Efficacy of Nitric Oxide in Stroke (ENOS) trial was an international randomised single-blind masked-endpoint partial-factorial trial. Patients admitted to hospital with acute stroke (within 48 hours) and raised blood pressure (between 140-220 mmHg) were randomised to seven days treatment with transdermal glyceryl trinitrate (GTN) or no GTN. A subset of patients who were already taking anti-hypertensive medication before trial entry were further assigned at random to continue or stop their medication. Patients were blind to whether they received GTN or no GTN through the use of a gauze dressing, which covered a similar part of the body if they were randomised to the no GTN group. It was not possible to mask patients for the decision to continue or stop their antihypertensive medication.

7.2.1.2 Summary of ENOS Results

The trial recruited 4011 participants from 173 sites, across 23 countries on five continents between July 2001 and October 2013. 2000 participants were assigned to receive GTN and 2011 were allocated to the no GTN group. Just over half the patients (2097, 52%) were taking anti-hypertensive medication before the trial and 1053 were allocated to the Continue and 1044 to the Stop arms of the trial. The trials safety analysis compared the overall rate of SAEs between treatment arms, and found no difference between GTN and no GTN and Continue and

Stop. The protocol, statistical analysis plan, and main results for ENOS are described in detail elsewhere.[200-202]

7.2.1.3 Adjudication of Investigator reported Serious Adverse Events in ENOS

Serious adverse events were defined in the protocol as adverse events that: were fatal, caused disability, were life threatening, led to hospital admission or prolonged discharge in a hospitalised patient, and/or resulted in birth defects in children born to trial participants. Excepting a small number of selected adverse events (such as headache), which were recorded but not adjudicated, non-serious adverse events were not recorded due to the established nature of the trial interventions. SAEs were recorded by site investigators using a web-based SAE form within 24 hours of the investigator being aware of the event. Site investigators were trained in SAE reporting before site initiation. Once five participants had been recruited by any participating trial site, there was a further site visit and ongoing data monitoring. Additional training was provided in yearly investigator meetings and regular trial newsletters also included hints and tips as well as frequently answered questions for SAE reporting.

It was not possible to blind site investigators to treatment allocation as they knew whether patients had received the GTN patch or not, and if patients were randomised to continue or stop their anti-hypertensive medication. The site investigator report included event classification, event diagnosis and evidence used to determine diagnosis, expectedness of event, and likely causality. Likely causality was

reported in five categories: *Definitely not*, *Unlikely* (to be), *Possibly*, *Probably* and *Definitely* related to treatment.

Two central independent expert adjudicators, who were blinded to treatment allocation, evaluated the investigator-reported SAEs. Each event was reviewed by a single adjudicator. The central adjudicators had access to the web-based SAE form provided by the site investigator, a list of known adverse reactions and evidence provided by investigators including clinical notes, cranial scans, biochemistry and bacteriology data. Adjudicators were able to request further information from the specific trial sites if this was required to carry out the assessment of the SAE. Adjudicators confirmed or re-categorised investigators' decisions on classification of event, event diagnosis, expectedness of event and likely causality. Any uncertainty was clarified by discussion with the chief investigator blinded to treatment allocation. The central adjudicator's decision was treated as the gold standard for this analysis and was the decision used in all ENOS trial analyses.

7.2.2 Statistical Methods

Continuous variables were summarised using mean and standard deviation, or median and interquartile range. Categorical variables were described using number (%). Mann-Whitney U tests and Chi-square tests were used to determine imbalance between treatment groups for investigator reported SAEs.

Unweighted kappa was used to estimate agreement between site investigators and central adjudicators on event classification and diagnosis. This agreement was determined overall, and also for UK sites and non-UK sites. Agreement between site investigators and central adjudicators on likely causality of SAEs was calculated using weighted kappa with symmetric weights. Weights were assigned as 1 if adjudicators and investigators agreed, as 0.75 if adjudicators and investigators causalities differed by one, as 0.5 if adjudicators and investigators causalities differed by two and 0 if adjudicators and investigators causalities differed by more than two. The likely causality categories were dichotomised into *Definitely not, Unlikely* and *Possibly* versus *Probably* and *Definitely*, and also into *Definitely not* and *Unlikely* versus *Possibly, Probably* and *Definitely*, with kappa used to estimate agreement for both dichotomisations.

The time taken to adjudicate SAEs was estimated as the difference between the time the adjudicator first assessed the event and the time the event was completed. If the adjudicator required further information, then the time taken to retrieve this information was included in the calculation of the time taken to adjudicate.

The number of patients with any SAE and the number of patients for each SAE diagnosis were compared by treatment arm using chi-squared tests for both site-investigator- and central-adjudicator-reported events. Binary logistic regression models were fitted for each SAE classification for both site-investigator- and central-adjudicator- reported events. RORs were calculated for each SAE classification. The ROR

was determined as the odds ratio using events classified by site investigators to the odds ratio using events classified by central adjudicators. Thus, a ROR > 1 indicated that central adjudication resulted in more events (worse outcome).

7.3 Results

Of 4011 patients enrolled into ENOS, 1031 (26%) were recorded as having had at least one SAE by site investigators. In total, site investigators reported 1473 SAEs, and of these, 1444 (98%) were classified as an SAE after assessment by central adjudicators.

The number of SAEs were similar between the GTN versus No GTN arms. Events were more likely to be moderate or mild than severe if a patient had been randomised to receive the GTN patch (Table 49). Investigators used clinical diagnostic evidence for roughly three-quarters of the SAEs (74%), with radiological (47%), biochemical (30%), haematological (26%) and ECG (23%) diagnostic evidence also used regularly. Adjudicators were certain of SAE diagnosis for the majority of investigator-reported events (76%), but indicated that their diagnosis was only possibly correct for 68 (5%) of the 1473 investigator-reported SAEs.

7.3.1 Agreement on SAE diagnosis

Agreement was excellent between central adjudicators and site investigators for the diagnosis of SAEs (Figure 21, Table 50, crude agreement = 88%, unweighted kappa $\kappa = 0.85$). The highest number of disagreements, ignoring the miscellaneous category, was between cardiovascular (25/337, 7%) and respiratory events (35/294, 12%). Overall agreement for the diagnosis of SAEs was marginally better for UK sites ($\kappa = 0.85$) compared to non-UK sites ($\kappa = 0.83$).

Table 49 Characteristics of Investigator reported serious adverse events and certainty of adjudicator's diagnosis

	All	GTN	No GTN	p-value
Patients	4011	2000	2011	
With SAE	1031	522	509	
SAEs	1473	781	692	
SAEs per patient				
Mean (SD)	1.5 (1.0)	1.6 (1.1)	1.4 (0.8)	0.063*
Median [IQR]	1 [1, 2]	1 [1, 2]	1 [1, 2]	
Min, Max	1, 9	1, 9	1, 6	
Timing of event				
Before treatment	14 (1%)	6 (1%)	8 (1%)	0.61†
During treatment	642 (44%)	335 (43%)	307 (44%)	
After treatment	817 (55%)	440 (56%)	377 (54%)	
Time from randomisation to event (days)				
Mean (SD)	22.5 (27.1)	22.7 (28.6)	22.1 (25.5)	0.97*
Median [IQR]	11 [3, 34]	12 [3, 33]	10 [4, 36]	
Min, Max	0, 383	0, 383	0, 148	
Diagnostic evidence used by investigators‡				
Pathological	94 (6%)	45 (6%)	49 (7%)	0.74†
Radiological	685 (47%)	376 (48%)	309 (45%)	
ECG	336 (23%)	178 (23%)	158 (23%)	
Bacteriology	168 (11%)	93 (12%)	75 (11%)	
Biochemistry	443 (30%)	232 (30%)	211 (30%)	
Haematology	383 (26%)	201 (26%)	182 (26%)	
Clinical	1095 (74%)	593 (76%)	502 (73%)	
Other	234 (16%)	115 (15%)	119 (17%)	
Intensity of event				
Mild	214 (15%)	133 (17%)	81 (12%)	0.003†
Moderate	448 (30%)	246 (31%)	202 (29%)	
Severe	810 (55%)	402 (51%)	408 (69%)	
Clinical outcome				
Recovered	479 (33%)	268 (34%)	211 (30%)	0.006†
Not yet recovered	542 (37%)	302 (39%)	240 (35%)	
Died	451 (31%)	211 (27%)	240 (35%)	
Certainty of adjudicators event diagnosis				
Definite	1126 (76%)	593 (76%)	533 (77%)	0.062†
Probable	276 (19%)	153 (20%)	123 (18%)	
Possible	68 (5%)	34 (4%)	34 (5%)	

Data are N(%) unless otherwise indicated.

GTN = Glyceryl trinitrate, ECG = echocardiogram

*p-value from Mann-Whitney U test

†p-value from chi-square test

‡Categories not mutually exclusive

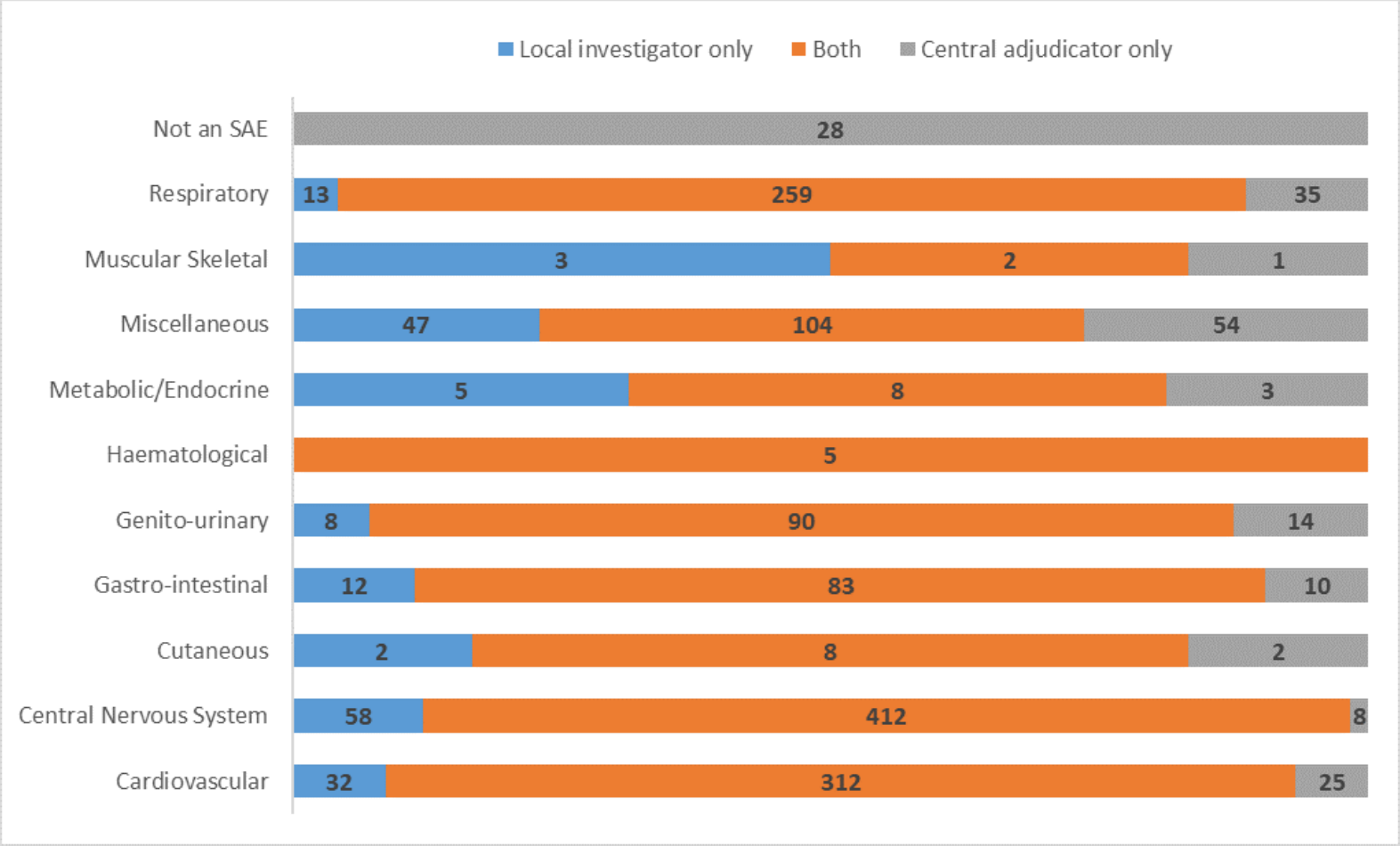
Table 50 Agreement on Diagnosis of serious adverse events between central adjudicators and site investigators

	Central Adjudicators											
Site Investigators	Cardiovascular	CNS	Cutaneous	Gastro-intestinal	Genito-urinary	Haematological	Metabolic / Endocrine	Miscellaneous	Muscular Skeletal	Respiratory	Not a SAE	**Total**
Cardiovascular	312	3	0	0	0	0	0	14	0	7	8	344
CNS	8	412	0	3	3	0	1	15	0	18	10	470
Cutaneous	0	0	8	0	0	0	0	2	0	0	0	10
Gastro-intestinal	0	0	0	83	0	0	0	7	0	1	4	95
Genito-urinary	0	0	1	0	90	0	0	5	0	1	1	98
Haematological	0	0	0	0	0	5	0	0	0	0	0	5
Metabolic / Endocrine	0	0	0	0	1	0	8	2	0	1	1	13
Miscellaneous	12	5	1	7	10	0	2	104	1	7	2	151
Muscular Skeletal	0	0	0	0	0	0	0	2	2	0	1	5
Respiratory	5	0	0	0	0	0	0	7	0	259	1	272
Not an SAE	0	0	0	0	0	0	0	0	0	0	0	0
Total	337	420	10	93	104	5	11	158	3	294	28	1463
Disagreements (%)	25 (7%)	8 (2%)	2 (20%)	10 (11%)	14 (13%)	0 (0%)	3 (27%)	54 (34%)	1 (33%)	35 (12%)	28 (100%)	180 (12%)

CNS = Central Nervous System. Crude agreement = 1283/1463 = 88%, unweighted kappa = 0.85 (0.82, 0.87)

Site Investigators did not record diagnosis of SAE for 10 events and so could these were not included in this analysis

Figure 21 Agreement on serious adverse event diagnosis between central adjudicators and local investigators



7.3.2 Agreement on SAE classification

Overall agreement on event classification was good (Table 51, crude agreement = 82%, unweighted kappa, κ = 0.78). The pulmonary embolism and other cardiovascular event subcategories had very high agreement, whilst the largest proportion of disagreements occurred in the extension of initial stroke, sudden cardiac death and pneumonia subcategories.

Within the cardiovascular diagnosis category, there was excellent agreement (Table 52, 337 events, crude agreement = 84%, unweighted kappa, κ = 0.82). Similarly, agreement within the central nervous system diagnosis category was good (Table 53, 420 events, crude agreement = 76%, unweighted kappa, κ = 0.72). Within this category, investigators identified 57 (79%) of the 72 complication of initial stroke events, 75 (87%) of the 86 recurrent stroke events and 74 (73%) of the 102 extension of initial stroke events. A large number of the incorrectly identified events for these three sub-categories should have been classified in one of the other two sub-categories. For complication, recurrence and extension this crossover was 47%, 82% and 79% respectively.

Agreement was only fair within the respiratory category (Table 54, 294 events, crude agreement = 73%, unweighted kappa, κ = 0.35), with only 192 of the 262 pneumonia events identified by investigators. 16 of these events were incorrectly classified as complication of initial stroke and 37 were classified by investigators as a chest infection.

Table 51 Agreement on classification of serious adverse events between central adjudicators and site investigators

	Central Adjudicators											
Site Investigators	Complication of initial stroke	Extension of initial stroke	Symptomatic intracerebral haemorrhage	Recurrent stroke	Myocardial infarction	Sudden cardiac death	Other cardiovascular event	Pulmonary embolism	Pneumonia	Other event	Not an SAE	**Total**
Complication of initial stroke	57	17	2	2	0	1	2	0	16	27	0	124
Extension of initial stroke	2	74	4	7	0	0	1	0	1	4	1	94
Symptomatic intracerebral haemorrhage	3	1	37	0	0	0	0	0	0	3	0	44
Recurrent stroke	5	5	3	75	0	0	0	1	0	5	0	94
Myocardial infarction	0	0	0	0	38	5	1	1	0	1	0	46
Sudden cardiac death	1	0	0	0	0	4	0	0	1	2	0	8
Other cardiovascular event	1	1	0	0	4	2	213	2	4	12	8	247
Pulmonary embolism	0	0	0	0	1	0	0	41	0	1	0	43
Pneumonia	0	0	0	0	0	0	3	0	192	4	1	200
Other event	3	4	1	2	3	0	13	1	46	472	18	563
Not an SAE	0	0	0	0	0	0	0	0	0	0	0	0
Total	72	102	47	86	46	12	233	46	260	531	28	1463
Disagreements (%)	15 (21%)	28 (27%)	10 (21%)	11 (13%)	8 (17%)	8 (67%)	20 (9%)	5 (11%)	68 (26%)	59 (11%)	28 (100%)	260 (18%)

Crude agreement = 1203/1463 = 82%, unweighted kappa = 0.78 (0.75, 0.80)

Table 52 Agreement on classification of cardiovascular serious adverse events between central adjudicators and site investigators

	Central Adjudicators									
Site Investigators	Complication of initial stroke	Extension of initial stroke	Recurrent stroke	Myocardial infarction	Sudden cardiac death	Other cardiovascular event	Pulmonary embolism	Pneumonia	Other event	**Total**
Complication of initial stroke	0	0	0	0	1	2	0	0	0	3
Extension of initial stroke	0	0	0	0	0	1	0	0	0	1
Recurrent stroke	0	0	0	0	0	0	1	0	0	1
Myocardial infarction	0	0	0	38	5	1	1	0	0	45
Sudden cardiac death	0	0	0	0	4	0	0	0	0	4
Other cardiovascular event	0	0	0	4	2	213	2	0	0	221
Pulmonary embolism	0	0	0	1	0	0	41	0	0	42
Pneumonia	0	0	0	0	0	3	0	0	0	3
Other event	0	0	0	3	0	13	1	0	0	17
Total	0	0	0	46	12	233	46	0	0	337
Disagreements (%)	0 (-)	0 (-)	0 (-)	8 (17%)	8 (67%)	20 (9%)	5 (11%)	0 (-)	0 (-)	41 (12%)

Crude agreement = 296/337 = 88%, unweighted kappa = 0.76 (0.69, 0.83)

Table 53 Agreement on classification of central nervous system serious adverse events between central adjudicators and site investigators

	Central Adjudicators							
Site Investigators	Complication of initial stroke	Extension of initial stroke	Symptomatic intracerebral haemorrhage	Recurrent stroke	Sudden cardiac death	Other cardiovascular event	Other event	**Total**
Complication of initial stroke	57	17	2	2	0	0	14	92
Extension of initial stroke	2	74	4	7	0	0	4	91
Symptomatic intracerebral haemorrhage	3	1	37	0	0	0	2	43
Recurrent stroke	5	5	3	75	0	0	3	91
Sudden cardiac death	1	0	0	0	0	0	0	1
Other cardiovascular event	1	1	0	0	0	0	0	2
Other event	3	4	1	2	0	0	90	100
Total	72	102	47	86	0	0	113	420
Disagreements (%)	15 (21%)	28 (27%)	10 (21%)	11 (13%)	0 (-)	0 (-)	23 (20%)	87 (21%)

Crude agreement = 333/420 = 79%, unweighted kappa = 0.74 (0.69, 0.79)

Table 54 Agreement on classification of respiratory serious adverse events between central adjudicators and site investigators

	Central Adjudicators							
Site Investigators	Complication of initial stroke	Extension of initial stroke	Sudden cardiac death	Other cardiovascular event	Pulmonary embolism	Pneumonia	Other event	**Total**
Complication of initial stroke	0	0	0	0	0	16	0	16
Extension of initial stroke	0	0	0	0	0	1	0	1
Sudden cardiac death	0	0	0	0	0	1	0	1
Other cardiovascular event	0	0	0	0	0	4	1	5
Pulmonary embolism	0	0	0	0	0	0	1	1
Pneumonia	0	0	0	0	0	192	1	193
Other event	0	0	0	0	0	46	31	77
Total	0	0	0	0	0	260	34	294
Disagreements (%)	0 (-)	0 (-)	0 (-)	0 (-)	0 (-)	68 (26%)	3 (9%)	71 (24%)

Crude agreement = 223/294 = 76%, unweighted kappa = 0.38 (0.30, 0.46)

7.3.3 Agreement on likely causality of SAEs

There was fair agreement on likely causality of SAEs between central adjudicators and site investigators, with disagreement for 644/1452 (44%) events (Table 55, Figure 22, weighted kappa, κ = 0.31). When categories were dichotomised into *Definitely not*/*Unlikely/Possibly* versus *Probably*/*Definitely* there was near perfect agreement (Table 56, crude agreement = 99%, kappa κ = 0.54). If categories instead were dichotomised into *Definitely not*/*Unlikely* versus *Possibly*/*Probably*/*Definitely*, then crude agreement remained high, but agreement over that expected by chance was reduced (Table 57, crude agreement = 90%, kappa κ = 0.33).

Table 55 Agreement on likely causality of serious adverse events between central adjudicators and site investigators

	Central Adjudicators					
Site Investigators	Definitely not	Unlikely	Possibly	Probably	Definitely	**Total**
Definitely not	558	154	28	0	0	740
Unlikely	325	223	47	1	0	596
Possibly	20	46	19	4	0	89
Probably	1	4	8	5	0	18
Definitely	0	0	2	4	3	9
Total	904	427	104	14	3	1452
Disagreements (%)	346 (38%)	204 (48%)	85 (82%)	9 (64%)	0 (0%)	644 (44%)

Crude agreement = 808/1452 = 56%, weighted kappa = 0.31 (0.27, 0.35), unweighted kappa = 0.20 (0.16, 0.24). 21 events did not have relatedness of treatment to SAE recorded by either adjudicators or investigators (Central Adjudicators = 19, Site Investigators = 2) and so were not included in this analysis

Figure 22 Agreement on likely causality of serious adverse events between central adjudicators and site investigators

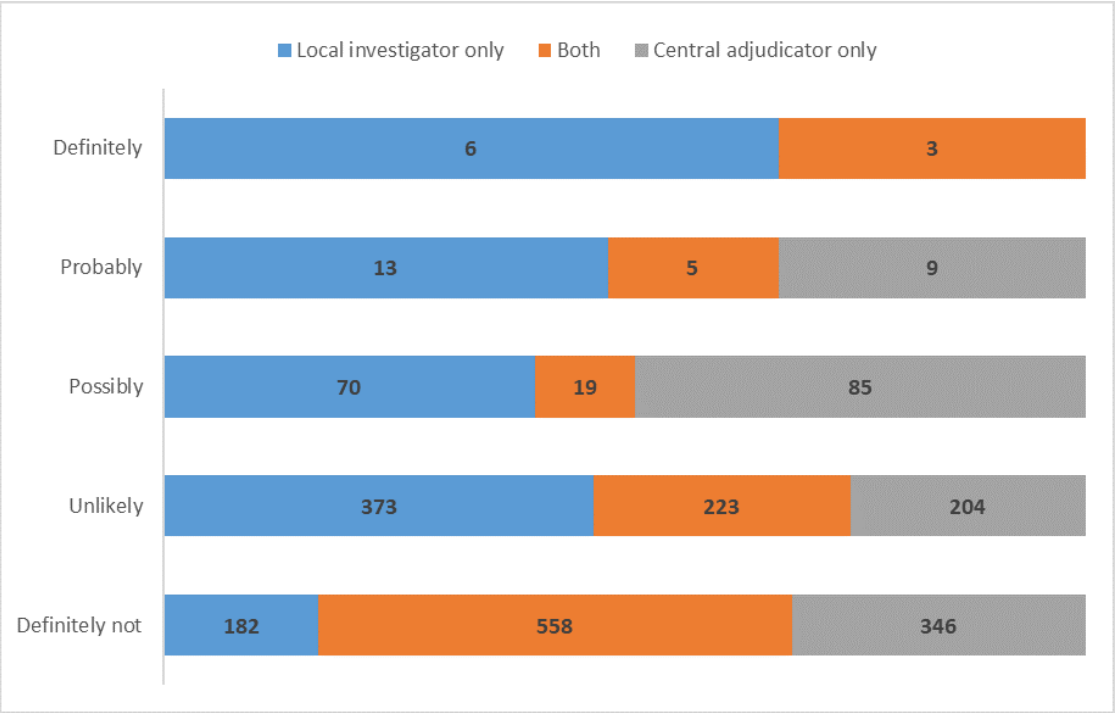


Table 56 Agreement on likely causality of SAEs between central adjudicators and site investigators – Dichotomous analysis 1

	Central Adjudicators		
Site Investigators	Definitely not, Unlikely or Possibly	Probably or Definitely	**Total**
Definitely not, Unlikely or Possibly	1420	5	1425
Probably or Definitely	15	12	27
Total	1435	17	1452
Disagreements (%)	15 (1%)	5 (29%)	20 (1%)

Crude agreement = 1432/1452 = 99%, unweighted kappa = 0.54 (0.49, 0.24). 21 events did not have relatedness of treatment to SAE recorded by either adjudicators or investigators (Central Adjudicators = 19, Site Investigators = 2) and so were not included in this analysis

Table 57 Agreement on likely causality of SAEs between central adjudicators and site investigators – Dichotomous analysis 2

	Central Adjudicators		
Site Investigators	Definitely not or Unlikely	Possibly, Probably or Definitely	**Total**
Definitely not or Unlikely	1260	76	1336
Possibly, Probably or Definitely	71	45	116
Total	1331	121	1452
Disagreements (%)	71 (5%)	76 (63%)	147 (10%)

Crude agreement = 1336/1452 = 90%, unweighted kappa = 0.32 (0.27, 0.38). 21 events did not have relatedness of treatment to SAE recorded by either adjudicators or investigators (Central Adjudicators = 19, Site Investigators = 2) and so were not included in this analysis

7.3.4 Time taken to adjudicate

Of the 1473 SAEs reported by site investigators, 15 had missing information for time of event completion. From the SAEs with complete data, 1297 (88%) were completed by adjudicators within 24 hours (Table 58). 127 (9%) events took at least 90 days for the adjudication to be completed as they required further information that was not readily available. This further information considerably extended the time taken for adjudication, and required resource from both the trial team and the team at the site where the information was being requested. For a number of events, this information was either not provided or not sufficient to complete the adjudication. For these events, the time was calculated at the time the adjudicator marked the event closed, even if it had not been resolved. It took 2882 days for one such case to be closed, with the event never resolved.

Table 58 Time taken for the adjudication of each serious adverse event

	Total (n=1458[1])
Timing of adjudication completion	
Adjudicator completed the event on same day	1297 (89%)
Adjudicator took more than one day to complete event	161 (11%)
Time taken to adjudicate event (days)	
Mean [SD]	46.8 [205.7]
Median [IQR]	0 [0, 0]
Min, max	0, 2882
If adjudicator took more than one day to complete event, time taken to adjudicate event (days)	
N	161
Mean [SD]	424.1 [473.5]
Median [IQR]	258 [50, 629]
Min, max	2, 2882

All data is N(%) unless otherwise indicated
[1]There were 15 events that had missing time of event completion, these events are not included

7.3.5 Effect of central adjudication on ENOS safety results

Central adjudication of SAEs had little impact compared with site investigators on the number of patients with any SAE reported for both the GTN vs No GTN and Continue vs Stop comparisons (Table 59). There were small differences between investigator- and adjudicator-reported SAEs for each SAE diagnosis, with evidence of a between-group effect differing for extension of initial stroke and recurrent stroke (both GTN vs No GTN comparison). However, these findings could be due to chance as the number of statistical tests carried out are high and the number of events did not substantially change for either extension or recurrence.

Furthermore, there was no evidence to suggest that central adjudication shifted the primary safety analysis (Table 60). Whilst there were small differences in the odds ratios between central adjudicators and site investigators for each classification and comparison, as seen with slight differences in the number of events in Table 48, all RORs contained the null value, 1, in their confidence intervals. Therefore, there is no evidence to suggest that the estimates for the ENOS safety analysis differ using the central adjudicator or site investigator reported events.

However, there were some classifications where site investigators identified events with limited accuracy. Investigators poorly identified complication of initial stroke, with 57 events classified that were not confirmed by adjudicators (Table 61).

Table 59 Number of patients with serious adverse events during follow-up at day 90 classified by site investigators and central adjudicators: GTN vs no GTN and Continue vs Stop

	GTN (n=2000)	No GTN (n=2011)	p-value*	Continue (n=1053)	Stop (n=1044)	p-value*
Any SAE						
Site Investigator	522	509	0.57	312	297	0.55
Central Adjudicator	520	502	0.45	310	294	0.52
Complication of initial stroke						
Site Investigator	67	53	0.18	26	36	0.19
Central Adjudicator	40	31	0.27	19	16	0.63
Extension of initial stroke						
Site Investigator	50	37	0.15	32	23	0.23
Central Adjudicator	59	37	0.021	33	23	0.19
Symptomatic intracranial haemorrhage						
Site Investigator	23	19	0.52	10	18	0.12
Central Adjudicator	27	17	0.13	11	16	0.32
Recurrent stroke						
Site Investigator	57	32	0.007	29	27	0.81
Central Adjudicator	47	35	0.17	25	27	0.80
Myocardial infarction						
Site Investigator	18	25	0.29	9	18	0.077
Central Adjudicator	19	22	0.65	11	16	0.32
Sudden cardiac death						
Site Investigator	3	5	0.48	1	3	0.31
Central Adjudicator	6	6	0.99	1	5	0.10
Other cardiovascular event						
Site Investigator	119	103	0.25	64	72	0.45
Central Adjudicator	109	97	0.37	60	66	0.55
Pulmonary embolism						
Site Investigator	26	17	0.16	10	11	0.81
Central Adjudicator	26	19	0.29	11	13	0.70
Pneumonia						
Site Investigator	89	100	0.44	74	50	0.030
Central Adjudicator	117	122	0.77	88	63	0.040
Other event						
Site Investigator	210	220	0.65	140	115	0.11
Central Adjudicator	212	212	0.95	134	119	0.35

GTN = Glyceryl trinitrate

*p-values are from a chi-squared test between treatment groups

Table 60 Impact of central adjudication on ENOS safety analysis: Ratio of Odds Ratios for serious adverse event classifications

	OR (95% CI) from Local Investigator data	OR (95% CI) from Central Adjudicator data	ROR (95% CI)
Any SAE			
GTN vs no GTN	0.72 (0.58, 0.91)	0.75 (0.60, 0.94)	0.96 (0.70, 1.32)
Continue vs Stop	0.81 (0.61, 1.08)	0.83 (0.62, 1.09)	0.98 (0.66, 1.47)
Complication of initial stroke			
GTN vs no GTN	1.19 (0.81, 1.73)	1.18 (0.73, 1.92)	1.00 (0.54, 1.85)
Continue vs Stop	0.66 (0.39, 1.12)	1.10 (0.56, 2.17)	0.60 (0.25, 1.42)
Extension of initial stroke			
GTN vs no GTN	1.27 (0.82, 1.97)	1.49 (0.97, 2.29)	0.85 (0.46, 1.57)
Continue vs Stop	1.34 (0.77, 2.34)	1.35 (0.78, 2.35)	0.99 (0.45, 2.17)
Symptomatic intracranial haemorrhage			
GTN vs no GTN	1.12 (0.61, 2.08)	1.46 (0.79, 2.71)	0.77 (0.32, 1.84)
Continue vs Stop	0.51 (0.23, 1.12)	0.62 (0.29, 1.36)	0.82 (0.27, 2.49)
Recurrent stroke			
GTN vs no GTN	1.71 (1.09, 2.67)	1.24 (0.79, 1.95)	1.37 (0.72, 2.58)
Continue vs Stop	1.02 (0.59, 1.75)	0.85 (0.48, 1.48)	1.20 (0.55, 2.63)
Myocardial infarction			
GTN vs no GTN	0.66 (0.35, 1.22)	0.78 (0.42, 1.45)	0.84 (0.35, 2.02)
Continue vs Stop	0.46 (0.20, 1.04)	0.62 (0.29, 1.36)	0.74 (0.24, 2.27)
Sudden cardiac death			
GTN vs no GTN	0.55 (0.13, 2.32)	0.91 (0.29, 2.83)	0.61 (0.10, 3.80)
Continue vs Stop	0.31 (0.03, 3.02)	0.18 (0.02, 1.57)	1.71 (0.08, 39.06)
Other cardiovascular event			
GTN vs no GTN	1.08 (0.81, 1.45)	1.03 (0.76, 1.38)	1.06 (0.70, 1.60)
Continue vs Stop	0.81 (0.56, 1.17)	0.81 (0.55, 1.19)	1.00 (0.59, 1.70)
Pulmonary embolism			
GTN vs no GTN	1.43 (0.77, 2.66)	1.25 (0.69, 2.29)	1.14 (0.48, 2.71)
Continue vs Stop	0.85 (0.36, 2.04)	0.77 (0.34, 1.75)	1.10 (0.34, 3.63)
Pneumonia			
GTN vs no GTN	0.80 (0.58, 1.08)	0.85 (0.64, 1.12)	0.94 (0.62, 1.43)
Continue vs Stop	1.49 (1.01, 2.20)	1.37 (0.96, 1.96)	1.09 (0.64, 1.85)
Other event			
GTN vs no GTN	0.86 (0.68, 1.09)	0.87 (0.69, 1.10)	1.00 (0.72, 1.39)
Continue vs Stop	1.21 (0.89, 1.64)	1.06 (0.78, 1.43)	1.14 (0.74, 1.75)

OR = Odds ratio, CI = Confidence interval, ROR = Ratio of odds ratios, GTN = Glyceryl trinitrate

All OR's from binary logistic regression models

Table 61 Agreement between site investigators and central adjudicators on the number of patients with serious adverse events during follow-up at day 90

	All (n=4011)	GTN (n=2000)	No GTN (n=2011)	Continue (n=1053)	Stop (n=1044)
Any SAE					
Both classify event	1022	520	502	310	294
Investigators classify event only	9	2	7	2	3
Adjudicators classify event only	0	0	0	0	0
Complication of initial stroke					
Both classify event	57	33	24	13	13
Investigators classify event only	57	31	26	12	20
Adjudicators classify event only	14	7	7	6	3
Extension of initial stroke					
Both classify event	68	43	25	25	17
Investigators classify event only	16	5	11	5	5
Adjudicators classify event only	27	15	12	8	5
Symptomatic intracranial haemorrhage					
Both classify event	34	19	15	10	14
Investigators classify event only	6	3	3	0	3
Adjudicators classify event only	7	6	1	1	1
Recurrent stroke					
Both classify event	70	41	29	21	23
Investigators classify event only	14	11	3	7	3
Adjudicators classify event only	10	4	6	4	3
Myocardial infarction					
Both classify event	33	12	21	9	13
Investigators classify event only	7	3	4	0	4
Adjudicators classify event only	6	5	1	2	2
Sudden cardiac death					
Both classify event	4	2	2	0	2
Investigators classify event only	3	1	2	1	0
Adjudicators classify event only	8	4	4	1	3
Other cardiovascular event					
Both classify event	160	87	73	47	48
Investigators classify event only	24	11	13	8	7
Adjudicators classify event only	14	5	9	6	3
Pulmonary embolism					
Both classify event	30	19	11	7	7
Investigators classify event only	0	0	0	0	0
Adjudicators classify event only	5	1	4	1	3
Pneumonia					
Both classify event	139	61	78	52	37
Investigators classify event only	5	1	4	4	1
Adjudicators classify event only	42	24	18	10	12
Other event					
Both classify event	246	112	134	74	67
Investigators classify event only	50	26	24	14	11
Adjudicators classify event only	48	20	28	13	18

GTN = Glyceryl trinitrate

7.4 Discussion

This chapter investigated the effect of central SAE adjudication in a large stroke trial with a single-blind design; here site assessors were not blinded to treatment allocation. The agreement between central adjudicators and site investigators was excellent for the diagnosis of events and good for classification of events. However, there was worse agreement on likely causality of SAEs when assessed on a 5-point scale, although this improved substantially when categories were dichotomised. Overall, adjudication of SAEs had no effect on the primary safety results of ENOS, with a similar number of adjudicator-reported and investigator-reported SAEs reported and similar odds ratios using either adjudicator-reported or investigator-reported SAEs.

However, this chapter did find that adjudication of SAEs has the potential to impact on the safety results of specific SAE classifications, in this case recurrent stroke and extension of initial stroke. There is a caveat needed with this finding, given that the number of events reported are low, the number of hypothesis tests conducted are high, and the actual number of events did not change dramatically in either case. In ENOS, adjudicators referred strictly to the protocol definitions for SAEs, whereas site investigators may not have followed these so rigorously, which could explain in part this finding. One potential approach for future studies could be to adjudicate a subset of events which are known to be difficult to diagnose, reducing the amount of adjudication needed whilst still ensuring the events with the highest chance of misclassification are adjudicated.

Furthermore, adjudication could be applied selectively, concentrating on major safety outcomes rather than all SAEs. Many trials, especially academically run, test treatments which already have a well-established safety profile, for which important safety issues can be predicted.[203] Therefore, relevant safety events can be listed at the trial design stage, and these can be chosen to be adjudicated, which would allow the number of events for adjudication to be greatly restricted and increase the chance that adjudication could be shown to impact meaningfully on the interpretation, since it limits the dilution of less relevant SAE 'noise'.

There was reasonable agreement between adjudicators and investigators for likely causality of SAEs. Investigators were more likely to report that SAEs were related to treatment than central adjudicators. After dichotomisation, events could be categorised as serious adverse reactions (SARs) which would be SAEs rated as probably or definitely related to treatment. In this case, agreement increased to 99%, and there would have been few SAEs that would have been incorrectly categorised as serious adverse reactions (SARs) and vice versa. This implies that even though agreement on likely causality was not high, the disagreements were confined to the related/not related to treatment categories, and there was minimal crossover, suggesting adjudication of SAEs may not be needed to ensure SAEs and SARs are categorised properly. The highest number of disagreements was found between adjudicators and investigators was for the *Possibly* category. Whilst assessing causality is subjective, the *Possibly* category is arguably the most subjective, and therefore it appears that this category is currently

unhelpful and not well understood. Furthermore, it could be argued that adjudicating causality is not relevant, given that safety reporting for serious adverse events in trials is required within a short pre-specified timeframe. Adjudication often takes place some weeks after the original event, and it is often the case that causality is not permitted to be downgraded, thus adjudication of causality has a limited impact in trials.

Whilst adjudication in trials may improve the accuracy of findings, the process comes at a considerable cost. For 161 of the SAEs assessed, adjudicators required more information than that provided by investigators, and this information took substantial time to obtain, in some cases with the adjudicators having to re-request data multiple times. Whilst only a portion of the time estimated would have actually been used to adjudicate SAEs, this highlights the sizeable task of implementing adjudication in clinical trials. This burden is shared between central trial and site teams, and must be funded appropriately.

A limitation of this study was that there was no quality control measure due to each SAE only being assessed by a single adjudicator. Whilst this meant the process used less resources, there was also a higher chance of random error[78], but level of agreement suggests this wasn't an issue in ENOS. Additionally, in ENOS, only events reported by investigators were adjudicated. Thus, after adjudication, the total number of SAEs could either stay the same, or decrease. This approach could potentially miss SAEs that were not identified by investigators during the trial. In fact, as investigators were not blinded to treatment allocation, there was the potential for SAEs to be selectively

reported. Granger et al.[43] suggests a potential alternative, where adjudicators also review “events that are ‘triggered’ by some feature that raises the suspicion of an event”. This method has been carried out previously, with computer algorithms used to identify events to adjudicate[60]. One strength of this study is that there was data on nearly 1500 SAEs from a large, well conducted randomised trial to use in this analysis. As well as this, the amount of missing data was low, with a high level of data completeness for both adjudicator- and investigator-assessed events. This study was designed and undertaken once the ENOS trial had finished, and so it was not possible that the presence of this study had any influence on investigators’ or adjudicators’ behaviour.

In summary, this chapter found that site investigators could accurately determine SAE diagnosis in a large international trial. Adjudication of investigator-reported SAEs had no impact on the main trial safety conclusions. This concludes the results of this thesis and the investigation into central adjudication in randomised stroke trials. The following chapter presents an overall discussion of the research conducted in this thesis in the context of previous research as well as discussing limitations of the research, directions of future work and general conclusions from this thesis.

Chapter 8 – Discussion

This thesis has presented research with the primary aim of answering the research question posed in Chapter 1: what are the benefits and costs of adjudicating outcomes in stroke trials? This chapter will briefly summarise Chapters 3-7 in turn and discuss these results in the context of previous research. I will then briefly highlight limitations with the research in this thesis, suggest potential avenues for future research and finally finish with an overall conclusion.

8.1 Summary of findings

In Chapter 3, 15 trials were identified in a systematic review that provided sufficient data to be included in a meta-analysis[194]. These studies, comprising 69,650 participants, were randomised stroke trials where central adjudicators and site investigators had both assessed the primary outcome. The primary result of the meta-analysis was a pooled Ratio of Treatment effects (RTE), which indicated that there was no evidence to suggest that the centrally adjudicated primary outcome provided a different treatment effect to the primary outcome assessed by site investigators (pooled RTE=1.02, 95% C.I: [0.95, 1.09]). All sensitivity analyses agreed with the primary finding. However, central adjudication did lead to fewer events being reported (pooled Odds Ratio (OR)=0.91, 95% C.I: [0.88, 0.95]), although this finding may have been driven by the fact that many of the studies only adjudicated site reported events (in these trials the number of events after adjudication can only reduce or stay the same).

During the systematic review 88 trials were identified that were potentially eligible to include, but did not report the outcome assessed by the site investigator. Thus, in Chapter 4, I presented the results from a methodological trial in which two methods of author contact were compared, to identify which method elicited the best initial response from the authors of the 88 trials[176]. The two methods selected to test were a short email with a protocol attached (Short) or a longer email without the protocol attached (Long). There was no evidence that either method differed in eliciting a response from the authors (Short vs Long, OR=1.10, 95% C.I: [0.36, 3.33]), and in a comparable timeframe (Short vs Long, hazard ratio=0.91, 95% C.I: [0.55, 1.51]). However, given the constrained sample size enforced by the systematic review, at 80% power the trial could detect a difference of 30 percentage points (OR=3.4), a magnitude of effect not commonly seen in trials. Thus, a larger sample size may have been able to detect a smaller difference between groups, if one existed.

A simulation study was carried out in Chapter 5, exploring the results of the main meta-analysis found in Chapter 3 as well as simulating hypothetical trials for further evaluation. For the first part of the simulation study, five trials that provided individual participant data had differential misclassification simulated to mimic the situation in which the site investigators are not blinded to treatment allocation and are biased in their outcome assessment, to favour treatment. The proportion of participants that were misclassified was increased until the confidence interval for the RTE did not include one, thus providing

evidence that the centrally adjudicated outcome gives a different treatment estimate compared with the biased site investigators outcome. For the trials with binary outcomes, the percentage of participants that needed to be misclassified for this to happen was between 2.1-6%. Thus, only a small proportion of participants need to have biased outcome assessment before central adjudication is vital in unblinded trials to ensure the trial results are valid.

In addition, the importance of adjudication for blinded trials was evaluated using simulation. Hypothetical trials were simulated, in which three levels of the treatment effect, here a risk ratio, were varied, as were five levels of the overall event rate and four levels of the sample size. Thus, there were 60 different hypothetical trial scenarios (three levels of treatment effect, five levels of event rate, four levels of sample size). For each scenario, non-differential misclassification was introduced to keep the treatment effect constant, and the proportion of events that needed to be non-differentially misclassified to make the treatment effect non-significant at 5% was calculated. This part of the simulation study found that for most of the scenarios, over 50% of the events needed to be non-differentially misclassified before a significant treatment effect became non-significant. This amount of random error is rare, and for stroke trials, implausible, suggesting that adjudication is not important for trials that have adequate blinding for outcome assessors.

In Chapter 6, the cost of central adjudication for nine of the trials included in the systematic review was estimated. This evaluation of cost

included estimating the cost of the adjudicators' time and the co-ordinating centre costs (e.g. website development, trial coordinator anonymising data, data coordinator chasing data queries with the adjudication), as well as the nominal fee that is usually paid to adjudicators per adjudication. In total, this corresponded to the adjudication process costing in excess of £100,000 for three of the nine trials included. This cost was then evaluated alongside the number of events that the adjudication process corrected (where the site investigators outcome was changed). For seven of the nine trials, at least one event was changed after adjudication, and the mean cost per event corrected was £2295.10. Therefore, the adjudication process does incur considerable cost, and if it is determined that the benefits of adjudication outweigh the costs, then this should be budgeted accordingly at the design stage for randomised trials.

The final chapter discussed here, Chapter 7, explored central adjudication of safety data in stroke trials. This chapter used a single case study, the Efficacy of Nitric Oxide in Stroke (ENOS) trial, in which both adjudicators and site investigators assessed serious adverse events (SAEs)[204]. In a similar fashion to Chapter 3, I found that the main safety results did not differ when SAEs were assessed by either adjudicators or site investigators. However, for more subjective outcomes, such as whether the SAE was caused by the trial allocated intervention, agreement was far worse between adjudicators and site investigators (Crude agreement = 56%, weighted kappa = 0.31). Therefore, it can be hypothesised that the results found in this thesis

will hold for the primary safety analysis in addition to the primary efficacy analysis. However, for more subjective outcomes, such as likely causality of SAEs to treatment allocation there may be additional benefit of central adjudication.

8.2 Findings in the context of previous research

In the discussion section for each chapter, I have already presented the results of this thesis in the context of previous research. In an effort to avoid repetition, I will summarise the main points of this thesis in relation to previous research in this area.

In this thesis, I found that central adjudication of the primary outcome in stroke trials did not alter treatment estimates in any of the 15 included trials, or when pooled in a meta-analysis. This result is in concordance with the two additional meta-analyses[44, 60] undertaken that compared adjudication to site assessment of outcomes, with all three meta-analyses finding very similar pooled estimates around the null value of one. However, these findings differ from three studies by Hrobjartsson that compare blinded and unblinded assessment of binary[30], time-to-event[33] and measurement scale outcomes[34]. All the studies by Hrobjartsson strictly compare blinded with unblinded outcome assessment, whilst in my review only 40% of the studies are comparing this, with the majority assessed by blinded site investigators. Moreover, Hrobjartsson only included outcomes defined to be subjective, in a similar fashion to the Cochrane review that also found the suggestion of an interaction effect by blinding status[60]. Through the simulation study,

I found that for unblinded trials, a small amount of differential misclassification can alter the treatment effect estimate and central adjudication is important to ensure trial results are correct. Conversely, it was shown that for trials with adequate blinding, a large amount of non-differential misclassification is needed before significant relative risks become non-significant. The importance of misclassification reduces as trial sample size increases and overall event rate increases up to 50%. For such trials, central adjudication has little impact and may be unnecessary. Thus, based on these findings and the results from previous research, it can be concluded that for large trials with blinded outcome assessment, the benefits of adjudication are small compared with its associated costs. Adjudication may not be necessary for these studies. However, for studies with no blinding, adjudication is important and should be implemented in future trials.

8.3 Limitations of this research

Similarly to the previous section, many limitations are given in each of the discussion chapters. Again, to avoid repetition, I will summarise the main limitations of this research briefly.

Firstly, the majority of the research in this thesis (Chapters 3, 5 and 6) is based on a group of fifteen trials (or a smaller subset of these). In fact, these represent only 17% (15/89) of the studies identified in the systematic review. It is possible that if more trials had provided data, the results of this research could have been different, or at least more

conclusive. However, the concordance between the 15 trials included was strong, with the individual findings from each trial similar.

Secondly, the simulation study could only use trials that provided individual patient data (IPD), and the cost-benefit analysis could only include studies that provided additional cost data. Specifying that all trials provide IPD could have resulted in less than 15 trials to include in the overall meta-analysis, so whilst I believe this is a limitation, I would not choose an alternative methodology if undertaking similar research in the future. The cost-benefit questionnaire was designed to be brief, with a very small number of questions and it could be completed in a short period of time. Thus, this further limitation is unfortunate but unavoidable, and obtaining cost data on the all of the studies was unrealistic.

Finally, this research makes the assumption throughout that it is the site investigators who are incorrect whenever there is conflicting opinion between adjudicators and site investigators. As mentioned in the introduction, adjudicators often only receive very limited patient data, and whilst they are often experts in their field, they may not have access to the same level of information that the site assessors have available (interaction with patient etc.). However, the assumption that one method of outcome assessment is gold standard (hence always correct) is neccessary for many of the methods employed in this thesis to be suitable. It is clearly more appropriate (and regularly assumed in previous research) that the centrally adjudicated outcome is gold standard. However, it is possible that on some occasions where site

investigators and central adjudicators disagree, that the latter are incorrect.

8.4 Direction of future work

The aim of the research presented in this thesis is to provide evidence to assist researchers in determining whether central adjudication should be included in future stroke trials. As described previously, there is potential to generalise these findings to adjudication in other clinical areas providing outcomes adjudicated are of a similar nature and the adjudication process is comparable. For example, adjudication also occurs in neonatal trials, with outcomes such as intraventricular haemorrhage grade[205], sepsis and necrotizing enterocolitis[206, 207] adjudicated. All three outcomes are arguably more subjective than many of the outcomes adjudicated in the stroke trials presented in this thesis. An evaluation of the adjudication process in a neonatal trial with both sepsis and necrotizing enterocolitis adjudicated found that similar results were obtained both with and without adjudication[208]. Additional blinded outcome assessment is also prevalent in many other clinical areas such as dermatology[209, 210]. For example, ulcers can be photographed and blinded assessors can measure the area of the ulcer in addition to an unblinded crude measurement that happens at the site[211]. Therefore, there is the potential to take similar methodology applied in this research programme to investigate other clinical areas where central adjudication is prevalent, or where additional blinded assessment of an outcome occurs. I would hypothesise, based on the results of this thesis, that additional assessment of outcomes would be

most beneficial where the outcomes adjudicated/assessed are complex, subjective and were assessed at site in an unblinded fashion. Thus, it would be interesting and important to explore these cases with future study.

Furthermore, the possibility of a prospective randomised adjudication study would be an alternative to the retrospective analysis that is contained in this thesis (and in most studies assessing the benefits of adjudication). Study within a trial (SWAT) methodology could be employed to prospectively test the value of adjudication of the primary outcome in an ongoing trial, with participants randomised to adjudication/no-adjudication in a separate study to the main trial. There may be concerns regarding whether this process could affect the main trial results, but all patients could be adjudicated, with the no-adjudication group having their adjudication occur after the SWAT has undertaken between-group analysis. A prospective adjudication study is planned for a cardiovascular trial[82]. The GLOBAL LEADERS Adjudication Sub-StudY (GLASSY) aims to quantify the level of agreement between site investigator and central adjudicator reported events. This study is prospective, and has been planned during set-up of the main trial, but does not randomise participants as I had suggested earlier. However, it still provides a framework from which to build on and extend to include randomisation of outcome assessment method as discussed earlier.

In addition to central adjudication, there is scope to explore additional areas involving research waste. Use of routinely collected data is

becoming increasingly popular, in part due to the ease in collecting outcome data on many thousands of patients. This data is starting to be utilised by clinical trials, as a method to save time, costs, and carry out large trials that would not be feasible using traditional data collection methods. The use of routine data could transform clinical research and studies have shown that participants can be matched to hospital records and that outcomes can be determined successfully[212], although there can be a large extent of missing data for some outcomes[213]. Mc Cord et al. call for future research to "focus on data quality validation, alternative research designs and how they affect outcome assessment"[214]. Thus, through employing similar methodology used in this thesis, trials can be analysed using both routinely and purposively collected data to investigate whether routinely collected data provides a suitable alternative, and for which kind of studies it can be used.

Finally, the methodological trial described in Chapter 4 could be extended to assess an alternative intervention, or a similar trial could be undertaken with a larger sample size (or with a sample size calculation carried out a prori to determine the number of authors who would need to be contacted). The methodological trial in this thesis compared short email and protocol with long email and no protocol, but it could be of interest to explore a factorial trial with email length (short/long) and protocol (attached/no attached) as factors. In fact, the study also offered all authors the chance to collaborate, thus giving them the possibility of academic incentives, and it could be explored in a further

study whether offering collaboration elicits a different response. An alternative study could explore other, potentially more useful, outcomes such as whether data is received. However, thought must go into what an appropriate intervention may be, and how to account for the differences in interaction after the intervention takes place if it involves initial contact like in the trial described in this thesis. Any new study should explore whether it would be possible to blind the trial team, as outcomes (especially regarding whether data is received) could be easily biased by beliefs of the trial team. Regardless, this area is an exciting one to pursue with a large amount of potential possibilities.

8.5 Reflection on this thesis

In reflection on the results of my thesis, it is my belief that central adjudication can still be a worthwhile process for stroke trials. However, there are some trials, possessing certain characteristics, in which I believe central adjudication of the primary outcome is not necessary. I do not believe that adjudication is useful for trials that have objective outcomes, where these outcomes are assessed in a blinded fashion. I would hypothesise (based on opinion and not upon the results of this thesis) that adjudication will be most useful for trials with subjective outcomes. It was shown in the simulation that an extreme (and implausible) amount of random error is required before modest true treatment effects are missed. Thus, one could argue that adjudication is not a worthwhile process for any trial with sufficient blinding of the outcome assessors regardless of objectivity of the outcome. However, I believe that this statement is too strong, and without new evidence to

support this claim, would temper it (as I have above) with the assertion that adjudication may still be useful for appropriately blinded trials provided that the outcome to be adjudicated is sufficiently subjective. I do believe that central adjudication is a very important clinical trial process for studies that do not have blinded outcome assessment. In Chapter 5, I gave examples of situations in which trials of this kind could provide completely different treatment effect estimates to the true treatment estimate with only a small amount of bias. Thus, for studies that do not have any (or sufficient) blinding for primary outcome assessment, I believe that adjudication of the primary outcome is important to ensure the validity of the trial findings.

In further reflections into the approach and methodology used in this thesis, I believe that I have used appropriate methodology and would follow a similar approach again to my primary analysis if I was to start this research programme again. However, this approach does not come without limitations, for example needing to assume that the adjudicated result is always correct, as discussed above. An alternative method to my accuracy based approach would be to explore the value of adjudication using a variance decomposition approach. For example, I could have approached this problem using generalisability theory, which would have relaxed the assumptions my accuracy based approach made. These variance decomposition approaches would permit a different evaluation into the importance of central adjudication, for example, assessing whether a subset of participants can be adjudicated rather than the entire trial population that exhibited an

event. However, an accuracy based approach does quantify an arguably more interpretable estimate of the difference between adjudication and site assessment, and this may be easier for trialists to understand and potentially act on. Indeed, all adjudication approaches that I am aware of to date have used such an approach. Both methodologies have their differences, with both having their own positives and negatives. Thus, it would be an interesting extension to this thesis, and an exciting foray into generalisability theory, to explore the impact of adjudication using a variance decomposition approach in addition to the work carried out in this thesis.

8.6 Conclusions

To conclude, this thesis has provided evidence about the benefits and costs of central adjudication of outcomes in stroke trials. Central adjudication of the primary outcome is of little importance for large blinded stroke trials, and in fact, it could be argued that it is unnecessary for all blinded stroke trials. This is in part due to the objective nature of the outcomes that are commonly used as the primary outcome in stroke trials. However, for open-label trials, those where blinding could become compromised, or those with subjective primary outcomes, central adjudication can be an important component in protecting the validity of the main trial results. Given that adjudication comes with non-trivial expense, it is important to have a clear rationale for its inclusion before implementing such a process in future studies.

References

1. Townsend, N., K. Wickramasinghe, P. Bhatnagar, K. Smolina, M. Nichols, J. Leal, R. Luengo-Fernandez, and M. Rayner, *Corony heart disease statistics*. 2012.
2. Stroke Association. *State of the nation*. 2017; Available from: https://www.stroke.org.uk/sites/default/files/state_of_the_nation_2017_final_1.pdf.
3. World Health Organisation. *The top 10 causes of death.* 2017; Available from: http://www.who.int/mediacentre/factsheets/fs310/en/.
4. Feigin, V.L., M.H. Forouzanfar, R. Krishnamurthi, G.A. Mensah, M. Connor, D.A. Bennett, A.E. Moran, R.L. Sacco, L. Anderson, T. Truelsen, M. O'Donnell, N. Venketasubramanian, S. Barker-Collo, C.M.M. Lawes, W. Wang, Y. Shinohara, E. Witt, M. Ezzati, M. Naghavi, C. Murray, I. on behalf of the Global Burden of Diseases, S. Risk Factors, and G.B.D.S.E.G. the, *Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010.* Lancet, 2014. **383**(9913): p. 245-254.
5. Luengo-Fernandez, R., J. Leal, and A. Gray, *UK research spend in 2008 and 2012: comparing stroke, cancer, coronary heart disease and dementia.* BMJ Open, 2015. **5**(4).
6. Barton, S., *Which clinical studies provide the best evidence?* BMJ, 2000. **321**(7256): p. 255.
7. U.S. National Institutes of Health. *Trends, Charts, and Maps*. 2017; Available from: https://clinicaltrials.gov/ct2/resources/trends#TypesOfRegisteredStudies.
8. Ali, M., P.M.W. Bath, J. Curram, S.M. Davis, H.-C. Diener, G.A. Donnan, M. Fisher, B.A. Gregson, J. Grotta, W. Hacke, M.G. Hennerici, M. Hommel, M. Kaste, J.R. Marler, R.L. Sacco, P. Teal, N.-G. Wahlgren, S. Warach, C.J. Weir, and K.R. Lees, *The*

Virtual International Stroke Trials Archive. Stroke, 2007. **38**(6): p. 1905.

9. Yew, K.S. and E. Cheng, *Acute Stroke Diagnosis.* American family physician, 2009. **80**(1): p. 33-40.
10. Jauch, E., S. Kasab, and B. Stettler. *Ischemic Stroke*. 2017; Available from: http://emedicine.medscape.com/article/1916852-overview.
11. Liebeskind, D. *Hemorrhagic Stroke*. 2017; Available from: http://emedicine.medscape.com/article/1916662-overview.
12. Nour, M. and D.S. Liebeskind, *Brain imaging in stroke: insight beyond diagnosis.* Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics, 2011. **8**(3): p. 330-339.
13. De Keyser, J., Z. Gdovinová, M. Uyttenboogaart, P.C. Vroomen, and G.J. Luijckx, *Intravenous Alteplase for Stroke.* Stroke, 2007. **38**(9): p. 2612.
14. Wahlgren, N., T. Moreira, P. Michel, T. Steiner, O. Jansen, C. Cognard, H.P. Mattle, W. van Zwam, S. Holmin, T. Tatlisumak, J. Petersson, V. Caso, W. Hacke, M. Mazighi, M. Arnold, U. Fischer, I. Szikora, L. Pierot, J. Fiehler, J. Gralla, F. Fazekas, and K.R. Lees, *Mechanical thrombectomy in acute ischemic stroke: Consensus statement by ESO-Karolinska Stroke Update 2014/2015, supported by ESO, ESMINT, ESNR and EAN.* International Journal of Stroke, 2015. **11**(1): p. 134-147.
15. Burn, J., M. Dennis, J. Bamford, P. Sandercock, D. Wade, and C. Warlow, *Long-term risk of recurrent stroke after a first-ever stroke. The Oxfordshire Community Stroke Project.* Stroke, 1994. **25**(2): p. 333-7.
16. Cramer, S.C., S.L. Wolf, H.P. Adams, Jr., D. Chen, A.W. Dromerick, K. Dunning, C. Ellerbe, A. Grande, S. Janis, M.G. Lansberg, R.M. Lazar, Y.Y. Palesch, L. Richards, E. Roth, S.I. Savitz, L.R. Wechsler, M. Wintermark, and J.P. Broderick, *Stroke Recovery and Rehabilitation Research: Issues, Opportunities,*

and the National Institutes of Health StrokeNet. Stroke, 2017. **48**(3): p. 813-819.

17. Yordanov, Y., A. Dechartres, I. Atal, V.-T. Tran, I. Boutron, P. Crequit, and P. Ravaud, *Avoidable waste of research related to outcome planning and reporting in clinical trials.* BMC Medicine, 2018. **16**(1): p. 87.
18. The Lancet. *The REWARD Statement*. 2014; Available from: http://www.thelancet.com/campaigns/efficiency/statement.
19. Chalmers, I., M.B. Bracken, B. Djulbegovic, S. Garattini, J. Grant, A.M. Gülmezoglu, D.W. Howells, J.P.A. Ioannidis, and S. Oliver, *How to increase value and reduce waste when research priorities are set.* The Lancet. **383**(9912): p. 156-165.
20. Ioannidis, J.P.A., S. Greenland, M.A. Hlatky, M.J. Khoury, M.R. Macleod, D. Moher, K.F. Schulz, and R. Tibshirani, *Increasing value and reducing waste in research design, conduct, and analysis.* The Lancet. **383**(9912): p. 166-175.
21. Salman, R.A.-S., E. Beller, J. Kagan, E. Hemminki, R.S. Phillips, J. Savulescu, M. Macleod, J. Wisely, and I. Chalmers, *Increasing value and reducing waste in biomedical research regulation and management.* The Lancet. **383**(9912): p. 176-185.
22. Chan, A.-W., F. Song, A. Vickers, T. Jefferson, K. Dickersin, P.C. Gøtzsche, H.M. Krumholz, D. Ghersi, and H.B. van der Worp, *Increasing value and reducing waste: addressing inaccessible research.* The Lancet. **383**(9913): p. 257-266.
23. Glasziou, P., D.G. Altman, P. Bossuyt, I. Boutron, M. Clarke, S. Julious, S. Michie, D. Moher, and E. Wager, *Reducing waste from incomplete or unusable reports of biomedical research.* The Lancet. **383**(9913): p. 267-276.
24. von Niederhäusern, B., G.H. Guyatt, M. Briel, and C. Pauli-Magnus, *Academic response to improving value and reducing waste: A comprehensive framework for INcreasing QUality In patient-oriented academic clinical REsearch (INQUIRE).* PLOS Medicine, 2018. **15**(6): p. e1002580.

25. Treweek, S., D.G. Altman, P. Bower, M. Campbell, I. Chalmers, S. Cotton, P. Craig, D. Crosby, P. Davidson, D. Devane, L. Duley, J. Dunn, D. Elbourne, B. Farrell, C. Gamble, K. Gillies, K. Hood, T. Lang, R. Littleford, K. Loudon, A. McDonald, G. McPherson, A. Nelson, J. Norrie, C. Ramsay, P. Sandercock, D.R. Shanahan, W. Summerskill, M. Sydes, P. Williamson, and M. Clarke, *Making randomised trials more efficient: report of the first meeting to discuss the Trial Forge platform.* Trials, 2015. **16**(1): p. 261.
26. Trial Forge. *A systematic approach to making trials more efficient*. 2016; Available from: http://www.trialforge.org/.
27. Kahan, B.C., S. Rehal, and S. Cro, *Blinded Outcome Assessment Was Infrequently Used and Poorly Reported in Open Trials.* PloS one, 2015. **10**(6): p. e0131926-e0131926.
28. Hrobjartsson, A., E. Forfang, M.T. Haahr, B. Als-Nielsen, and S. Brorson, *Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding.* Int J Epidemiol, 2007. **36**(3): p. 654-63.
29. Bello, S., H. Moustgaard, and A. Hróbjartsson, *The risk of unblinding was infrequently and incompletely reported in 300 randomized clinical trial publications.* Journal of Clinical Epidemiology, 2014. **67**(10): p. 1059-1069.
30. Hrobjartsson, A., A.S. Thomsen, F. Emanuelsson, B. Tendal, J. Hilden, I. Boutron, P. Ravaud, and S. Brorson, *Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors.* BMJ, 2012. **344**: p. e1119.
31. Lambert, J., *Statistics in brief: how to assess bias in clinical studies?* Clinical orthopaedics and related research, 2011. **469**(6): p. 1794-1796.
32. Moustgaard, H., S. Bello, F.G. Miller, and A. Hrobjartsson, *Subjective and objective outcomes in randomized clinical trials: definitions differed in methods publications and were often*

absent from trial reports. J Clin Epidemiol, 2014. **67**(12): p. 1327-34.

33. Hrobjartsson, A., A.S. Thomsen, F. Emanuelsson, B. Tendal, J.V. Rasmussen, J. Hilden, I. Boutron, P. Ravaud, and S. Brorson, *Observer bias in randomized clinical trials with time-to-event outcomes: systematic review of trials with both blinded and non-blinded outcome assessors.* Int J Epidemiol, 2014. **43**(3): p. 937-48.
34. Hrobjartsson, A., A.S. Thomsen, F. Emanuelsson, B. Tendal, J. Hilden, I. Boutron, P. Ravaud, and S. Brorson, *Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors.* Cmaj, 2013. **185**(4): p. E201-11.
35. Viswanathan, M., N. Berkman, D. Dryden, and L. Hartling, *Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank.* 2013: p. Appendix A, Approaches to Assessing the Risk of Bias in Studies.
36. Higgins, J., D. Altman, and J. Sterne, *Cochrane Handbook for Systematic Review os Interventions, Chapter 8: Assessing risk of bias in included studies*. 2011.
37. Hansson, L., T. Hedner, and B. Dahlof, *Prospective randomized open blinded end-point (PROBE) study. A novel design for intervention trials. Prospective Randomized Open Blinded End-Point.* Blood Press, 1992. **1**(2): p. 113-9.
38. Mansournia, M.A., J.P.T. Higgins, J.A.C. Sterne, and M.A. Hernán, *Biases in Randomized Trials: A Conversation Between Trialists and Epidemiologists.* Epidemiology (Cambridge, Mass.), 2017. **28**(1): p. 54-59.
39. Karanicolas, P.J., F. Farrokhyar, and M. Bhandari, *Practical tips for surgical research: blinding: who, what, when, why, how?* Canadian journal of surgery. Journal canadien de chirurgie, 2010. **53**(5): p. 345-348.

40. Mackintosh, D.R. and V. Molloy, *Differences in clinical trial conduct in US and EU investigational sites.* The Quality Assurance Journal, 2001. **5**(1): p. 13-17.
41. Ford, R., M. O'Neal, S. Moskowitz, and J. Fraunberger, *Adjudication Rates between Readers in Blinded Independent Central Review of Oncology Studies.* Journal of Clinical Trials, 2016. **6**(5).
42. Dechartres, A., I. Boutron, C. Roy, and P. Ravaud, *Inadequate planning and reporting of adjudication committees in clinical trials: Recommendation proposal.* Journal of Clinical Epidemiology. **62**(7): p. 695-702.
43. Granger, C.B., V. Vogel, S.R. Cummings, P. Held, F. Fiedorek, M. Lawrence, B. Neal, H. Reidies, L. Santarelli, R. Schroyer, N.L. Stockbridge, and Z. Feng, *Do we need to adjudicate major clinical events?* Clin Trials, 2008. **5**(1): p. 56-60.
44. Pogue, J., S.D. Walter, and S. Yusuf, *Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs.* Clinical Trials, 2009. **6**(3): p. 239-251.
45. Serebruany, V.L. and S.D. Fortmann, *International central adjudication committee in the PLATO trial: independent body of experts or friendly family picnic?* Expert Review of Cardiovascular Therapy, 2015. **13**(8): p. 867-869.
46. US Food and Drug Administration. *Guidance for Clinical Trial Sponsors. Establishment and Operation of Clinical Trial Data Monitoring Committees.* 2006. [Accessed 30 July 2019]; Available from: http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm127073.pdf.
47. European Medicines Agency. *Committee for Medicinal Products for Human Use - Guideline on Data Monitoring Committees*. 2005. [Accessed 30 July 2019]; Available from: http://osp.od.nih.gov/sites/default/files/resources/WC500003635.pdf).

48. Seltzer, J.H., T. Heise, P. Carson, D. Canos, J.C. Hiatt, P. Vranckx, T. Christen, and D.E. Cutlip, *Use of endpoint adjudication to improve the quality and validity of endpoint assessment for medical device development and post marketing evaluation: Rationale and best practices. A report from the cardiac safety research consortium.* Am Heart J, 2017. **190**: p. 76-85.
49. Krumholz-Bahner, S., M. Garibbo, K.A. Getz, and B.E. Widler, *An Overview and Analysis Regarding the Use of Adjudication Methods in EU and US Drug Approvals.* Therapeutic Innovation & Regulatory Science, 2015. **49**(6): p. 831-839.
50. Hata, J., H. Arima, S. Zoungas, G. Fulcher, C. Pollock, M. Adams, J. Watson, R. Joshi, A.P. Kengne, T. Ninomiya, C. Anderson, M. Woodward, A. Patel, G. Mancia, N. Poulter, S. MacMahon, J. Chalmers, B. Neal, and A.C.G. on behalf of the, *Effects of the Endpoint Adjudication Process on the Results of a Randomised Controlled Trial: The ADVANCE Trial.* PLoS ONE, 2013. **8**(2): p. e55807.
51. Ninomiya, T., G. Donnan, N. Anderson, C. Bladin, B. Chambers, G. Gordon, N. Sharpe, J. Chalmers, M. Woodward, and B. Neal, *Effects of the End Point Adjudication Process on the Results of the Perindopril Protection Against Recurrent Stroke Study (PROGRESS).* Stroke, 2009. **40**(6): p. 2111.
52. Godolphin, P.J., T. Hepburn, N. Sprigg, L. Walker, E. Berge, R. Collins, J. Gommans, G. Ntaios, S. Pocock, K. Prasad, J.M. Wardlaw, P.M. Bath, and A.A. Montgomery, *Central masked adjudication of stroke diagnosis at trial entry offered no advantage over diagnosis by local clinicians: Secondary analysis and simulation.* Contemporary clinical trials communications, 2018. **12**: p. 176-181.
53. Kidwell Chelsea, S., S. Liebeskind David, S. Starkman, and L. Saver Jeffrey, *Trends in Acute Ischemic Stroke Trials Through the 20th Century.* Stroke, 2001. **32**(6): p. 1349-1359.

54. Walter, S.D., D.J. Cook, G.H. Guyatt, and D. King, *Outcome assessment for clinical trials: How many adjudicators do we need?* Controlled Clinical Trials, 1997. **18**(1): p. 27-42.
55. Stuck, A.K., E. Fuhrer, A. Limacher, M. Méan, and D. Aujesky, *Adjudication-related processes are underreported and lack standardization in clinical trials of venous thromboembolism: a systematic review.* Journal of Clinical Epidemiology. **67**(3): p. 278-284.
56. Kahan, B.C., B. Feagan, and V. Jairath, *A comparison of approaches for adjudicating outcomes in clinical trials.* Trials, 2017. **18**(1): p. 266.
57. Curb, J.D., A. McTiernan, S.R. Heckbert, C. Kooperberg, J. Stanford, M. Nevitt, K.C. Johnson, L. Proulx-Burns, L. Pastore, M. Criqui, and S. Daugherty, *Outcomes ascertainment and adjudication methods in the women's health initiative.* Annals of Epidemiology, 2003. **13**(9, Supplement): p. S122-S128.
58. Arnold, D.M., F. Lauzier, C. Rabbat, N. Zytaruk, B. Barlow Cash, F. Clarke, D. Heels-Ansdell, G. Guyatt, S.D. Walter, A. Davies, and D.J. Cook, *Adjudication of bleeding outcomes in an international thromboprophylaxis trial in critical illness.* Thrombosis Research. **131**(3): p. 204-209.
59. Saunders, L., N. Zytaruk, J. Southon, B.B. Cash, K. Lutz, F. Clarke, S. Duchesne, K. Krolicki, and M. Heels-Ansdell. *Methods Center Adjudication Costs For 5 Morbidity Outcomes In A Thromboprophylaxis Trial*. in *AMERICAN JOURNAL OF RESPIRATORY AND CRITICAL CARE MEDICINE*. 2011. Am Thoracic Soc.
60. Ndounga Diakou, L.A., L. Trinquart, A. Hróbjartsson, C. Barnes, A. Yavchitz, P. Ravaud, and I. Boutron, *Comparison of central adjudication of outcomes and onsite outcome assessment on treatment effect estimates.* Cochrane Database of Systematic Reviews, 2016(3).

61. Paul, M. and L. Leibovici, *Systematic review or meta-analysis? Their place in the evidence hierarchy.* Clinical Microbiology and Infection, 2014. **20**(2): p. 97-100.
62. Cooper, H. and E. Patall. *The Relative Benefits of Meta-Analysis Conducted With Individual Participant Data Versus Aggregated Data*. 2009; 2:[165-176]. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=ovftj&NEWS=N&AN=00060744-200906000-00006.
63. Polanin, J.R., *Efforts to retrieve individual participant data sets for use in a meta-analysis result in moderate data sharing but many data sets remain missing.* Journal of Clinical Epidemiology, 2018. **98**: p. 157-159.
64. Jaspers, G.J. and P.L.J. Degraeuwe, *A failed attempt to conduct an individual patient data meta-analysis.* Systematic reviews, 2014. **3**: p. 97-97.
65. Meursinge Reynders, R., L. Ladu, and N. Di Girolamo, *Contacting of authors by systematic reviewers: protocol for a cross-sectional study and a survey.* Systematic reviews, 2017. **6**(1): p. 249-249.
66. Taichman, D.B., P. Sahni, A. Pinborg, L. Peiperl, C. Laine, A. James, S.-T. Hong, A. Haileamlak, L. Gollogly, F. Godlee, F.A. Frizelle, F. Florenzano, J.M. Drazen, H. Bauchner, C. Baethge, and J. Backus, *Data Sharing Statements for Clinical Trials — A Requirement of the International Committee of Medical Journal Editors.* New England Journal of Medicine, 2017. **376**(23): p. 2277-2279.
67. Hrynaszkiewicz, I., V. Khodiyar, A.L. Hufton, and S.-A. Sansone, *Publishing descriptions of non-public clinical datasets: proposed guidance for researchers, repositories, editors and funding organisations.* Research Integrity and Peer Review, 2016. **1**(1): p. 6.
68. Savage, C.J. and A.J. Vickers, *Empirical Study of Data Sharing by Authors Publishing in PLoS Journals.* PLOS ONE, 2009. **4**(9): p. e7078.

69. Lundh, A., L.T. Krogsbøll, and P.C. Gøtzsche, *Access to data in industry-sponsored trials.* The Lancet, 2011. **378**(9808): p. 1995-1996.
70. Wicherts, J.M., M. Bakker, and D. Molenaar, *Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results.* PLOS ONE, 2011. **6**(11): p. e26828.
71. Ahmed, I., A.J. Sutton, and R.D. Riley, *Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey.* BMJ, 2012. **344**: p. d7762.
72. Veroniki, A.A., S.E. Straus, H. Ashoor, L.A. Stewart, M. Clarke, and A.C. Tricco, *Contacting authors to retrieve individual patient data: study protocol for a randomized controlled trial.* Trials, 2016. **17**(1): p. 138.
73. Veroniki, A.A., H.M. Ashoor, S.P.C. Le, P. Rios, L.A. Stewart, M. Clarke, D. Mavridis, S.E. Straus, and A.C. Tricco, *Retrieval of individual patient data depended on study characteristics: A randomised controlled trial.* Journal of Clinical Epidemiology, 2019.
74. Ohmann, C., R. Banzi, S. Canham, S. Battaglia, M. Matei, C. Ariyo, L. Becnel, B. Bierer, S. Bowers, L. Clivio, M. Dias, C. Druml, H. Faure, M. Fenner, J. Galvez, D. Ghersi, C. Gluud, T. Groves, P. Houston, G. Karam, D. Kalra, R.L. Knowles, K. Krleža-Jerić, C. Kubiak, W. Kuchinke, R. Kush, A. Lukkarinen, P.S. Marques, A. Newbigging, J. O'Callaghan, P. Ravaud, I. Schlünder, D. Shanahan, H. Sitter, D. Spalding, C. Tudur-Smith, P. van Reusel, E.-B. van Veen, G.R. Visser, J. Wilson, and J. Demotes-Mainard, *Sharing and reuse of individual participant data from clinical trials: principles and recommendations.* BMJ Open, 2017. **7**(12): p. e018647.
75. Polanin, J.R. and M. Terzian, *A data-sharing agreement helps to increase researchers' willingness to share primary data: results*

from a randomized controlled trial. J Clin Epidemiol, 2019. **106**: p. 60-69.

76. Teunis, T., S.P.F.T. Nota, and J.H. Schwab, *Do corresponding authors take responsibility for their work? A covert survey.* Clinical orthopaedics and related research, 2015. **473**(2): p. 729-735.
77. Kelley, G.A., K.S. Kelley, and Z.V. Tran, *Retrieval of missing data for meta-analysis: a practical example.* International journal of technology assessment in health care, 2004. **20**(3): p. 296-299.
78. Simunovic, N., S. Walter, P.J. Devereaux, S. Sprague, G.H. Guyatt, E. Schemitsch, P. Tornetta, 3rd, D. Sanders, M. Swiontkowski, and M. Bhandari, *Outcomes assessment in the SPRINT multicenter tibial fracture trial: Adjudication committee size has trivial effect on trial results.* J Clin Epidemiol, 2011. **64**(9): p. 1023-33.
79. Mahaffey, K.W., R.A. Harrington, M. Akkerhuis, N.S. Kleiman, L.G. Berdan, B.S. Crenshaw, B.E. Tardiff, C.B. Granger, I. DeJong, M. Bhapkar, P. Widimsky, R. Corbalon, K.L. Lee, J.W. Deckers, M.L. Simoons, E.J. Topol, R.M. Califf, and P.I. for the, *Disagreements between central clinical events committee and site investigator assessments of myocardial infarction endpoints in an international clinical trial: review of the PURSUIT study.* Current Controlled Trials in Cardiovascular Medicine, 2001. **2**(4): p. 187-194.
80. Serebruany, V.L., *Delays of event adjudication in the TRITON trial.* Cardiology, 2010. **115**(3): p. 217-20.
81. Serebruany, V.L. and D. Atar, *Viewpoint: Central adjudication of myocardial infarction in outcome-driven clinical trials--common patterns in TRITON, RECORD, and PLATO?* Thromb Haemost, 2012. **108**(3): p. 412-4.
82. Leonardi, S., A. Franzone, R. Piccolo, E. McFadden, P. Vranckx, P. Serruys, E. Benit, C. Liebetrau, L. Janssens, M. Ferrario, A. Zurakowski, R.-J. van Geuns, M. Dominici, K. Huber, T.

Slagboom, P. Buszman, L. Bolognese, C. Tumscitz, K. Bryniarski, A. Aminian, M. Vrolix, I. Petrov, S. Garg, C. Naber, J. Prokopczuk, C. Hamm, G. Steg, D. Heg, P. Juni, S. Windecker, and M. Valgimigli, *Rationale and design of a prospective substudy of clinical endpoint adjudication processes within an investigator-reported randomised controlled trial in patients with coronary artery disease: the GLOBAL LEADERS Adjudication Sub-StudY (GLASSY).* BMJ Open, 2019. **9**(3): p. e026053.

83. Zhao, W. and K. Pauls, *Architecture design of a generic centralized adjudication module integrated in a web-based clinical trial management system.* Clin Trials, 2016. **13**(2): p. 223-33.
84. Held, C., *When do we need clinical endpoint adjudication in clinical trials?* Ups J Med Sci, 2019. **124**(1): p. 42-45.
85. Kjoller, E., J. Hilden, P. Winkel, S. Galatius, N.J. Frandsen, G.B. Jensen, J. Fischer Hansen, J. Kastrup, C.M. Jespersen, P. Hildebrandt, H.J. Kolmos, and C. Gluud, *Agreement between public register and adjudication committee outcome in a cardiovascular randomized clinical trial.* Am Heart J, 2014. **168**(2): p. 197-204.e1-4.
86. Sepehrvand, N., Y. Zheng, P.W. Armstrong, R. Welsh, S.G. Goodman, W. Tymchak, F. Khadour, M. Chan, D. Weiss, and J.A. Ezekowitz, *Alignment of site versus adjudication committee–based diagnosis with patient outcomes: Insights from the Providing Rapid Out of Hospital Acute Cardiovascular Treatment 3 trial.* Clinical Trials, 2015. **13**(2): p. 140-148.
87. Petersen, J.L., G. Haque, A.S. Hellkamp, G.C. Flaker, N.A. Mark Estes, 3rd, F.E. Marchlinski, J.H. McAnulty, A.J. Greenspon, R.A. Marinchak, K.L. Lee, G.A. Lamas, and K.W. Mahaffey, *Comparing classifications of death in the Mode Selection Trial: agreement and disagreement among site investigators and a clinical events committee.* Contemp Clin Trials, 2006. **27**(3): p. 260-8.

88. Mahaffey, K.W., C. Held, D.M. Wojdyla, S.K. James, H.A. Katus, S. Husted, P.G. Steg, C.P. Cannon, R.C. Becker, R.F. Storey, N.S. Khurmi, J.C. Nicolau, C.M. Yu, D. Ardissino, A. Budaj, J. Morais, D. Montgomery, A. Himmelmann, R.A. Harrington, and L. Wallentin, *Ticagrelor effects on myocardial infarction and the impact of event adjudication in the PLATO (Platelet Inhibition and Patient Outcomes) trial.* J Am Coll Cardiol, 2014. **63**(15): p. 1493-9.

89. Mahaffey, K.W., R.A. Harrington, M. Akkerhuis, N.S. Kleiman, L.G. Berdan, B.S. Crenshaw, B.E. Tardiff, C.B. Granger, I. DeJong, M. Bhapkar, P. Widimsky, R. Corbalon, K.L. Lee, J.W. Deckers, M.L. Simoons, E.J. Topol, R.M. Califf, and P.I. for the, *Systematic adjudication of myocardial infarction end-points in an international clinical trial.* Current Controlled Trials in Cardiovascular Medicine, 2001. **2**(4): p. 180-186.

90. Mahaffey, K.W., J.L. Wampole, A. Stebbins, L.G. Berdan, D. McAfee, T.L. Rorick, J.K. French, N.S. Kleiman, C.M. O'Connor, E.A. Cohen, C.B. Granger, and P.W. Armstrong, *Strategic lessons from the clinical event classification process for the Assessment of Pexelizumab in Acute Myocardial Infarction (APEX-AMI) trial.* Contemp Clin Trials, 2011. **32**(2): p. 178-87.

91. Kirwan, B.A., J. Lubsen, S. de Brouwer, N. Danchin, A. Battler, A. Bayes de Luna, P.H. Dunselman, S. Glasser, P.J. Koudstaal, G. Sutton, F.J. van Dalen, and P.A. Poole-Wilson, *Diagnostic criteria and adjudication process both determine published event-rates: the ACTION trial experience.* Contemp Clin Trials, 2007. **28**(6): p. 720-9.

92. McGarvey, L.P., S. Magder, D. Burkhart, S. Kesten, D. Liu, R.C. Manuel, and D.E. Niewoehner, *Cause-specific mortality adjudication in the UPLIFT(R) COPD trial: findings and recommendations.* Respir Med, 2012. **106**(4): p. 515-21.

93. McGarvey, L.P., M. John, J.A. Anderson, M. Zvarich, and R.A. Wise, *Ascertainment of cause-specific mortality in COPD:*

operations of the TORCH Clinical Endpoint Committee. Thorax, 2007. **62**(5): p. 411-5.

94. Wise, R.A., P.R. Kowey, G. Austen, A. Mueller, N. Metzdorf, A. Fowler, and L.P. McGarvey, *Discordance in investigator-reported and adjudicated sudden death in TIOSPIR.* ERJ Open Research, 2017. **3**(1): p. 00073-2016.
95. Bolland, M.J., A. Barber, R.N. Doughty, A. Grey, G. Gamble, and I.R. Reid, *Differences between self-reported and verified adverse cardiovascular events in a randomised clinical trial.* BMJ Open, 2013. **3**(3).
96. US Food and Drug Administration. *Guidance for Industry. Diabetes Mellitus - Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes*. 2008 [Accessed 30 July 2019]; Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071627.pdf).
97. European Medicines Agency. *Committee for Medicinal Products for Human Use - Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus*. 2012 [Accessed 30 July 2019]; Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129256.pdf).
98. Easton, J.D., H. Denison, S.R. Evans, M. Knutsson, P. Amarenco, G.W. Albers, P. Ladenvall, K. Minematsu, C.A. Molina, Y. Wang, K.L. Wong, and S.C. Johnston, *Estimated treatment effect of ticagrelor versus aspirin by investigator-assessed events compared with judgement by an independent event adjudication committee in the SOCRATES trial.* Int J Stroke, 2019: p. 1747493019851282.
99. Internet Stroke Center. *Modified Rankin Scale (mRS)*. 2011 [cited 2019]; Available from: http://www.strokecenter.org/wp-content/uploads/2011/08/modified_rankin.pdf.
100. Quinn, T., *Improving outcome assessment for clinical trials in stroke*. 2010, University of Glasgow.

101. McArthur, K.S., *Improving efficiency in stroke trials: an exploration of methods to improve the use of the modified Rankin Scale in acute stroke trials*. 2014, University of Glasgow.
102. McArthur, K.S., P.C.D. Johnson, T.J. Quinn, P. Higgins, P. Langhorne, M.R. Walters, C.J. Weir, J. Dawson, and K.R. Lees, *Improving the Efficiency of Stroke Trials.* Stroke, 2013. **44**(12): p. 3422.
103. McArthur, K., Y. Fan, Z. Pei, and T. Quinn, *Optimising outcome assessment to improve quality and efficiency of stroke trials.* Expert Rev Pharmacoecon Outcomes Res, 2014. **14**(1): p. 101-11.
104. Lopez-Cancio, E., M. Salvat, N. Cerda, M. Jimenez, J. Codas, L. Llull, S. Boned, L.M. Cano, B. Lara, C. Molina, E. Cobo, A. Davalos, T.G. Jovin, and J. Serena, *Phone and Video-Based Modalities of Central Blinded Adjudication of Modified Rankin Scores in an Endovascular Stroke Trial.* Stroke, 2015. **46**(12): p. 3405-10.
105. Berge, E., R.A.-S. Salman, H.B. van der Worp, C. Stapf, P. Sandercock, N. Sprigg, M.R. Macleod, P.J. Kelly, P.J. Nederkoorn, G.A. Ford, M. Arnold, E. Berge, E. Diez-Tejedor, D. Jatuzis, P.J. Kelly, D.W. Krieger, P.J. Nederkoorn, P. Sandercock, C. Stapf, C. Weimar, G.A. Ford, and R.A.-S. Salman, *Increasing value and reducing waste in stroke research.* The Lancet Neurology, 2017. **16**(5): p. 399-408.
106. Chandler, J., J.P. Higgins, J.J. Deeks, C. Davenport, and M.J. Clarke, *Cochrane Handbook for Systematic Reviews of Interventions*, ed. J. Higgins, et al. 2017: Cochrane.
107. Antman, E.M., J. Lau, B. Kupelnick, F. Mosteller, and T.C. Chalmers, *A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatments for Myocardial Infarction.* JAMA, 1992. **268**(2): p. 240-248.

108. Montori, V.M., N.L. Wilczynski, D. Morgan, and R.B. Haynes, *Optimal search strategies for retrieving systematic reviews from Medline: analytical survey.* BMJ, 2005. **330**(7482): p. 68.
109. Stevinson, C. and D.A. Lawlor, *Searching multiple databases for systematic reviews: added value or diminishing returns?* Complement Ther Med, 2004. **12**(4): p. 228-32.
110. Zheng, M.H., X. Zhang, Q. Ye, and Y.P. Chen, *Searching additional databases except PubMed are necessary for a systematic review.* Stroke, 2008. **39**(8): p. e139; author reply e140.
111. Bramer, W.M., M.L. Rethlefsen, J. Kleijnen, and O.H. Franco, *Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study.* Systematic Reviews, 2017. **6**(1): p. 245.
112. Scherer, R.W., J.J. Meerpohl, N. Pfeifer, C. Schmucker, G. Schwarzer, and E. von Elm, *Full publication of results initially presented in abstracts.* Cochrane Database of Systematic Reviews, 2018(11).
113. Godin, K., J. Stapleton, S.I. Kirkpatrick, R.M. Hanning, and S.T. Leatherdale, *Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada.* Systematic reviews, 2015. **4**: p. 138-138.
114. Adams, J., F.C. Hillier-Brown, H.J. Moore, A.A. Lake, V. Araujo-Soares, M. White, and C. Summerbell, *Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies.* Systematic Reviews, 2016. **5**(1): p. 164.
115. Saleh, A.A., M.A. Ratajeski, and M. Bertolet, *Grey Literature Searching for Health Sciences Systematic Reviews: A Prospective Study of Time Spent and Resources Utilized.* Evidence based library and information practice, 2014. **9**(3): p. 28-50.

116. Haddaway, N.R., A.M. Collins, D. Coughlin, and S. Kirk, *The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching.* PLoS One, 2015. **10**(9): p. e0138237.
117. Hopewell, S., M. Clarke, C. Lefebvre, and R. Scherer, *Handsearching versus electronic searching to identify reports of randomized trials.* Cochrane Database Syst Rev, 2007(2): p. Mr000001.
118. Richards, D., *Handsearching still a valuable element of the systematic review.* Evid Based Dent, 2008. **9**(3): p. 85.
119. Crowther, M., W. Lim, and M.A. Crowther, *Systematic review and meta-analysis methodology.* Blood, 2010. **116**(17): p. 3140.
120. Morrison, A., J. Polisena, D. Husereau, K. Moulton, M. Clark, M. Fiander, M. Mierzwinski-Urban, T. Clifford, B. Hutton, and D. Rabb, *The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies.* Int J Technol Assess Health Care, 2012. **28**(2): p. 138-44.
121. Neimann Rasmussen, L. and P. Montgomery, *The prevalence of and factors associated with inclusion of non-English language studies in Campbell systematic reviews: a survey and meta-epidemiological study.* Systematic Reviews, 2018. **7**(1): p. 129.
122. Kwon, Y., M. Lemieux, J. McTavish, and N. Wathen, *Identifying and removing duplicate records from systematic review searches.* Journal of the Medical Library Association : JMLA, 2015. **103**(4): p. 184-188.
123. Qi, X., M. Yang, W. Ren, J. Jia, J. Wang, G. Han, and D. Fan, *Find Duplicates among the PubMed, EMBASE, and Cochrane Library Databases in Systematic Review.* PLOS ONE, 2013. **8**(8): p. e71838.
124. Borah, R., A.W. Brown, P.L. Capers, and K.A. Kaiser, *Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry.* BMJ Open, 2017. **7**(2): p. e012545.

125. Mateen, F.J., J. Oh, A.I. Tergas, N.H. Bhayani, and B.B. Kamdar, *Titles versus titles and abstracts for initial screening of articles for systematic reviews.* Clinical epidemiology, 2013. **5**: p. 89-95.
126. Robson, R.C., B. Pham, J. Hwee, S.M. Thomas, P. Rios, M.J. Page, and A.C. Tricco, *Few studies exist examining methods for selecting studies, abstracting data, and appraising quality in a systematic review.* Journal of Clinical Epidemiology, 2019. **106**: p. 121-135.
127. Mathes, T., P. Klaßen, and D. Pieper, *Frequency of data extraction errors and methods to increase data extraction quality: a methodological review.* BMC medical research methodology, 2017. **17**(1): p. 152-152.
128. Buscemi, N., L. Hartling, B. Vandermeer, L. Tjosvold, and T.P. Klassen, *Single data extraction generated more errors than double data extraction in systematic reviews.* J Clin Epidemiol, 2006. **59**(7): p. 697-703.
129. Campbell, M., S.V. Katikireddi, A. Sowden, J.E. McKenzie, and H. Thomson, *Improving Conduct and Reporting of Narrative Synthesis of Quantitative Data (ICONS-Quant): protocol for a mixed methods study to develop a reporting guideline.* BMJ Open, 2018. **8**(2): p. e020064.
130. Popay, J., H. Roberts, A. Sowden, M. Petticrew, L. Arai, M. Rodgers, N. Britten, K. Roen, and S. Duffy, *Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme*. 2006.
131. Higgins, J., T. Lasserson, J. Chandler, D. Tovey, and R. Churchill, *Methodological expectations of cochrane intervention reviews.* London: Cochrane, 2016. **5**.
132. Melendez-Torres, G.J., A. O'Mara-Eves, J. Thomas, G. Brunton, J. Caird, and M. Petticrew, *Interpretive analysis of 85 systematic reviews suggests that narrative syntheses and meta-analyses are incommensurate in argumentation.* Research Synthesis Methods, 2017. **8**(1): p. 109-118.

133. Deeks, J.J., D.G. Altman, and M.J. Bradburn, *Statistical Methods for Examining Heterogeneity and Combining Results from Several Studies in Meta-Analysis. In Systematic Reviews in Health Care*. Systematic Reviews in Health Care: Meta-Analysis in Context, Second Edition, ed. M. Egger, G. Davey Smith, and D.G. Altman. 2001: Wiley.
134. Riley, R.D., J.P.T. Higgins, and J.J. Deeks, *Interpretation of random effects meta-analyses.* BMJ, 2011. **342**: p. d549.
135. Mantel, N. and W. Haenszel, *Statistical aspects of the analysis of data from retrospective studies of disease.* J Natl Cancer Inst, 1959. **22**(4): p. 719-48.
136. Yusuf, S., R. Peto, J. Lewis, R. Collins, and P. Sleight, *Beta blockade during and after myocardial infarction: An overview of the randomized trials.* Progress in Cardiovascular Diseases, 1985. **27**(5): p. 335-371.
137. Greenland, S. and A. Salvan, *Bias in the one-step method for pooling study results.* Statistics in Medicine, 1990. **9**(3): p. 247-252.
138. Cochran, W.G., *The combination of estimates from different experiments.* Biometrics, 1954. **10**: p. 101-129.
139. Huedo-Medina, T.B., J. Sanchez-Meca, F. Marin-Martinez, and J. Botella, *Assessing heterogeneity in meta-analysis: Q statistic or I2 index?* Psychol Methods, 2006. **11**(2): p. 193-206.
140. Higgins, J.P. and S.G. Thompson, *Quantifying heterogeneity in a meta-analysis.* Stat Med, 2002. **21**(11): p. 1539-58.
141. Higgins, J. and S. Green, *Chapter 9.5.2: Identifying and measuring heterogeneity. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011].* The Cochrane Collaboration, 2011.
142. Higgins, J.P.T., S.G. Thompson, J.J. Deeks, and D.G. Altman, *Measuring inconsistency in meta-analyses.* BMJ (Clinical research ed.), 2003. **327**(7414): p. 557-560.
143. DerSimonian, R. and N. Laird, *Meta-analysis in clinical trials.* Control Clin Trials, 1986. **7**(3): p. 177-88.

144. Borenstein, M., L.V. Hedges, J.P.T. Higgins, and H.R. Rothstein, *A basic introduction to fixed-effect and random-effects models for meta-analysis.* Research Synthesis Methods, 2010. **1**(2): p. 97-111.
145. Bown, M.J. and A.J. Sutton, *Quality Control in Systematic Reviews and Meta-analyses.* European Journal of Vascular and Endovascular Surgery, 2010. **40**(5): p. 669-677.
146. Thompson, S.G. and J.P. Higgins, *How should meta-regression analyses be undertaken and interpreted?* Stat Med, 2002. **21**(11): p. 1559-73.
147. Thornton, A. and P. Lee, *Publication bias in meta-analysis: its causes and consequences.* Journal of Clinical Epidemiology, 2000. **53**(2): p. 207-216.
148. Egger, M., G.D. Smith, M. Schneider, and C. Minder, *Bias in meta-analysis detected by a simple, graphical test.* BMJ, 1997. **315**(7109): p. 629.
149. Sterne, J.A.C., A.J. Sutton, J.P.A. Ioannidis, N. Terrin, D.R. Jones, J. Lau, J. Carpenter, G. Rücker, R.M. Harbord, C.H. Schmid, J. Tetzlaff, J.J. Deeks, J. Peters, P. Macaskill, G. Schwarzer, S. Duval, D.G. Altman, D. Moher, and J.P.T. Higgins, *Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials.* BMJ, 2011. **343**: p. d4002.
150. Sterne, J.A.C., P. Jüni, K.F. Schulz, D.G. Altman, C. Bartlett, and M. Egger, *Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research.* Statistics in Medicine, 2002. **21**(11): p. 1513-1524.
151. Ndounga Diakou, L.A., L. Trinquart, A. Hróbjartsson, C. Barnes, A. Yavchitz, P. Ravaud, and I. Boutron, *Comparison of central adjudication of outcomes and onsite outcome assessment on treatment effect estimates: Protocol.* Cochrane Database of Systematic Reviews, 2015(7).

152. Bae, J.-M., *Meta-epidemiology.* Epidemiology and health, 2014. **36**: p. e2014019-e2014019.
153. Murad, M.H. and Z. Wang, *Guidelines for reporting meta-epidemiological methodology research.* Evidence-based medicine, 2017. **22**(4): p. 139-142.
154. Fleiss, J.L., B. Levin, and M. Cho Paik, *Statistical Methods for Rates and Proportions*. 3rd ed. Wiley Series in Probability and Statistics. 2003: Wiley.
155. Goodman, L.A. and W.H. Kruskal, *Measures of association for cross classifications.* Journal of the American Statistical Association, 1954. **49**: p. 732-769.
156. Rogot, E. and I.D. Goldberg, *A proposed index for measuring agreement in test-retest studies.* Journal of Chronic Diseases, 1966. **19**(9): p. 991-1006.
157. Ranganathan, P., C.S. Pramesh, and R. Aggarwal, *Common pitfalls in statistical analysis: Measures of agreement.* Perspectives in clinical research, 2017. **8**(4): p. 187-191.
158. McHugh, M.L., *Interrater reliability: the kappa statistic.* Biochemia medica, 2012. **22**(3): p. 276-282.
159. Cohen, J., *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 1960. **20**: p. 37-46.
160. Chmura Kraemer, H., V.S. Periyakoil, and A. Noda, *Kappa coefficients in medical research.* Stat Med, 2002. **21**(14): p. 2109-29.
161. Landis, J.R. and G.G. Koch, *The Measurement of Observer Agreement for Categorical Data.* Biometrics, 1977. **33**(1): p. 159-174.
162. Cohen, J., *Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.* 1968. **70**: p. 213-220.
163. Tang, W., J. Hu, H. Zhang, P. Wu, and H. He, *Kappa coefficient: a popular measure of rater agreement.* Shanghai archives of psychiatry, 2015. **27**(1): p. 62-67.

164. Fleiss, J.L., *Measuring nominal scale agreement among many raters.* 1971. **76**: p. 378-382.
165. Shrout, P.E. and J.L. Fleiss, *Intraclass correlations: uses in assessing rater reliability.* Psychol Bull, 1979. **86**(2): p. 420-8.
166. McGraw, K.O. and S.P. Wong, *Forming inferences about some intraclass correlation coefficients.* 1996. **1**: p. 30-46.
167. Bland, J.M. and D.G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement.* Lancet, 1986. **1**(8476): p. 307-10.
168. Agresti, A., *Analysis of ordinal categorical data*. Vol. 656. 2010: John Wiley & Sons.
169. Nunn, A., P.M. Bath, and L.J. Gray, *Analysis of the Modified Rankin Scale in Randomised Controlled Trials of Acute Ischaemic Stroke: A Systematic Review.* Stroke Research and Treatment, 2016. **2016**: p. 7.
170. Zelen, M., *The analysis of several 2× 2 contingency tables.* Biometrika, 1971. **58**(1): p. 129-137.
171. Fahrmeir, L., T. Kneib, S. Lang, and B. Marx, *Regression: Models, Methods and Applications*. 2013.
172. Pitman, E.J.G., *A Note on Normal Correlation.* Biometrika, 1939. **31**(1/2): p. 9-12.
173. Zacks, S. and H. Solomon, *On testing and estimating the interaction between treatments and environmental conditions in binomial experiments: The case of two stations.* Communications in Statistics - Theory and Methods, 1976. **5**(3): p. 197-223.
174. C. Heilbron, D., *The Analysis of Ratios of Odds Ratios in Stratified Contingency Tables*. Vol. 37. 1981.
175. Moher, D., A. Liberati, J. Tetzlaff, D.G. Altman, and P.G. The, *Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement.* PLOS Medicine, 2009. **6**(7): p. e1000097.
176. Godolphin, P.J., P.M. Bath, and A.A. Montgomery, *Short email with attachment versus long email without attachment when contacting authors to request unpublished data for a systematic*

review: a nested randomised trial. BMJ Open, 2019. **9**(1): p. e025273.

177. Higgins, J. and S. Green, *Chapter 8: Assessing risk of bias in included studies. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011].* The Cochrane Collaboration, 2011.
178. Barnett, H.J.M., D.W. Taylor, M. Eliasziw, A.J. Fox, G.G. Ferguson, R.B. Haynes, R.N. Rankin, G.P. Clagett, V.C. Hachinski, D.L. Sackett, K.E. Thorpe, H.E. Meldrum, and J.D. Spence, *Benefit of Carotid Endarterectomy in Patients with Symptomatic Moderate or Severe Stenosis.* New England Journal of Medicine, 1998. **339**(20): p. 1415-1425.
179. Barnett, H.J.M., D.W. Taylor, R.B. Haynes, D.L. Sackett, S.J. Peerless, G.G. Ferguson, A.J. Fox, R.N. Rankin, V.C. Hachinski, D.O. Wiebers, and M. Eliasziw, *Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis.* N Engl J Med, 1991. **325**(7): p. 445-53.
180. VITATOPS Trial Study Group, *B vitamins in patients with recent transient ischaemic attack or stroke in the VITAmins TO Prevent Stroke (VITATOPS) trial: a randomised, double-blind, parallel, placebo-controlled trial.* Lancet Neurol, 2010. **9**(9): p. 855-65.
181. ESPRIT Study Group, *Aspirin plus dipyridamole versus aspirin alone after cerebral ischaemia of arterial origin (ESPRIT): randomised controlled trial.* Lancet, 2006. **367**(9523): p. 1665-73.
182. PROGRESS Collaborative Group, *Randomised trial of a perindopril-based blood-pressure-lowering regimen among 6,105 individuals with previous stroke or transient ischaemic attack.* Lancet, 2001. **358**(9287): p. 1033-41.
183. Bath, P.M., L.J. Woodhouse, J.P. Appleton, M. Beridze, H. Christensen, R.A. Dineen, L. Duley, T.J. England, K. Flaherty, D. Havard, S. Heptinstall, M. James, K. Krishnan, H.S. Markus, A.A. Montgomery, S.J. Pocock, M. Randall, A. Ranta, T.G. Robinson, P. Scutt, G.S. Venables, and N. Sprigg, *Antiplatelet therapy with*

aspirin, clopidogrel, and dipyridamole versus clopidogrel alone or aspirin and dipyridamole in patients with acute cerebral ischaemia (TARDIS): a randomised, open-label, phase 3 superiority trial. The Lancet, 2018. **391**(10123): p. 850-859.

184. SPS3 Study Group, *Blood-pressure targets in patients with recent lacunar stroke: the SPS3 randomised trial.* The Lancet, 2013. **382**(9891): p. 507-515.
185. Berge, E., M. Abdelnoor, P.H. Nakstad, and P.M. Sandset, *Low molecular-weight heparin versus aspirin in patients with acute ischaemic stroke and atrial fibrillation: a double-blind randomised study. HAEST Study Group. Heparin in Acute Embolic Stroke Trial.* Lancet, 2000. **355**(9211): p. 1205-10.
186. Bonati, L.H., J. Dobson, R.L. Featherstone, J. Ederle, H.B. van der Worp, G.J. de Borst, W.P.T.M. Mali, J.D. Beard, T. Cleveland, S.T. Engelter, P.A. Lyrer, G.A. Ford, P.J. Dorman, and M.M. Brown, *Long-term outcomes after stenting versus endarterectomy for treatment of symptomatic carotid stenosis: the International Carotid Stenting Study (ICSS) randomised trial.* The Lancet, 2015. **385**(9967): p. 529-538.
187. Giugliano, R.P., C.T. Ruff, E. Braunwald, S.A. Murphy, S.D. Wiviott, J.L. Halperin, A.L. Waldo, M.D. Ezekowitz, J.I. Weitz, J. Špinar, W. Ruzyllo, M. Ruda, Y. Koretsune, J. Betcher, M. Shi, L.T. Grip, S.P. Patel, I. Patel, J.J. Hanyok, M. Mercuri, and E.M. Antman, *Edoxaban versus Warfarin in Patients with Atrial Fibrillation.* New England Journal of Medicine, 2013. **369**(22): p. 2093-2104.
188. Hosomi, N., Y. Nagai, T. Kohriyama, T. Ohtsuki, S. Aoki, T. Nezu, H. Maruyama, N. Sunami, C. Yokota, K. Kitagawa, Y. Terayama, M. Takagi, S. Ibayashi, M. Nakamura, H. Origasa, M. Fukushima, E. Mori, K. Minematsu, S. Uchiyama, Y. Shinohara, T. Yamaguchi, and M. Matsumoto, *The Japan Statin Treatment Against Recurrent Stroke (J-STARS): A Multicenter, Randomized, Open-label, Parallel-group Study.* EBioMedicine, 2015. **2**(9): p. 1071-8.

189. Johnston, S.C., P. Amarenco, G.W. Albers, H. Denison, J.D. Easton, S.R. Evans, P. Held, J. Jonasson, K. Minematsu, C.A. Molina, Y. Wang, and K.S.L. Wong, *Ticagrelor versus Aspirin in Acute Stroke or Transient Ischemic Attack.* New England Journal of Medicine, 2016. **375**(1): p. 35-43.
190. Jovin, T.G., A. Chamorro, E. Cobo, M.A. de Miquel, C.A. Molina, A. Rovira, L. San Román, J. Serena, S. Abilleira, M. Ribó, M. Millán, X. Urra, P. Cardona, E. López-Cancio, A. Tomasello, C. Castaño, J. Blasco, L. Aja, L. Dorado, H. Quesada, M. Rubiera, M. Hernandez-Pérez, M. Goyal, A.M. Demchuk, R. von Kummer, M. Gallofré, and A. Dávalos, *Thrombectomy within 8 Hours after Symptom Onset in Ischemic Stroke.* New England Journal of Medicine, 2015. **372**(24): p. 2296-2306.
191. Weimar, C., K. Bilbilis, J. Rekowski, T. Holst, F. Beyersdorf, M. Breuer, M. Dahm, A. Diegeler, A. Kowalski, S. Martens, F.W. Mohr, J. Ondrášek, B. Reiter, P. Roth, R. Seipelt, M. Siggelkow, G. Steinhoff, A. Moritz, M. Wilhelmi, G. Wimmer-Greinecker, H.-C. Diener, H. Jakob, C. Ose, A. Scherag, and S.C. Knipp, *Safety of Simultaneous Coronary Artery Bypass Grafting and Carotid Endarterectomy Versus Isolated Coronary Artery Bypass Grafting: A Randomized Clinical Trial.* Stroke, 2017. **48**(10): p. 2769-2775.
192. Ranta, A., S. Dovey, M. Weatherall, D. Dea, J. Gommans, and M. Tilyard, *Cluster randomized controlled trial of TIA electronic decision support in primary care.* Neurology, 2015. **84**(15): p. 1545.
193. Johnston, S.C., J.D. Easton, M. Farrant, W. Barsan, R.A. Conwit, J.J. Elm, A.S. Kim, A.S. Lindblad, and Y.Y. Palesch, *Clopidogrel and Aspirin in Acute Ischemic Stroke and High-Risk TIA.* New England Journal of Medicine, 2018. **379**(3): p. 215-225.
194. Godolphin Peter, J., M. Bath Philip, A. Algra, E. Berge, M. Brown Martin, J. Chalmers, L. Duley, M. Eliasziw, J. Gregson, P. Greving Jacoba, J. Hankey Graeme, N. Hosomi, S.C. Johnston, E. Patsko, A. Ranta, M. Sandset Per, J. Serena, C. Weimar, A.

Montgomery Alan, and on behalf of the Adjudicating Outcomes in Stroke Trials Collaboration, *Outcome Assessment by Central Adjudicators Versus Site Investigators in Stroke Trials.* Stroke, 2019. **0**(0): p. STROKEAHA.119.025019.

195. Schulz, K.F., D.G. Altman, and D. Moher, *CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials.* BMJ, 2010. **340**: p. c332.
196. Selph, S.S., A.D. Ginsburg, and R. Chou, *Impact of contacting study authors to obtain additional data for systematic reviews: diagnostic accuracy studies for hepatic fibrosis.* Systematic Reviews, 2014. **3**: p. 107-107.
197. Teunis, T., S.P.F.T. Nota, and J.H. Schwab, *Do Corresponding Authors Take Responsibility for Their Work? A Covert Survey.* Clinical Orthopaedics and Related Research®, 2015. **473**(2): p. 729-735.
198. Wolfe, N., P.C. Gøtzsche, and L. Bero, *Strategies for obtaining unpublished drug trial data: a qualitative interview study.* Systematic reviews, 2013. **2**(1): p. 31.
199. Hicks, K.A., K.W. Mahaffey, R. Mehran, S.E. Nissen, S.D. Wiviott, B. Dunn, S.D. Solomon, J.R. Marler, J.R. Teerlink, A. Farb, D.A. Morrow, S.L. Targum, C.A. Sila, M.T.T. Hai, M.R. Jaff, H.V. Joffe, D.E. Cutlip, A.S. Desai, E.F. Lewis, C.M. Gibson, M.J. Landray, A.M. Lincoff, C.J. White, S.S. Brooks, K. Rosenfield, M.J. Domanski, A.J. Lansky, J.J.V. McMurray, J.E. Tcheng, S.R. Steinhubl, P. Burton, L. Mauri, C.M. O'Connor, M.A. Pfeffer, H.M.J. Hung, N.L. Stockbridge, B.R. Chaitman, and R.J. Temple, *2017 Cardiovascular and Stroke Endpoint Definitions for Clinical Trials.* Circulation, 2018. **137**(9): p. 961-972.
200. Bath, P.M., A. Houlton, L. Woodhouse, N. Sprigg, J. Wardlaw, and S. Pocock, *Statistical analysis plan for the 'Efficacy of Nitric Oxide in Stroke' (ENOS) trial.* Int J Stroke, 2014. **9**(3): p. 372-4.
201. Bath, P.M., L. Woodhouse, P. Scutt, K. Krishnan, J.M. Wardlaw, D. Bereczki, N. Sprigg, E. Berge, M. Beridze, V. Caso, C. Chen, H. Christensen, R. Collins, A. El Etribi, A.C. Laska, K.R. Lees, S.

Ozturk, S. Phillips, S. Pocock, H.A. de Silva, S. Szatmari, and S. Utton, *Efficacy of nitric oxide, with or without continuing antihypertensive treatment, for management of high blood pressure in acute stroke (ENOS): a partial-factorial randomised controlled trial.* Lancet, 2015. **385**(9968): p. 617-28.

202. *Glyceryl trinitrate vs. control, and continuing vs. stopping temporarily prior antihypertensive therapy, in acute stroke: rationale and design of the Efficacy of Nitric Oxide in Stroke (ENOS) trial (ISRCTN99414122).* Int J Stroke, 2006. **1**(4): p. 245-9.
203. Hesse, K., R.L. Fulton, A.H. Abdul-Rahim, and K.R. Lees, *Characteristic Adverse Events and Their Incidence Among Patients Participating in Acute Ischemic Stroke Trials.* Stroke, 2014. **45**(9): p. 2677-2682.
204. Godolphin, P.J., A.A. Montgomery, L.J. Woodhouse, D. Bereczki, E. Berge, R. Collins, E. Díez-Tejedor, J. Gommans, K.R. Lees, S. Ozturk, S. Phillips, S. Pocock, K. Prasad, S. Szatmari, Y. Wang, P.M. Bath, N. Sprigg, and E.I. on behalf of the, *Central adjudication of serious adverse events did not affect trial's safety results: Data from the Efficacy of Nitric Oxide in Stroke (ENOS) trial.* PLOS ONE, 2018. **13**(11): p. e0208142.
205. Duley, L., J. Dorling, A. Pushpa-Rajah, S.J. Oddie, C.W. Yoxall, B. Schoonakker, L. Bradshaw, E.J. Mitchell, and J.A. Fawke, *Randomised trial of cord clamping and initial stabilisation at very preterm birth.* Archives of Disease in Childhood - Fetal and Neonatal Edition, 2018. **103**(1): p. F6.
206. Abbott, J., J. Berrington, U. Bowler, E. Boyle, J. Dorling, N. Embleton, E. Juszczak, A. Leaf, L. Linsell, S. Johnson, K. McCormick, W. McGuire, T. Roberts, B. Stenson, and G. The Sift Investigators, *The Speed of Increasing milk Feeds: a randomised controlled trial.* BMC Pediatrics, 2017. **17**(1): p. 39.
207. Griffiths, J., P. Jenkins, M. Vargova, U. Bowler, E. Juszczak, A. King, L. Linsell, D. Murray, C. Partlett, M. Patel, J. Berrington, J. Dorling, N.D. Embleton, P.T. Heath, S. Oddie, W. McGuire, S.

Ainsworth, E. Boyle, P. Clarke, S. Craig, K. Johnson, H. Mactier, T. Scorrer, M. Ledwidge, I. Story, G. Holder, P. Ohadike, S. Ellis, R. Vaikute, G. Gowda, H. Yates, S. Garg, E. Pilling, C. Roehr, D. Batra, D. Gibson, M. Johnson, Y. Kumar, D. Bartle, C. Peters, D. Quine, R. Gupta, J. Matthes, N. Kennea, P. Reynolds, R. Geethanath, S. Janakiraman, V. Vasu, and C.M. Manjunatha, *Enteral lactoferrin supplementation for very preterm infants: a randomised placebo-controlled trial.* The Lancet, 2019. **393**(10170): p. 423-433.

208. Partlett, C., L. Linsell, O. Hewer, E. Juszczak, and J. Dorling, *The impact of blinded endpoint review on the incidence of primary short-term outcomes in the SIFT trial*. 2017, BMC. p. Article: P414.

209. Ashby, R.L., R. Gabe, S. Ali, P. Saramago, L.H. Chuang, U. Adderley, J.M. Bland, N.A. Cullum, J.C. Dumville, C.P. Iglesias, A.R. Kang'ombe, M.O. Soares, N.C. Stubbs, and D.J. Torgerson, *VenUS IV (Venous leg Ulcer Study IV) - compression hosiery compared with compression bandaging in the treatment of venous leg ulcers: a randomised controlled trial, mixed-treatment comparison and decision-analytic model.* Health Technol Assess, 2014. **18**(57): p. 1-293, v-vi.

210. Ormerod, A.D., K.S. Thomas, F.E. Craig, E. Mitchell, N. Greenlaw, J. Norrie, J.M. Mason, S. Walton, G.A. Johnston, and H.C. Williams, *Comparison of the two most commonly used treatments for pyoderma gangrenosum: results of the STOP GAP randomised controlled trial.* BMJ : British Medical Journal, 2015. **350**: p. h2958.

211. Patsko, E., P.J. Godolphin, K.S. Thomas, T. Hepburn, E.J. Mitchell, F.E. Craig, P.M. Bath, and A.A. Montgomery, *Investigating the effect of independent, blinded digital image assessment on the STOP GAP trial.* Trials, 2017. **18**(1): p. 53.

212. Wright-Hughes, A., E. Graham, D. Cottrell, and A. Farrin, *Routine hospital data - is it good enough for trials? An example using England's Hospital Episode Statistics in the SHIFT trial of*

Family Therapy vs. Treatment as Usual in adolescents following self-harm. Clinical trials (London, England), 2018. **15**(2): p. 197-206.

213. Kilburn, L.S., M. Aresu, J. Banerji, P. Barrett-Lee, P. Ellis, and J.M. Bliss, *Can routine data be used to support cancer clinical trials? A historical baseline on which to build: retrospective linkage of data from the TACT (CRUK 01/001) breast cancer trial and the National Cancer Data Repository.* Trials, 2017. **18**(1): p. 561-561.
214. Mc Cord, K.A., R. Al-Shahi Salman, S. Treweek, H. Gardner, D. Strech, W. Whiteley, J.P.A. Ioannidis, and L.G. Hemkens, *Routinely collected data for randomized trials: promises, barriers, and implications.* Trials, 2018. **19**(1): p. 29.

Appendix A – PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	95
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	97
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	56-62
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	98-100
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	99
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	99-100
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	100-104
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	101

Section/topic	#	Checklist item	Reported on page #
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	104-106
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	106-108
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	107
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	108-109
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	110
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	109-111
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	110-111
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	112-113
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	114-115
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	119-123

Section/topic	#	Checklist item	Reported on page #
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	124-126
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	126-129 & 132-133
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	128
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	127-132
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	134-135
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	136
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	136-137
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	NA

Appendix B – Full search strategy

Search Strategy (06 November 2018):

EMBASE (1669):

1. exp randomised controlled trial/
2. exp animal/ not human/
3. exp stroke/
4. ((adjudicat* or endpoint or outcome or review or classification or central or event)adj2(adjudicat* or committee or panel or review)).mp.
5. 1 not 2
6. 3 and 4 and 5

MEDLINE (249):

1. exp Randomized controlled trial/
2. exp animals/ not exp humans/
3. exp Stroke/
4. ((adjudicat* or endpoint or outcome or review or classification or central or event)adj2(adjudicat* or committee or panel or review)).mp.
5. 1 not 2
6. 3 and 4 and 5

PsycINFO (249):

1. exp *Clinical Trials/
2. exp *Drug Therapy/
3. exp *Evidenced Based Practice/
4. exp *Treatment Effectiveness Evaluation/
5. exp *cerebrovascular accidents/
6. ((adjudicat* or endpoint or outcome or review or classification or central or event)adj2(adjudicat* or committee or panel or review)).mp.
7. 1 or 2 or 3 or 4
8. 5 and 6 and 7

CENTRAL (2395):

"stroke" AND (adjudicat* OR "advisory committee" OR "outcome assessment" OR "endpoint committee" OR "endpoint review" OR "outcome committee" OR "outcome panel" OR "outcome review" OR "review committee" OR "review panel" OR "classification committee" OR " classification panel" OR "central committee" OR "central panel" OR "central review" OR "event committee" OR "event panel" OR “end point committee” OR “end-point committee”)

Web of Science (1477):

"stroke" AND (adjudicat* OR "advisory committee" OR "outcome assessment" OR "adjudication committee" OR "adjudication panel" OR "endpoint committee" OR "endpoint review" OR "endpoint adjudication"

OR "outcome adjudication" OR "outcome committee" OR "outcome panel" OR "outcome review" OR "review committee" OR "review panel" OR "classification committee" OR " classification panel" OR "central adjudication" OR "central committee" OR "central panel" OR "central review" OR "event adjudication" OR "event committee" OR "event panel")

Google Scholar (300 (only first 300 selected)):

(“randomised” OR “randomized”) AND "stroke" AND ("adjudication" OR "outcome assessment" OR "event panel" OR "endpoint committee" OR "review committee" OR "central review” OR “event committee”)

Appendix C – Data Extraction Form

Data extraction form:

Review title: A comparison of independent adjudication and site assessment of stroke outcomes and its impact on treatment effect estimates in randomised trials: systematic review and meta-analysis

Trial ID:	**Reference ID(s) and type:**	
Data extractor:	**Date form completed:**	**Year of main study publication:**
Main study title:		
Main study author(s):		
Main study reference:		
Additional reference material sourced: *(ID and type)*		

1. Study eligibility:

Study characteristics:	**Eligibility criteria**	**Eligibility criteria met?** **Yes** **No** **Unclear**	**Location in reference**
Type of study:	Randomised Controlled Trial	☐ ☐ ☐	
Participants:	Patients being treated for stroke or being given an intervention to prevent stroke	☐ ☐ ☐	
Intervention method:	Independent adjudication of the primary outcome	☐ ☐ ☐	
Comparison method:	Site-assessment of the same primary outcome	☐ ☐ ☐	
INCLUDE: ☐		**EXCLUDE:** ☐	
Reason for exclusion:			
Notes:			

DO NOT PROCEED IF STUDY IS EXCLUDED FROM REVIEW

1. Study information:

Type of study design: Parallel: ☐ Factorial: ☐ Cluster: ☐ Crossover: ☐ Other: ☐		Notes:
Type of stroke trial: Acute: ☐ Rehab: ☐ Pri-prevention: ☐ Sec-prevention: ☐ Other: ☐		Notes:
Describe the intervention(s)?		**Type of intervention:**
Describe the comparator(s)?		**Type of comparator:**
Trial setting:		
Describe setting:		No. of countries: No. of trial centres:
Sample size:		Notes:
Total sample size:	Sample size in each trial arm:	
Patients randomised: Total randomised:	Number randomised to each trial arm:	Notes:
Description of outcome adjudicated and assessed by site investigators:		**Review outcome group:**

2. Outcome information:

Type of outcome:	Notes:	**Objectiveness of outcome:**
Binary ☐ Ordinal ☐ Nominal ☐ Continuous ☐		
Number and description of comparisons: *(multiple comparisons expected if factorial or 3-arm trial)*		
Description of statistic, measure used and method of calculating C.I: *(e.g. Odds ratio, OLR and Normal)* *Comparison 1:* *Comparison 2:* *Comparison 3:*		
Site assessed treatment effect, N, standard error and/or confidence interval: *Comparison 1:* *Comparison 2:* *Comparison 3:*		
Adjudicator-assessed treatment effect, N, standard error and/or confidence interval: *Comparison 1:* *Comparison 2:* *Comparison 3:*		

IF THERE ARE MULTIPLE COMPARISONS, THE REMAINDER OF THIS FORM MAY BE COMPLETED MULTIPLE TIMES IF THE ADJUDICATOR INFORMATION/RISK OF BIAS DIFFERS FOR SOME COMPARISONS

3. Adjudicator information:

Comparison:

Blinding status of adjudicators:		Notes:
To treatment: Blinded ☐ Not Blinded ☐	To site-assessment Blinded ☐ Not Blinded ☐	
Number and profession of adjudicators:		Notes:
Number:	Profession (tick all that apply): Doctor ☐ Nurse ☐ Other HP ☐ HP Ot ☐	☐
Describe training, if any, that adjudicators received:		
Describe what information was provided to adjudicators: *(medical notes, results of tests/procedures performed, case report form, video footage)*		
How were disagreements between adjudicators dealt with: *(single adjudicator so no disagreement, consensus decision, additional adjudicator reviewed information)*		

4. Assessment of risk of bias:

Comparison:					
Domain:	**Risk of bias:**			**Support for judgement:** *(Include direct quotes where available with explanatory comments)*	**Location in reference:** *(Ref & Page)*
	Low	**High**	**Unclear**		
Random sequence generation					
Allocation concealment					
Blinding of participants and personnel					
Blinding of outcome assessment					
Incomplete outcome data					
Selective outcome reporting?					
Selection of participants to adjudicate					
Notes:					

Appendix D – Data Extraction Manual

Data extraction manual:

The aim of this manual is to provide explanation about what needs to be extracted and where to find out the required information. Each stroke trial will have a unique Study ID, and there may be multiple references included from each study. These will be grouped prior to data extraction, so only one form needs to be completed per study.

You may need to search through multiple references to extract the required information. If appropriate, then you may search for additional material (protocol paper, statistical analysis plan, secondary analysis paper etc.) in order to complete the data extraction. If any additional material is sourced then please take a record of this material in the space provided. There are *Notes* boxes throughout which enable you to write any further information relevant to the field. Please make a note of any missing or unclear information as you complete the data extraction form and record this data in the Missing information log.

Throughout the form there are a number of greyed out boxes. These are for central completion, so there is no need to extract this information. If you have any queries then do not hesitate to contact me on Peter.Godolphin@nottingham.ac.uk or 0115 82 31563.

Trial ID: Each study (stroke trial) is given a unique ID (T01)

Reference ID and type: Each trial may contain a number of references, these are given a unique ID (R0001) from the order stored in EndNote (alphabetical by author). The reference type will be details such as main study paper, protocol paper, secondary analysis etc.

Main study: The main study refers to the main stroke trial that the references belongs to. If this is unclear to determine or only a secondary analysis paper/meta-analysis is obtained then treat the most appropriate paper as the *main study*.

Main study title: Give the name of the trial (or acronym if available).

Main study author(s): Record the primary/corresponding author – if it is unclear, multiple authors can be recorded.

Additional reference material sourced: List any additional references sourced that were not identified in the systematic review and their type (see Reference ID and type), for these additional references their ID will be allocated in order of study ID (A0001).

----

1. Study eligibility:

Eligibility criteria: The listed criteria are a summary of what is needed for each study characteristic, consult the protocol for more details. If the study meets ALL of these criteria then it can be included in this review, otherwise exclude.

Location in reference: Please list the reference ID and the page/line/section where this evidence was found – sufficient detail should be given so it is easily identified.

Reason for exclusion: Make a note of the reason why you did not think that this particular study should be included in this review.

Page 2 onwards need only be completed if the study is being included in the review.

----

2. Study information:

Type of study design: The type of study design should be given and described. If the study is a factorial trial then please describe the factors, if the study is cluster-randomised please describe what the clusters are and if the study is a crossover trial then please describe the order of treatments (e.g. ABC).

Type of stroke trial: Record the type of stroke trial and describe details about the population (e.g. Primary prevention stroke trial for patients who have already suffered a myocardial infarction). Note that Pri-prevention and Sec-Prevention refer to Primary and Secondary Prevention respectively.

Trial setting: Please indicate whether the trial was national or international, and give more specific information (e.g. Europe or India) if it is easily obtainable.

---

3. Outcome information:

Type of outcome: Specify is the outcome is binary, categorical (ordinal/nominal) or continuous. Describe the categories if appropriate.

Number and description of comparisons: If there are multiple comparisons of this outcome then describe what they are and assign each comparison a number (C1, C2 etc.) which will be used in the next three questions to differentiate between them.

Description of statistic, measure used and method of calculating C.I: Please describe the statistical analysis carried out for the primary outcome, what kind of measure the primary outcome was and how the confidence interval was calculated (e.g. Unadjusted Cox's proportional hazards model, relative risk reduction was the measure used which was (1-HR) x 100, C.I was 95% using Normal approximation.).

Treatment effect, N, standard error and/or confidence interval: Give the number of participants used in the analysis, the treatment effect and either the standard error or the confidence interval (with % e.g. 95%, 99%). If easy to obtain, then additional summary information is preferred (unadjusted mean split by treatment group if adjusted mean difference). If the outcome is binary then also provide the number of events in each treatment group.

---

4. Adjudicator information:

Blinding status of adjudicators: Record if the adjudicators were aware which treatments participants received, or if they were aware what the site assessment was for the outcome adjudicated

Profession of adjudicators: If the profession is not listed, then describe.

----

5. Assessment of risk of bias:

Consult the Cochrane handbook for more guidance on risk of bias (http://handbook.cochrane.org/chapter_8/table_8_5_d_criteria_for_judging_risk_of_bias_in_the_risk_of.htm)

Random sequence generation:
Low – the investigators describe a random component in the sequence generation process. For example, referring to a random-number table, minimisation etc.
High – the investigators describe a non-random component in the sequence generation process. For example, sequence generated by odd or even date of birth.
Unclear – insufficient information provided to make a decision

Allocation concealment:
Low – participants and investigators enrolling participants could not foresee assignment because an appropriate method was used to conceal allocation. For example, central allocation including telephone/web-based allocation.
High – it was possible for participants or investigators enrolling participants to know the allocation. For example, using envelopes without appropriate safeguards
Unclear – Not sufficient information reported

Blinding of participants and personnel:
Low – participants and personnel blinded to treatment allocation (or “double blind” trial)
High – participants and personnel not blinded to treatment allocation
Unclear – it is not clear from the records if the participants and personnel were appropriately blinded to treatment allocation

Blinding of outcome assessment:
Low – outcome assessors blinded to treatment allocation (or “double blind” trial)
High – outcome assessors not blinded to treatment allocation
Unclear – it is not clear from the records if the outcome assessors were appropriately blinded to treatment allocation

Incomplete outcome data:
Low – no missing outcome data, or missing outcome data not related to true outcome, or missing outcome data balanced across groups with similar reasons for loss of information, or missing data has been imputed using appropriate methods.
High – Missing data likely to be related to true outcome and there is an imbalance between extent of missing data between groups or per-protocol analysis carried out with a substantially reduced population or an inappropriate method of imputation.
Unclear – Insufficient reporting or study doesn’t address this outcome

Selective outcome reporting?
Low – The study protocol is available and all pre-specified outcomes have been reported in the pre-specified way.
High – Not all of the studies pre-specified outcomes have been reported or one of the outcomes was not pre-specified or one of the pre-specified outcomes is reported using a different method to that described in the protocol.
Unclear – Insufficient information to judge

Selection of participants to adjudicate:
Low – All participants, all suspected participants, or only site reported participants identified by blinded assessors are selected for adjudication
High – Only site reported participants which are identified by unblinded assessors are selected for adjudication
Unclear – It is not clear how participants are selected for adjudication

Appendix E – Missing Information Log

Missing information log:

Review title: A comparison of independent adjudication and site assessment of stroke outcomes and its impact on treatment effect estimates in randomised trials: systematic review and meta-analysis

Study ID:
Missing/Unclear Information:
Study ID:
Missing/Unclear Information:
Study ID:
Missing/Unclear Information:
Study ID:
Missing/Unclear Information:
Study ID:
Missing/Unclear Information:
Study ID:
Missing/Unclear Information:

Appendix F – Data Sharing and Use Agreement

DATA SHARING AND USE AGREEMENT

Between:

<The sponsor>; and

The University of Nottingham a body incorporated in England and Wales by Royal Charter with registration number RC000664 whose principal offices are at University Park, Nottingham NG7 2RD ("Nottingham")

Background:

This agreement governs the terms on which access will be granted to the trial data detailed below.

"Approved Project"	**The project entitled "Adjudicated Outcomes in Stroke Trials (AOST)" as more particularly described in the approved research proposal attached with this agreement**
"Data Requester"	**Peter Godolphin**
"Data"	**Details of dataset** **The data outlined below from all participants over the duration of their participation in the <trial name> trial:** • **Data to reproduce the trial primary analysis** • **Adjudicated outcome data** • **Unadjudicated site data**
"Data Custodian"	**Name of CI**

Agreed terms:

1. In consideration of the obligations accepted by Nottingham for <sponsor institution>, <sponsor institution> grants to Nottingham for the duration of the Approved Project a non-exclusive, personal and non-transferable licence to use the Data only for the Approved Project.

2. Nottingham shall store and use the Data strictly in accordance with the terms specified in this agreement and all applicable laws and regulations.

3. Nottingham undertakes to keep the Data confidential and to use or disclose the Data only to the extent necessary for the Approved Project or as otherwise required by law or a court of competent jurisdiction.

4. Nottingham will not attempt to identify from the Data, or communicate with, any of the trial participants. Nottingham shall not link or attempt to link the Data to other data or information without prior written consent from <sponsor institution>. In the event that Nottingham performs any act that results in the Data becoming personal data in Nottingham's hands, Nottingham shall (a) immediately notify <sponsor institution>; (b) process the data as <sponsor institution>'s data processor strictly in accordance with <sponsor institution>'s instructions; and (c) indemnify <sponsor institution> against any claims, liabilities, costs, expenses, damages and losses suffered or incurred by <sponsor institution> arising out of or in connection with any third party claims brought against <sponsor institution> arising from the breach by Nottingham of its obligations under this clause.

5. Nottingham shall not transfer, disclose or make available the Data, in whole or in part, to any third party.

6. Nottingham will implement all appropriate administrative, organisational, physical and technical measures to prevent the unauthorised access, use or disclosure of the Data other than as permitted by this agreement. In particular (and without limiting the generality of the preceding wording) Nottingham shall not make physical or electronic copies of the Data except to the extent reasonably necessary for the Approved Project and shall ensure that any copy of the Data stored on movable media (including laptops) is password-protected and fully encrypted and that any copy of the Data stored on networked or non-networked hard drives is properly protected with firewall and controlled access permissions.

7. Nottingham shall keep a record of where each copy of the Data is stored and who has access to the Data. Nottingham shall provide <sponsor institution> with a copy of this record on request.

8. Nottingham will notify <sponsor institution> as soon as reasonably practicable in the event that it becomes aware of any unauthorised access, use or disclosure of the Data.

9. Nottingham shall restrict access to the Data to the Data Requester and those staff and students comprising the Data Requester's research team named in the approved research proposal appended to this agreement, and will ensure that those staff and students are aware of and comply with the terms of this agreement. Nottingham is at all times responsible for such staff and students' compliance with the obligations set out in this agreement.

10. The Data remains the property of <sponsor institution> at all times and no transfer of or licence of rights in the Data from <sponsor institution> to

Nottingham shall be implied other than the licence set out in clause 1 of this agreement. This agreement does not restrict <sponsor institution>'s rights in the Data, including (without limitation) its right to make the Data available to other institutions or to publish any document relating to or including the Data.

11. The results of the Approved Project shall belong to Nottingham. Nottingham shall notify <sponsor institution> of the results. <sponsor institution> will keep the results confidential and not disclose the results to any third party prior to their publication without the prior written consent of Nottingham.

12. Nottingham will refer to the <trial name> study and acknowledge the <trial name> investigators, funder and where appropriate other contributors to the Data in any publication arising from the use of the Data, in a form to be agreed with <sponsor institution> (and taking into account any requirements of the funder of the <trial name> study).

13. <sponsor institution> may terminate this agreement with immediate effect by written notice to Nottingham in the event that Nottingham:

    - breaches any remediable term of this agreement and has failed to remedy such breach within sixty (60) days after receipt of a written request from <sponsor institution> to remedy the breach; or

    - commits any material or persistent breach of this agreement; or

    - enters into administration or some other formal insolvency procedure, enters into liquidation, is wound up, or ceases to carry on business.

14. As soon as reasonably practicable on completion of the Approved Project or the termination of this agreement (whichever is earlier), Nottingham shall permanently delete all copies of the Data from its hard drives and movable media and destroy all physical copies of the Data.

15. Although <sponsor institution> tries to take reasonable care in collecting, compiling and sharing data, trial data is often contributed by third parties and cannot be guaranteed to be free from errors, omissions or inaccuracies. Accordingly, <sponsor institution> will not be liable to Nottingham for any loss arising from any reliance placed on the Data by Nottingham.

16. Notwithstanding anything else contained in this agreement and except for the indemnity in Clause 4, either Party will in no event be liable to the other under or otherwise in connection with the subject matter of this agreement for:

    - loss of profits, contracts, business, anticipated savings or reputation, the cost of any substitute supply, in each case whether direct or indirect; or

    - indirect or consequential loss.

    For the avoidance of doubt, nothing in this clause 14 or elsewhere in this agreement shall exclude either Party's liability for personal injury or death

caused by its negligence, for fraud, or for any other liability that cannot be limited or excluded by law.

17. English law shall apply to this agreement, and the English courts shall have exclusive jurisdiction over any matter relating to it.

SIGNED for and on behalf of

<SPONSOR INSTITUTION>

Name:

Position:

Date:

Acknowledged by the Data Custodian:

Name:

Position:

Date:

SIGNED for and on behalf of

THE UNIVERSITY OF NOTTINGHAM

Name: Alan Montgomery

Position: Professor of Medical Statistics and Clinical Trials

Date:

Acknowledged by the Data Requester:

Name: Peter Godolphin

Position: NIHR Doctoral Research Fellow

Date:

Appendix G: CONSORT Checklist

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	138
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	139
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	140-142
	2b	Specific objectives or hypotheses	140
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	144-145
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	NA
Participants	4a	Eligibility criteria for participants	142-144
	4b	Settings and locations where the data were collected	142-144

Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	145-147 & Appendix H & Appendix I
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	147-148
	6b	Any changes to trial outcomes after the trial commenced, with reasons	NA
Sample size	7a	How sample size was determined	148
	7b	When applicable, explanation of any interim analyses and stopping guidelines	NA
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	145
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	145
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	145
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	145
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	144

	11b	If relevant, description of the similarity of interventions	NA
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	148-149
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	148-149
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	150-152
	13b	For each group, losses and exclusions after randomisation, together with reasons	NA
Recruitment	14a	Dates defining the periods of recruitment and follow-up	7/8
	14b	Why the trial ended or was stopped	NA
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	150
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	152-154
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	152-154
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	NA
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	153

Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	NA
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	158-159
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	156-158
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	156-158
Other information			
Registration	23	Registration number and name of trial registry	NA
Protocol	24	Where the full trial protocol can be accessed, if available	NA
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	NA

Appendix H: Short Invitation email

Dear (ENTER NAME HERE),

I am writing to invite you to join an international collaboration investigating the usefulness of adjudicating outcomes in stroke trials. We would like to include your trial (ENTER TRIAL NAME) in a systematic review. The collaboration is led by myself, Professor Philip Bath and Professor Alan Montgomery, all at the University of Nottingham.

It is currently unclear in what circumstances adjudication is required to produce unbiased estimates of effect. In this project, entitled Adjudicating Outcomes in Stroke Trials (AOST), we wish to compare and combine analyses of the primary outcome in trials using adjudicated and unadjudicated data.

We hope that you will be interested and willing to collaborate in this project. The AOST protocol is attached for further information. We look forward to hearing from you and thank you for taking the time to read this invitation.

Best wishes,

Peter Godolphin

NIHR Doctoral Research Fellow

Appendix I: Long invitation email

Dear (ENTER NAME HERE),

I am writing to invite you to join an international collaboration investigating the usefulness of adjudicating outcomes in stroke trials. We would like to include your trial (ENTER TRIAL NAME) in a systematic review. The collaboration is led by myself, Professor Philip Bath and Professor Alan Montgomery, all at the University of Nottingham.

As you are aware, a range of outcomes are commonly adjudicated in stroke trials. It is currently unclear in what circumstances adjudication is required to produce unbiased estimates of effect. In this project, entitled Adjudicating Outcomes in Stroke Trials (AOST), we aim to address this question by combining information about adjudication from a number of stroke trials.

We wish to compare analyses of the primary outcome in trials using adjudicated and unadjudicated data. Therefore to collaborate in this project, we would like you to provide one of two things, either:

(1) Summary results from your trial (outcome in each arm, between group measure of effect and measure of variance/confidence interval) for the primary outcome based on: (i) adjudicated data, and (ii) unadjudicated data

(2) Anonymised individual participant data for key variables from your study to allow us to carry out analysis of the primary

outcome using adjudicated and unadjudicated data. For this we will need:

- Adjudicated primary outcome
- Unadjudicated primary outcome
- Treatment arm, covariates or other data used in primary analysis
- Statistical analysis plan or other details of primary analysis

Data, and any accompanying materials, can be shared in any spreadsheet format (Excel, Stata, SAS etc.) and we will re-format the data to ensure that it fits our database. Data can be shared using your preferred method, for example secure download, email or CD/memory stick.

We aim to publish the results in a high impact peer-reviewed journal. All collaborators will see the final results and manuscript for comment/interpretation, and the results will be published under the banner of AOST with collaborators listed by name. Our group at Nottingham has considerable experience in working with individual participant data from large, international randomised trials (including projects such as BASC, DISC, NEMAS, OAST, OT, OA-COG), and of protecting the interests of the collaborators who have shared their data/results. Your data/results will not be used for any purpose other than relating to AOST.

We believe that this study will address the lack of evidence about when adjudication of outcomes may or may not be needed in stroke trials to

ensure robust results. We hope that you will be interested and willing to collaborate by sharing your data or results and (should you wish to) helping in interpretation of the results and their publication. Please let me know if you would like a copy of the AOST protocol.

We look forward to hearing from you and thank you for taking the time to read this invitation.

Best wishes,

Peter Godolphin

NIHR Doctoral Research Fellow

Appendix J: Adjudicating Outcomes in Stroke Trials Protocol

A comparison of central adjudication and site assessment of outcomes in randomised stroke trials: summary protocol

The 'Adjudicating Outcomes in Stroke Trials' Collaboration (AOST)

Peter Godolphin[1,2]

NIHR Doctoral Research Fellow

Philip Bath[1]

Stroke Association Professor of Stroke Medicine

Alan Montgomery[2]

Professor of Medical Statistics and Clinical Trials

[1]*Stroke Trials Unit, Division of Clinical Neuroscience, University of Nottingham*

[2]*Nottingham Clinical Trials Unit, University of Nottingham*

AIM

The aim of this research programme is to answer the question:
What are the benefits and costs of adjudicating outcomes in randomised stroke trials?

BACKGROUND

Central adjudication in clinical trials refers to the re-evaluation of clinical trial data by independent experts in order to provide information about key outcomes or safety issues. The use of adjudication in clinical trials

has become commonplace in many clinical areas, in particular cardiovascular disease with over 80% of trials in this area reported to include adjudication[1]. Central adjudication can be conducted by a single expert or a panel of experts, and the adjudicators typically assess primary outcomes determined by the assessors at trial sites to decide whether or not they agree, or to categorise them in a specific way.

Description of the problem or issue

Detection bias can be an issue in randomised trials, especially if outcome assessors (often at trial sites) are not blinded to treatment, with the potential to exaggerate treatment effect estimates[2]. As adjudicators are blinded to treatment allocation whenever possible[3], central adjudication can help to reduce bias and play a vital role in ensuring trial results are valid.

Furthermore, adjudication can help to improve accurate classification of events as well as controlling for differential misclassification[4]. This is especially important in trials with few events, where a small degree of misclassification can have a large impact on study findings[5, 6]. Many researchers who report on adjudication mention the reassurance an adjudication committee brings being “of substantial importance”[7, 8]. A simulation of the adjudication of outcome assessment in stroke trials has found that increasing the reliability of modified Rankin scores using adjudication could result in 20% reductions in sample size[9].

However, it is often the case that adjudicators only have access to limited patient data, and never physically interact with the patients, unlike many assessors at trial sites[4]. The view held by some is that, although experts in their field, the opinions of the adjudicators may not hold more validity than those of the hospital staff who have overseen the patient and made a decision based on a number of reasons, rather than solely numerical data or scans[5].

Also, adjudication can be an expensive process and may be an inefficient use of limited research funding. In 2012, £56M of government funding was allocated to stroke research and it is the aim of the Stroke Association to double the research funding assigned to stroke by 2020. Therefore, it is important to identify scenarios where adjudication is vital, to ensure new trials produce high quality evidence, and where adjudication is not required.

The current evidence

A Cochrane systematic review[11] compared central adjudication of subjective binary outcomes with outcome assessment at trial sites and investigated their effect on treatment estimates. It found 47 studies eligible to be included in a meta-analysis and concluded that, on average, there was no difference in treatment effect estimates between adjudicators and site assessment (Ratio of Odds Ratios: 1.00, 95% C.I: 0.97 to 1.04).

Secondary analyses investigating the impact of the adjudication process on a single trial have found mixed results, with some indicating

that adjudication is a vital process[12, 13] and others showing that it had no discernible result on any trial findings[7, 8].

There has been a substantial amount of research on the effect of adjudication of cardiovascular outcomes, in particular myocardial infarction[4, 5, 12, 14-16]. However, few studies have investigated how adjudication impacts on stroke outcomes. Ninomiya et al.[7] evaluated the effects of the end point adjudication process on a stroke trial, where adjudicators analysed type of stroke and cause of death. The study concluded that 'the end point adjudication process … had no substantive impact on the effect estimates obtained from the trial' and that 'with clinical research costs escalating, it is a requisite on researchers to ensure that scarce resources are applied as effectively as possible' as well as asking for 'formal quantitative estimates of the likelihood of an adjudication process influencing trial conclusions'.

Why it is important to carry out this research programme

A range of outcomes are commonly adjudicated in stroke trials. Thus, we need to address the lack of collective evidence around the importance of adjudicating outcomes in stroke trials. It is currently unclear in what circumstances adjudication is required to produce unbiased estimates of effect.

Furthermore, adjudication may be costly and can result in a significant delay in a clinical trial. The data are sent to an adjudicator who reviews it and sends back their decision; this is then entered and combined with the site information[4, 5]. If there is disagreement between the

adjudicator and site, then a second adjudicator is often asked to evaluate the data to determine the final outcome. However, trials use a variety of methods to deal with disagreements[17]. In cardiovascular trials, with many thousands of patients, this process can be time-consuming and expensive. Adjudication of bleeding outcomes in an international thromoprophylaxis trial was estimated to cost £80,000 for 1,737 suspected events[18, 19]. In a large cardiovascular trial, adjudication was carried out on all suspected cases of nonfatal myocardial infarction, nonfatal stroke and cardiovascular death at an estimated cost of over £500,000 for 7,911 diagnoses[8].

Therefore research is needed to understand the role adjudication has in stroke trials. In three linked Work Packages (WPs), this research programme will aim to address the following question: *What are the benefits and costs of adjudicating outcomes in stroke trials?*

METHODS

WP1: Systematic review and meta-analyses

Aim: To assess the effects of adjudicator-assessed outcomes versus site-assessed outcomes on the treatment effect estimates for stroke trials

This review will include randomised stroke trials that have adjudicated their primary outcome. Adjudicated secondary outcomes, including adverse events will be collected to enable further exploration of the importance of adjudication for different types of outcomes. For detailed

methodology of the systematic review, please consult the PROSPERO[20].

To assess the impact of adjudication on trial conclusions we will calculate the ratio of treatment effects for each trial. This will be calculated using the treatment estimate using site-assessed outcomes versus the treatment estimate using adjudicator-assessed outcomes.

The decision to pool data will be based on an assessment of heterogeneity between the included studies of the same outcome. If the data are not suitable to be pooled than a narrative description of the included studies will be given. If the data are suitable to be combined then it will be pooled using a separate meta-analysis for each different type of outcome.

The data needed for this review (and subsequent work packages) are either (1) summary results from the eligible trial, or (2) individual participant data in order to compare analysis of the primary outcome using adjudicated and unadjudicated data.

(1) Results from the eligible trial:

a. Outcome in each arm (if binary/categorical N (%)) using unadjudicated and adjudicated data

b. Between group measure of effect using both unadjudicated and adjudicated data

c. Measure of variance or C.I for between group measure of effect using both unadjudicated and adjudicated data

(2) Individual participant data:

a. Adjudicated primary outcome

b. Unadjudicated primary outcome

c. Treatment arm, covariates or other data used in primary analysis

d. Statistical analysis plan or other details of primary analysis

WP2: Statistical simulation to investigate the effect on bias and precision of adjudicated versus unadjudicated outcomes

Aim: To investigate how adjudication impacts on the bias and precision of unadjudicated outcomes and how this affects the sample size needed for a trial

A series of statistical simulations will be created based on real trial data received as part of WP1 and will cover a variety of trial types, settings and outcomes. Simulations will be adapted to introduce both differential and non-differential misclassification. This will allow an exploration of how adjudication of outcomes impacts on the bias and precision of unadjudicated treatment effects in these differing situations. These findings can help to inform how the adjudication process affects sample size when there is differential, non-differential or no misclassification present.

Further analysis will explore how varying agreement between adjudicator-assessed and site-assessed outcomes will impact on the

sample size of stroke trials. Different adjudication methods will be simulated, including adjudication of site-reported events and adjudication of all participants. This will help to provide an understanding of how different adjudication methods can directly impact on statistical power, hence sample size.

WP3: Cost-effectiveness

Aim: To investigate the cost-effectiveness of adjudication in stroke trials

Collaborators who have provided results/data used in WP1/WP2 will be asked if they would like to participate in this work package. In WP3 we will send out a questionnaire with questions designed to provide information about the cost of adjudication. Examples of questions are as follows:

- How many people were in the adjudication committee and what were their professional backgrounds?
- How long did each committee member spend adjudicating?
- Was anyone else involved in the adjudication process, e.g. data entry clerk, statistician, trial co-ordinator, programmer?
- Was any payment made to adjudicators for the adjudication?

Standardised estimates of costs will be assigned centrally using this information, and these costs will be in today's prices regardless of the age of the trial to allow comparison.

Effectiveness will be measured by the number of outcomes corrected in the adjudication process. The costs associated with adjudication will be

estimated from direct costs: fee paid to adjudicators, time spent by adjudicators, travel costs, other staff costs etc.

Data Sharing

Collaborators will be sharing data with the AOST collaboration and with no other group or individual. Specifically, AOST will not share data with any third party. Data will be stored on a secure desktop computer in a locked room. The computer is backed up daily to a central University server.

Publication

Data will not be used for any purpose other than to do with AOST. The results from this research programme will be published under the banner of AOST with the collaborators listed by name. Each trial will be allocated a number of author slots depending on trial size. Trials of size 1-99 will be allocated one slot, trials of size 100-999 will be allocated two slots, trials of size 1000-9999 will be allocated three slots and trials that recruit at least 10000 participants will be allocated four slots.

The manuscripts will be written by a 'Writing Committee' and then distributed to all Collaborators for comment, interpretation, changes, additions etc. Any collaborators that contribute to the publication will be listed in the Writing Committee. The results will be submitted to major stroke/trial methodology conference(s) for presentation, and submitted for publication in major quality peer-reviewed journal(s).

An example of our publication policy is as follows:

Banner: AOST

Writing committee: Peter Godolphin, contributing author 1, contributing author 2, …, Philip Bath, Alan Montgomery

Collaborators: Trial 1 (232 participants) Author 1, Author 2, Trial 2 (2094 participants) Author 3, Author 4, Author 5, etc.

Experience of the Nottingham Stroke Trials Unit

Nottingham Stroke Trials Unit has considerable experience in working will individual patient data within collaborations such as AOST. This includes projects such as BASC (blood pressure in acute stroke collaboration), DISC (dipyridamole), NEMAS (NXY-059, animal data), OAST (optimising the analysis of stroke trials), OT (community occupational therapy), OA-COG (optimising the analysis of cognition), and of protecting the interests of the collaborators who shared their data.

References

1. Dechartres A, Boutron I, Roy C, Ravaud P. Inadequate planning and reporting of adjudication committees in clinical trials: recommendation proposal. Journal of clinical epidemiology. 2009;62(7):695-702.

2. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded

outcome assessors. BMJ. 2012 Feb 27;344:e1119. PubMed PMID: 22371859. Epub 2012/03/01. eng.

3. Walter SD, Cook DJ, Guyatt GH, King D. Outcome assessment for clinical trials: How many adjudicators do we need? Controlled Clinical Trials. 1997 1997/02/01;18(1):27-42.

4. Granger CB, Vogel V, Cummings SR, Held P, Fiedorek F, Lawrence M, et al. Do we need to adjudicate major clinical events? Clinical Trials. 2008;5(1):56-60.

5. Pogue J, Walter SD, Yusuf S. Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs. Clinical Trials. 2009;6(3):239-51.

6. Serebruany VL, Fortmann SD. International central adjudication committee in the PLATO trial: independent body of experts or friendly family picnic? Expert review of cardiovascular therapy. 2015;13(8):867-9.

7. Ninomiya T, Donnan G, Anderson N, Bladin C, Chambers B, Gordon G, et al. Effects of the end point adjudication process on the results of the Perindopril Protection Against Recurrent Stroke Study (PROGRESS). Stroke. 2009;40(6):2111-5.

8. Hata J, Arima H, Zoungas S, Fulcher G, Pollock C, Adams M, et al. Effects of the endpoint adjudication process on the results of a randomised controlled trial: the ADVANCE trial. PloS one. 2013;8(2):e55807.

9. McArthur K, Fan Y, Pei Z, Quinn T. Optimising outcome assessment to improve quality and efficiency of stroke trials. Expert review of pharmacoeconomics & outcomes research. 2014;14(1):101-11.

10. Luengo-Fernandez R, Leal J, Gray A. UK research spend in 2008 and 2012: comparing stroke, cancer, coronary heart disease and dementia. BMJ Open. 2015 April 1, 2015;5(4).

11. Ndounga Diakou LA, Trinquart L, Hróbjartsson A, Barnes C, Yavchitz A, Ravaud P, et al. Comparison of central adjudication of outcomes and onsite outcome assessment on treatment effect estimates. The Cochrane Library.

12. Mahaffey KW, Harrington RA, Akkerhuis M, Kleiman NS, Berdan LG, Crenshaw BS, et al. Disagreements between central clinical events committee and site investigator assessments of myocardial infarction endpoints in an international clinical trial: review of the PURSUIT study. Current controlled trials in cardiovascular medicine. 2001 Jul 17;2(4):187-94. PubMed PMID: 11806794. Pubmed Central PMCID: PMC57750. Epub 2002/01/25. Eng.

13. Näslund U, Grip L, Fischer-Hansen J, Gundersen T, Lehto S, Wallentin L. The impact of an end-point committee in a large multicentre, randomized, placebo-controlled clinical trial: results with and without the end-point committee's final decision on end-points. European heart journal. 1999;20(10):771.

14. Mahaffey KW, Harrington RA, Akkerhuis M, Kleiman NS, Berdan LG, Crenshaw BS, et al. Systematic adjudication of myocardial infarction end-points in an international clinical trial. Current controlled trials in cardiovascular medicine. 2001 Jul 17;2(4):180-6. PubMed PMID: 11806793. Pubmed Central PMCID: PMC57749. Epub 2002/01/25. Eng.

15. Mahaffey KW, Wampole JL, Stebbins A, Berdan LG, McAfee D, Rorick TL, et al. Strategic lessons from the clinical event classification process for the Assessment of Pexelizumab in Acute Myocardial Infarction (APEX-AMI) trial. Contemporary clinical trials. 2011;32(2):178-87.

16. Mahaffey KW, Held C, Wojdyla DM, James SK, Katus HA, Husted S, et al. Ticagrelor effects on myocardial infarction and the impact of event adjudication in the plato (platelet inhibition and patient outcomes) trial. Journal of the American College of Cardiology. 2014;63(15):1493-9.

17. Stuck AK, Fuhrer E, Limacher A, Méan M, Aujesky D. Adjudication-related processes are underreported and lack standardization in clinical trials of venous thromboembolism: a systematic review. Journal of clinical epidemiology. 2014;67(3):278-84.

18. Arnold DM, Lauzier F, Rabbat C, Zytaruk N, Cash BB, Clarke F, et al. Adjudication of bleeding outcomes in an international thromboprophylaxis trial in critical illness. Thrombosis research. 2013;131(3):204-9.

19. Saunders L, Zytaruk N, Southon J, Cash BB, Lutz K, Clarke F, et al., editors. Methods Center Adjudication Costs For 5 Morbidity Outcomes In A Thromboprophylaxis Trial. AMERICAN JOURNAL OF RESPIRATORY AND CRITICAL CARE MEDICINE; 2011: Am Thoracic Soc.

20. Godolphin P, Montgomery A, Bath P. A comparison of independent adjudication and site assessment of outcomes and its impact on treatment effect estimates in randomised stroke trials: systematic review and meta-analyses 2017. Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42017056731

Appendix K – Cost of Adjudication Questionnaire

Cost of Adjudication Questionnaire:

Introduction

This short questionnaire asks for information about the resources required for outcome adjudication in your trial. Please try to answer the questions as fully and as accurately as possible. These data will form the basis of a planned paper on cost-effectiveness of adjudication which you will be invited to co-author in due course.

If you have any questions please contact Peter Godolphin on peter.godolphin@nottingham.ac.uk.

Email address

Enter email

1. Trial name

Dropdown option

Adjudication

This section collects information about who the adjudicators were, time required to adjudicate events, and adjudicators' approximate salary.

2. How many events were adjudicated in your trial?

Free text (must be number)

3. For the events in Q2, was each event adjudicated…

Once

Twice

Other (please specify)

4. Please list the profession of the adjudicators (tick all that apply)

Medically qualified health professional

Other health professional

Not a health professional

Other (please specifiy)

5. For professions ticked in Q4, what was the approximate total number of adjudications completed by each? (Please give a number or a percentage)

Free text

6. On average, how much time did each adjudication require? (Please take into account accessing files, asking for further information and actual time adjudicating)

Free text

7. For professions ticked in Q3, what is the approximate annual salary in today's costs for each? (Please include currency)

Free text

8. Additional comments or information about this section

Free text

Disagreements

This section is about disagreements between adjudicators, how long these took to resolve and who it involved. If disagreements were not possible (each event only adjudicated once) or there were no disagreements between adjudicators in your trial then please skip this section.

9. How many adjudicated events resulted in a disagreement between adjudicators? (Please give a number, or percentage)

Free text

10. How many people were involved in resolving each disagreement?

2

3

Other (please specify)

11. On average, how much time did each adjudicator spend resolving each disagreement? (Please take into account time for further adjudication, discussion or consensus decision)

Free text

12. Additional comments or information about this section

Free text

Comments

13. Please give any further comments, or information about any costs not covered in the above questions.

Free text