OPTIMISING THE ANALYSIS OF STROKE TRIALS

By Laura Jayne Gray MSc. BSc. (Hons)

Thesis submitted to the University of Nottingham

For the degree of Doctorate of Philosophy

May 2008

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

CHAPTER 2

GENERAL METHODS

CHAPTER 3

RESULTS

COMPARISON OF UNIVARIATE STATISTICAL METHODS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Most large acute stroke trials have shown no treatment effect. Functional outcome is routinely used as the primary outcome in stroke trials. This is usually analysed using a binary analysis, e.g. death or dependency versus independence. This project assessed which statistical approaches are most efficient in analysing functional outcome data from stroke trials.

Fifty five data sets from 47 (54,173 patients) completed randomised trials were assessed. Re-analysing this data with a variety of statistical approaches showed that methods which retained the ordinal nature of functional outcome data were statistically more efficient than those which collapsed the data into two or more groups. Ordinal logistic regression, t-test, robust rank test, bootstrapping the difference in mean rank, or the Wilcoxon test are recommended. When assessing sample size, using ordinal logistic regression to analyse data instead of a binary outcome can reduce the sample size needed for a given power by 28%. Ordinal methods may not be appropriate for trials of treatments which not only increase the proportion of patients having a good outcome but also have an increase in hazard, such as thrombolytics.

Adjusting the analysis performed for prognostic factors can have an additional effect on sample size. Re-analysing data from 23 stroke trials (25,674 patients), where covariate data was supplied, showed that ordinal logistic regression adjusted for age, sex and baseline stroke severity reduced the sample size needed for a given statistical power by around 37%. Alternatively trialists could increase the statistical power to find an effect for a given sample

size, as it is argued that stroke trials have been too small and therefore underpowered.

Stroke prevention trials also routinely collect binary data, e.g. stroke/no stroke. Converting this data into ordinal outcomes, e.g. fatal stroke/non-fatal stroke/no stroke and analysing these with a method which takes into account the ordered nature of the data also increases the statistical power to find a treatment effect. This method also provides additional information on the effect of treatment on the severity of events.

Using ordinal methods of analysis may improve the design and statistical analysis of both acute and stroke prevention trials. Smaller trials would help stroke developments by reducing time to completion, study complexity, and financial expense.

ACKNOWLEDGEMENTS

The work in this thesis was carried out under the supervision of Professor Philip Bath, Division of Stroke Medicine, University of Nottingham; he was the architect of both the acute and prevention OAST projects. I would like to thank him for his unending optimism in the OAST project and for his encouragement throughout.

I would also like to thank Timothy Collier, Professor Stuart Pocock and Dr James Carpenter, all from the Medical Statistics Unit, London School of Hygiene and Tropical Medicine, for their statistical and programming guidance with this project. I acknowledge Dr Chamila Geeganage, Division of Stroke Medicine, University of Nottingham, for collating the data used in the OAST prevention analysis, and her for her help with this part of the project; and Dr Gillian Sare Division of Stroke Medicine, University of Nottingham, for her medical input and help in updating the HRT example used in Chapter 7. I would like to thank all of the OAST collaborators for sharing the data with this project and reviewing draft manuscripts.

I would also like to show appreciation to all my friends, past and present, at the Division of Stroke Medicine for their support and encouragement over the years. Margaret Adrian, Dr Claire Allen, Alison Columbine, Fiona Hammonds, Dr Gillian Sare, Sally Utton and Graham Watson for taking the time to proof read my thesis. Wim Clarke, Sharon Ellender (for tea making), Dr Timothy England, Tanya Payne, Hazel Sayers, Beverly Whysall and Dr Mark Willmot for their advice, encouragement and friendship. And last but not least my research side-kick Dr Nikola Sprigg for many years of friendship both inside and outside of work and for always being on the other end of the phone!

I would also like to thank Dr Kelly Handley for her statistical ear and pedantic nature and Katie Pike for reading and commenting on my thesis.

A big thank you to all my family and friends....too many to mention, you know who you are.

LIST OF COLLABORATORS

The following collaborators provided individual patient data from their trial, and commented on draft manuscripts:

Abciximab: H Adams (USA), K Dougherty (USA); W Hacke (Germany),

ASK: G Donnan (Australia)

ASSIST 07 & 10: S Davis (Australia)

ATLANTIS A & B: G Albers, S Hamilton (USA)

BEST Pilot & Main: D Barer (UK)

Citicoline 1, 7, 10, 18: A Davalos (Spain)

Corr: S Corr (UK)

Dover Stroke Unit: P Langhorne (UK)

DCLHb: P Koudstaal, R Saxena (Netherlands)

DESTINY: E Juettler, W Hacke (Germany)

Ebselen: T Yamaguchi (Japan)

ECASS II: W Hacke, E Bluhmki (Germany)

Factor VII: S Mayer (USA), K Begtrup (Denmark)

FISS: R Kay (Hong Kong)

FOOD 3: M Dennis (UK)

Gilbertson: L Gilbertson (UK)

INWEST: N-G Wahlgren, N Ahmed (Sweden)

IST: P Sandercock (UK)

Kuopio Stroke Unit: J Sivenius (Finland)

Logan: P Logan (UK)

MAST-I: L Candelise (Italy), J Wardlaw (UK)

Minocycline: Y Lampl, M Boaz (Israel)

Newcastle Stroke Unit: H Rodgers (UK)

NINDS: J Marler (USA)

Nottingham Stroke Unit: N Lincoln, P Berman (UK)

Parker: C Parker (UK)

RANNTAS I & II, STIPAS, TESS I & II: P Bath (UK), B Musch (USA)

Statin withdrawal: J Castillo (Spain)

Walker 1 & 2: M Walker (UK)

Young: J Young, A Forster (UK)

We thank the patients who took part in these studies, and the trialists who shared their data.

LIST OF ABBREVIATIONS

ADL	Activities of daily living
AF	Atrial fibrillation
ANOVA	Analysis of variance
BI	Barthel Index
CEA	Carotid endarterectomy
CHD	Coronary heart disease
CI	Confidence interval
CI	Chief Investigator
CONSORT	Consolidated Standards of Reporting Trials
CT	Computer tomography
CVD	Cerebrovascular disease
DCLHb	Diaspirin cross-linked haemoglobin
EADL	Extended activities of daily living
FAST	Face-Arm-Speech Test
GEE	Generalised estimating equations
GO	Global outcome
GOS	Glasgow Outcome Scale
HRT	Hormone replacement therapy
ICF	International Classification of Functioning, Disability and Health
ICIDH	International Classification of Impairments, Disabilities and Handicaps
IMPACT	International Mission for Prognosis and Clinical Trial
IPD	Individual patient data
IQR	Interquartile range
LACI	Lacunar infarction
MI	Myocardial infarction

MRI	Magnetic resonance imaging
mRS	Modified Rankin Scale
NIHSS	National Institute of Health Stroke Scale
OAST	Optimising Analysis of Stroke Trials
OHS	Oxford Handicap Scale
OR	Odds ratio
OT	Occupational therapy
PACI	Partial anterior circulation infarction
PE	Pulmonary embolism
PEG	Percutaneous endoscopic gastrostomy
PI	Principal Investigator
POCI	Posterior circulation infarction
PT	Physiotherapy
SU	Stroke unit
TACI	Total anterior circulation infarction
TIA	Transient ischaemic attack
UA	Unstable angina
VISTA	Virtual Stroke Trials Archive
VTE	Venous thromboembolism
WHO	World Health Organisation
3Q	Three Questions Outcome

To Mum, Dad, Theresa, Nicola and Dicky Mint

"I put my heart and my soul into my work, and have lost my mind in the process"

Vincent Van Gogh

CHAPTER 1

INTRODUCTION

PUBLICATIONS/PRESENTATIONS CONTRIBUTING TO THIS CHAPTER

Gray L.J, Sprigg N, Bath P.M.W, Boysen G, De Deyn P, Leys D, O'Neill D, Ringelstein EB, for the TAIST Investigators (2007) Sex differences in quality of life in stroke survivors: data from the 'Tinzaparin in Acute Ischaemic Stroke Trial' (TAIST). *Stroke. 38 (11):2960-4.*

Gray L.J, Sprigg N, Bath P.M.W, Sørensen P, Lindenstrøm E, Boysen G, De Deyn P.P, Friis P, Leys D, Marttila R, Olsson J-E, O'Neill D, Ringelstein B, MD; van der Sande J-J, Turpie A.G.G, for the TAIST Investigators (2006) Significant variation in mortality and functional outcome after acute ischaemic stroke between western countries: data from the 'Tinzaparin in Acute Ischaemic Stroke Trial' (TAIST). *Journal of Neurology, Neurosurgery and Psychiatry 77: 327-333.*

Sprigg N, **Gray L.J**, Bath P.M.W, Lindenstrøm E, Boysen G, De Deyn P.P, Friis P, Leys D, Marttila R, Olsson J-E, O'Neill D, Ringelstein B, van der Sande J-J, Turpie A.G.G, for the TAIST Investigators (2007) Stroke severity, early recovery and outcome are each related with clinical classification of stroke: data from the Tinzaparin in Acute Ischaemic Stroke Trial' (TAIST) *Journal of Neurological sciences. 254(1-2):54-9.*

Sprigg N, **Gray L.J**, Bath P.M.W, Lindenstrøm E, Boysen G, De Deyn P.P, Friis P, Leys D, Marttila R, Olsson J-E, O'Neill D, Ringelstein B, van der Sande J-J, Turpie A.G.G, for the TAIST Investigators (2006) Relationship between outcome and baseline blood pressure, pulse pressure, heart rate and rate-pressure product in acute ischaemic stroke: data from the 'Tinzaparin in Acute Ischaemic Stroke Trial' (TAIST). *Journal of Hypertension **24**(7), 1413-1417.*

Bath P.M.W, **Gray L.J** (2005) Hormone replacement therapy and subsequent stroke: a meta analysis. *British Medical Journal,* ***330****, 342*

Abstract republished in Nature Clinical Practice Cardiovascular Medicine (2005) 2, 119. Abstract republished in Journal of Neurology, Neurosurgery and Psychiatry (2005) 76, 838

Halkes P.H.A, **Gray L.J**, Bath P.M.W, Bousser M-G, Diener H-C, Guiraud-Chaumeil B, Yatsu F, Algra A, on behalf of the Dipyridamole in Stroke Collaboration (DISC) (2007) Dipyridamole plus aspirin in the secondary prevention after TIA or stroke of arterial origin: a meta analysis by risk using individual patient data from randomised trials. *Journal of Neurology, Psychiatry and Neurosurgery DOI:10.1136/jnnp.2008.143875.*

Bath P.M.W, **Gray L.J** (2007) Should Data Monitoring Committees assess efficacy when considering safety in trials in acute stroke? *International Journal of Clinical Practice. 61 (10): 1749-1755.*

1.1 INTRODUCTION

This chapter will briefly introduce the key themes of this thesis: stroke, measuring outcome, and clinical trials in stroke. Section 1.4 will review in detail the research carried into the statistical analysis of functional outcome scales so far. The final section will outline the main aims of this project.

1.2 STROKE

The World Health Organisation (WHO) define stroke as "rapidly developed clinical signs of focal or global disturbance of cerebral function, lasting more than 24 hours or until death, with no apparent cause other than of vascular origin" (WHO MONICA Project Principal Investigators, 1988). In lay terms, a stroke can be thought of as a brain attack, which comes on very suddenly. During a stroke the blood supply to part of the brain may be cut off, this loss can cause brain cells to be damaged. These damaged brain cells can affect bodily functions. For example, if damage occurs in the part of the brain which controls limb function, movement of the limb could be affected (The Stroke Association, 2008). The severity of a stroke can vary dramatically from recovery in a day to severe disability or death (Warlow, 1998). Stroke is a collective term for several types of brain injury, of which there are two main types, ischaemic (inadequate blood flow) and haemorrhagic (a bleed) (see Figure 1.1).

Ischaemic strokes are the most common type of stroke, accounting for around 85% of the total number (NHS direct, 2001). Ischaemic stroke occurs when an artery supplying blood to the brain becomes blocked, and therefore interrupts the blood supply to the brain. Brain tissue starved of blood will die (cerebral infarction).

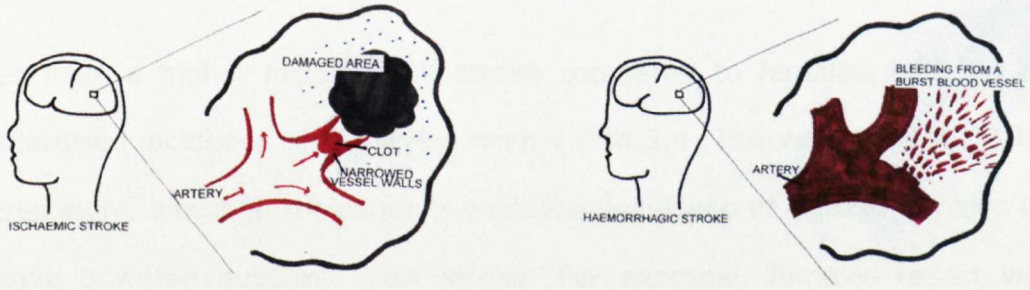
There are four causes of an ischaemic stroke:

- Embolism, where a blood clot formed in another part of the body (usually the heart) travels through the bloodstream to the brain (20%).
- Thrombosis, where a blood clot forms in a main artery leading to the brain or within the brain (50%).
- Lacunar stroke, which occurs when small vessels deep within the brain become blocked (25%).
- Other causes, such as arterial dissection, arteritis, and infective endocarditis, account for the remaining 5% of ischaemic strokes.

A haemorrhagic stroke occurs when a blood vessel in or around the brain bursts, accounting for 15% of all strokes (Bamford et al., 1990).

FIGURE 1.1

Diagrams of an ischaemic and haemorrhagic stroke, taken from http://www.strokerehabunit.ie/en/AboutStroke/DifferentTypesofStroke/



A transient ischaemic attack (TIA) is a related condition which does not fall within the definition of a stroke. It is sometimes called a 'mini-stroke' as it starts like a stroke but lasts for less than 24 hours and leaves no lasting symptoms (Warlow et al., 1996).

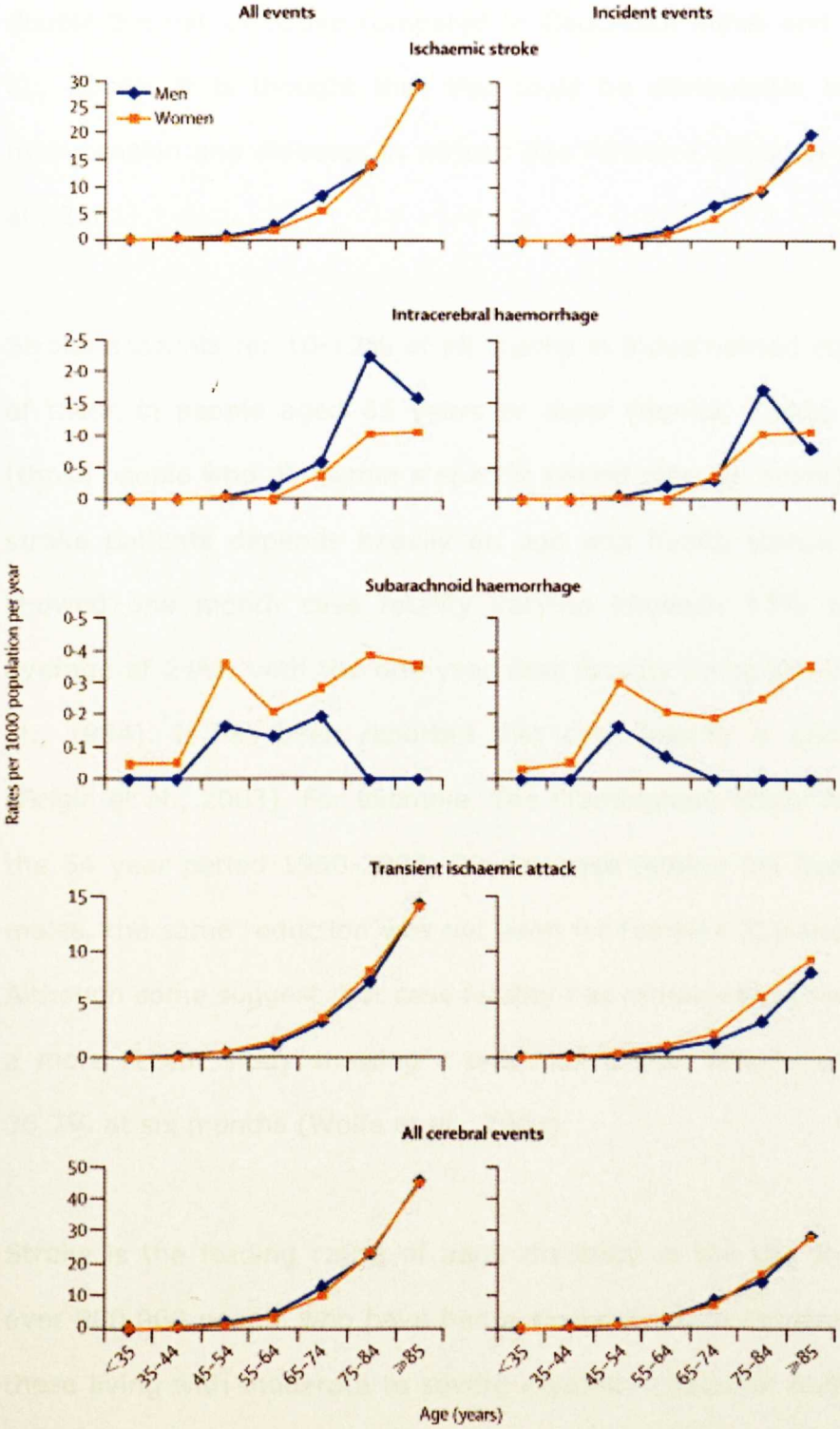
1.2.1 Epidemiology

Stroke is the third most common cause of death in the United Kingdom (UK), preceded by cancer and myocardial infarction (heart attack), with one in four men and one in five women expected to have a stroke by the age of 85 (Wolfe, 2000). Incidence measures the number of new cases in one year divided by the number at risk (Bland, 2000). The incidence of stroke rises exponentially with increasing age. Once aged over 55 years the incidence of stroke doubles with each successive decade (Wolfe, 2000), with an incidence of three per 10,000 when aged 30-40 increasing 100 fold to 300 per 10,000 when aged 80-90 (Bonita et al., 1984). Figure 1.2 shows age specific rates for cerebrovascular events taken from the "Oxford Vascular Study". This was an observational study looking at acute vascular events occurring in Oxfordshire between 2002 and 2005. This shows that the incidence of all events, apart from subarachnoid haemorrhage, increase with age for both males and females (Rothwell et al., 2005).

Males have a higher incidence of stroke compared to females, with an age-standardised incidence ratio varying from 1.2 to 2.4 (Thorvaldsen et al., 1995). Interestingly, although males have a greater incidence of stroke, females tend to have a worse outcome after stroke. For example, females report worse quality of life post stroke compared to males (Gray et al., 2007). There are many possible reasons for this difference, including higher levels of atrial fibrillation (irregular heart beat) and hypertension (high blood pressure) in females prior to their stroke (Di Carlo et al., 2003), and differences in their in-hospital care. For example, males are more likely to receive thrombolytic therapy, which is a highly efficacious clot busting treatment for ischaemic stroke (Warner Gargano et al., 2008).

FIGURE 1.2

Age-specific rates for cerebrovascular events by sex.



Reprinted from The Lancet, 366. Rothwell PM et al. Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (Oxford Vascular Study), 1773-1783., Copyright (2005), with permission from Elsevier.

Differences in incidence rates are also apparent across ethnic groups. For example, African and African-Caribbean males and females have approximately double the risk of stroke compared to Caucasian males and females (Kakar et al., 2006). It is thought that this could be attributable to higher levels of hypertension and diabetes in African and African-Caribbean patients (Sacco et al., 2001).

Stroke accounts for 10-12% of all deaths in industrialised countries, with 88% of these in people aged 65 years or older (Bonita, 1992). The case fatality (those people who die within a specific period after an event) at one month for stroke patients depends heavily on age and health status. In 1984 a study showed one month case fatality varying between 17% and 34% with an average of 24%; with the one year case fatality being around 42% (Bonita et al., 1984). It has been reported that case fatality is decreasing over time (Feigin et al., 2003). For example, The Framingham Study found that between the 54 year period 1950-2004, 30 day case fatality fell from 23% to 14% in males, the same reduction was not seen for females (Carandang et al., 2006). Although some suggest that case fatality has remained constant over time, with a more recent study showing a one month case fatality of 25.7%, rising to 36.7% at six months (Wolfe et al., 2002).

Stroke is the leading cause of adult disability in the UK. In 2005 there were over 900,000 people who have had a stroke living in England, with 300,000 of these living with moderate to severe disability (National Audit Office, 2005). A study comparing outcome after ischaemic stroke across eleven countries, found that in the UK, at six months, post stroke 21% of patients had died, 63% were still dependent on others and 37% were living in an institution (Gray et al., 2006). Those in the UK also reported greater levels of dependency and poorer

quality of life after stroke than other western countries, even after adjustment for case mix and service quality markers (Gray et al., 2006, Gray et al., 2008).

1.2.2 Symptoms

Strokes affect different people in different ways depending on the type of stroke, the area of the brain affected and the severity. The most common symptoms are: numbness or weakness of the face, arm and/or leg weakness (normally on one side of the body), confusion, difficulty speaking, difficulty with vision, dizziness and sudden severe headaches. In the late 1990's the Face-Arm-Speech Test (FAST) was developed to help rapidly identify those suffering from a stroke. This involves checking individuals for facial weakness, arm weakness and speech problems. The use of this test has been shown to increase diagnosis of stroke by paramedics (Harbison et al., 2003). The FAST test has since been advertised to the public by the Stroke Association to encourage people to ring 999 on seeing these symptoms to allow prompt care. Less common symptoms include: nausea, fever, vomiting, loss of consciousness, fainting or convulsions (Warlow et al., 1996).

1.2.3 Diagnosis

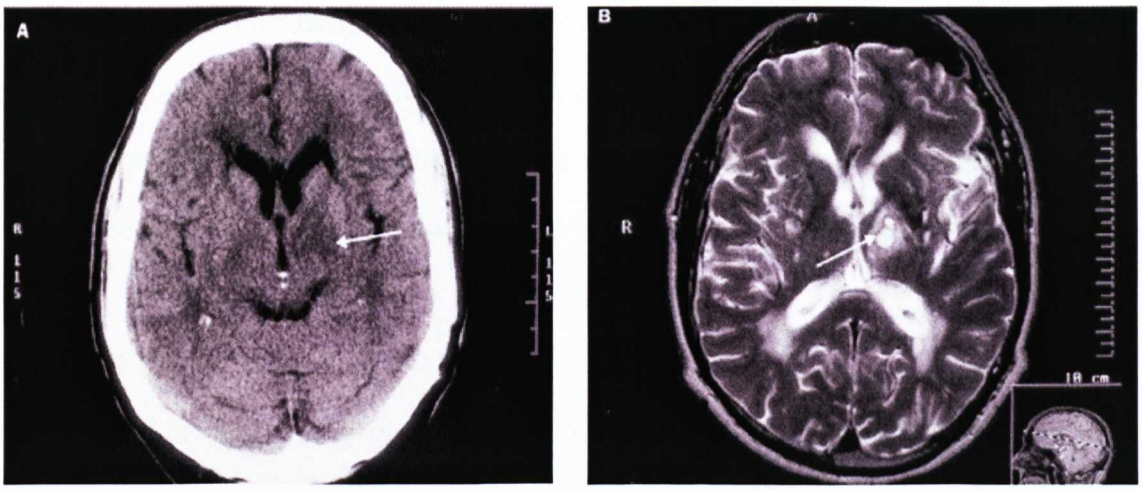
Diagnosis of stroke has three main elements; history, clinical examination and imaging. After initial stabilisation it is imperative that a history is obtained from either the patient or a relative. This is to establish the time of onset (important for treatment options), possible causes, presence of risk factors and history of any cardiac disorders (Vuadens and Bogousslavsky, 1998). The clinical examination is usually directed at confirming cardiovascular disease. The doctor will carry out a general examination (blood pressure etc) and then a full detailed neurological examination. The neurological examination will assess cranial nerves, meningeal signs, motor system, posture and gait, reflexes, coordination, sensation and cognitive function. Once the clinical examination has taken place a clinical diagnosis should have been made (de Freitas and Bogousslavsky, 1997).

Investigations are then carried out to confirm the type and cause of stroke. Imaging (either by cranial computed tomography (CT) scan or magnetic resonance imaging (MRI), see Figure 1.3 for an example of the two scan types) is the most accurate method for distinguishing between ischaemic and haemorrhagic stroke. This is important to determine as haemorrhagic strokes are treated differently (Jager, 2000). The new National Stroke Strategy for the UK states that patients with potential strokes should be imaged within 24 hours of onset (Department of Health, 2007). Scanning patients very early allows doctors to treat ischaemic strokes with a thrombolytic agent, a powerful clot busting drug which is only licensed to be given within the first three hours of stroke onset.

FIGURE 1.3

A comparison of CT and MR scan of a mild lacunar stroke.

CT Scan MRI Scan



Reprinted from The Lancet, 362. 9391, Warlow C et al. Stroke, 1211-1224., Copyright (2003), with permission from Elsevier.

1.2.4 Prognostic factors

A prognostic factor is a situation, condition, or a characteristic of a patient, that can be used to estimate the chance of recovery from a disease, or the chance of the disease recurring (i.e. the patients' prognosis). Prognostic factors which are used to assess prognosis in stroke patients include; type of stroke, stroke subtype, level of consciousness, severity of the stroke and age.

As previously discussed, stroke patients can be grouped as ischaemic or haemorrhagic. Patients with haemorrhagic strokes have a five times higher case fatality compared to those with an ischaemic stroke (Bamford et al., 1990). Once a CT scan has confirmed diagnosis, those with ischaemic stroke can then be further sub classified. In 1991 a classification for sub groups of ischaemic stroke was developed, this is often referred to as the Bamford Classification (Bamford et al., 1991).

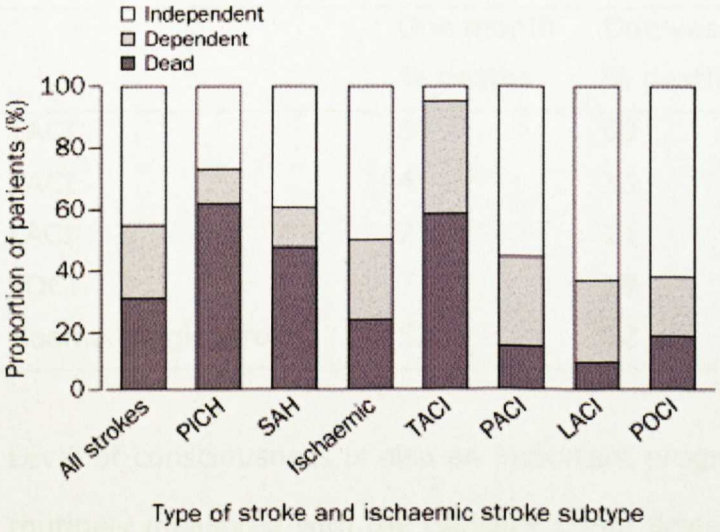
Four sub groups of ischaemic stroke were established:

- Total anterior circulation infarction (TACI) ~ 20% of patients
- Partial anterior circulation infarction (PACI) ~ 30% of patients
- Posterior circulation infarction (POCI) ~ 25% of patients
- Lacunar infarction (LACI) ~ 25% of patients

The prognosis of patients who fall into these categories is very different, and therefore this classification can be used as a prognostic factor.

FIGURE 1.4

Proportion of patients who are dead, dependent, or independent a year after first stroke by type of stroke and by clinical subtype of ischaemic stroke.



Reprinted from The Lancet, 362. 9391, Warlow C et al. Stroke, 1211-1224., Copyright (2003), with permission from Elsevier.

Patients with a TACI have suffered a large infarct with both cortical and sub cortical involvement, with slow recovery (Sprigg et al., 2007). These patients have the worse prognosis with high mortality (see Figure 1.4). Patients with a PACI are more likely to have recurrent strokes, while patients with POCI are at the greatest risk of a recurrent stroke later in the first year after initial onset. Patients with POCI have the best chance of a good functional outcome post

stroke. Patients with LACI have suffered from small infarcts, but can still remain substantially disabled. Table 1.1 shows the percentage of deaths in each sub group at one month and one year (Bamford et al., 1991, Ebrahim and Harwood, 2003).

There is a very small minority of patients (around 1%) with ischaemic stroke who do not fall into either category.

TABLE 1.1

Clinical stroke subtype and mortality (Bamford et al., 1991, Ebrahim and Harwood, 2003).

	One month % deaths	One year % deaths
TACI	39	60
PACI	4	16
LACI	2	11
POCI	7	19
Haemorrhagic stroke	52	62

Level of consciousness is also an important prognostic factor. Consciousness is routinely measured with the Glasgow Coma Scale, patients are scored between three (deep unconsciousness) and 15 (normal state) (Teasdale and Jennett, 1974). The Glasgow Coma Scale is highly related to both mortality at two weeks and outcome at three months (Weir et al., 2003). Severity is related to level of consciousness, with patients with more severe stroke tending to have a lower level of consciousness. The National Institute of Health Stroke Scale (NIHSS) (Brott et al., 1989) is a well validated measure of stroke severity and has been shown to be strongly related to outcome at both seven days and three months. A higher score on the NIHSS reflects greater severity and it has been shown that for every unit increase on the NIHSS, the likelihood of a good

outcome at seven days is decreased by 24%, and by 17% at three months (Adams et al., 1999).

The increasing risk of stroke with increasing age is well documented, but age is also an important prognostic factor. A study looking at producing models for predicting prognosis found that the chance of surviving a stroke decreases by 3% with every year increase in age at stroke onset. It was also found that the odds of becoming independent after a stroke also decrease with increasing age (Odds Ratio (OR) 0.95, 95% Confidence Interval (CI) 0.93-0.97) (Counsell et al., 2002).

Other factors which can be used to predict early mortality are high blood pressure (Sprigg et al., 2006), raised blood glucose, raised haematocrit, atrial fibrillation, pupil changes, gaze paresis, abnormal breathing, abnormal body temperature and meningeal irritation (Ebrahim and Harwood, 2003).

1.2.5 Treatment

Currently there are four interventions which have been shown in randomised controlled trials to improve outcome in acute stroke: admission to a stroke unit; treatment with aspirin; treatment with thrombolytic therapy and most recently, decompressive surgery for those with cerebral oedema. Admission to a stroke unit can be used to treat patients with both ischaemic and haemorrhagic stroke, whereas aspirin, thrombolytic therapy and decompressive surgery may only be used in ischaemic cases. Unfortunately, there have been no definitive clinical trials which have demonstrated beneficial medication for patients with haemorrhagic stroke. If the bleed is life threatening, then surgical evacuation of the clot can be considered (Warlow et al., 1996).

Stroke units combine acute stroke care with rehabilitation. In 1997 a systematic review was carried out on studies which looked at stroke unit care. The review found that stroke units gave a reduction in death (OR 0.83, 95% CI 0.69 to 0.98), poor outcome (death or dependency) (OR 0.69, 95% CI 0.59 to 0.82) and death or institutionalisation (OR 0.75, 95% CI 0.65 to 0.87) (Stroke unit trialists' collaboration, 1997).

Treatment with aspirin has been shown to have a limited effect, but has wide utility. A data pooling project found that acute treatment with aspirin showed a reduction in the combined outcome of death or non-fatal recurrent stroke of one per 1000 patients treated (Chen et al., 2000).

In contrast, treatment with thrombolytics has been shown to be highly effective but with limited availability. Thrombolytic treatment aims to break down the clot and restore blood flow to the damaged part of the brain, and in doing this reduce the area of brain damage and therefore improve outcome (Warlow et al., 1996). The National Institute of Neurological Disorders and Stroke (NINDS) trial showed that treatment within three hours of onset improved outcome at three months, with an 11-13% absolute increase in the chance of minimum or no disability (The National Institute Of Neurological Disorders And Stroke rt-Pa Stroke Study Group, 1995).

Decompressive surgery involves removing a skull flap to alleviate intra-cranial pressure and remove the risk of death from pressure building up in the brain. A meta analysis of three trials (one of these is still ongoing) showed that patients in the surgery group had a better outcome and improved survival (Vahedi et al., 2007).

1.2.6 Prevention

Most strokes are thought to be preventable; there are four reasons for this. Firstly, variations in time and place, both within and between countries suggest that stroke risk is changeable. Secondly, observational studies have shown that migrants adopt the risk of their host environment. Thirdly, personal characteristics are associated with the gradient of stroke risk (i.e. the lower the level of the risk factor the lower the occurrence of stroke). Lastly, experimental evidence from randomised controlled trials demonstrates that stroke incidence is reduced following the reduction of stroke risk factors (Ebrahim and Harwood, 2003, Marmot and Poulter, 1992). A risk factor is defined as something that predisposes a person to a morbid event (Millikan et al., 1987). Risk factors for stroke can be split into two groups, those that can be modified, and those that are non-modifiable. Modifiable risk factors include: high blood pressure, cigarette smoking (Shinton and Beevers, 1989), heart disease, diabetes, hormone replacement therapy use (Bath and Gray, 2005), and alcohol consumption (Wolf, 1998). Non modifiable risk factors include: age, sex, family history, and ethnicity (Wolf, 1998).

High blood pressure is a major modifiable risk factor. Blood pressure is calculated using two measurements, one when the heart beats (systolic) and one when the heart relaxes (diastolic). Both systolic and diastolic blood pressure have been shown to be positively and independently associated with the primary incidence of stroke. Reducing systolic blood pressure by 5.8 mmHg has been shown to lead to a 42% reduction in the incidence of stroke (Collins et al., 1990). Similarly, for diastolic blood pressure between the range of 70-110 mmHg, the risk of stroke doubles with each increase of 7.5 mmHg (MacMahon et al., 1990). Blood pressure may be reduced by losing weight, eating a healthy diet of low saturated fat, cholesterol and salt, being more physically active and

lowering alcohol intake. Although modification of these factors can have an effect on blood pressure, this effect is generally modest. For example, a 10 kg drop in body weight may reduce systolic blood pressure by 6-16 mmHg, and 30 minutes of daily exercise leads to a reduction of around 3.3 mmHg (Bhatt et al., 2007), therefore many patients will require blood pressure lowering therapy.

1.2.7 Secondary prevention

The term secondary prevention refers to preventing further strokes in patients who have already suffered a stroke. Those who have suffered from a stroke or a TIA are at a higher risk of having a recurrent stroke than those who have not. A population based study found that after TIA or minor stroke the risk of recurrence was around 8-12% at seven days, 12-15% at one month and 17-19% at three months, with the higher rates being seen in those with minor stroke compared to TIA (Coull et al., 2004). The "Early use of Existing Preventive Strategies for Stroke" (EXPRESS) study showed that early treatment after TIA or minor stroke could reduce the risk of early recurrence by 80% (Rothwell et al., 2007).

There are different treatment options available for the prevention of secondary strokes depending on the cause of the initial event. If patients have suffered from an ischaemic stroke there are medications available that may block the formation of further blood clots and therefore reduce the risk of further strokes. The most widely used treatment of this type is aspirin, which can reduce the risk of stroke by around 13-22% (Antithrombotic Trialists' Collaboration, 2002). There are other alternative therapies that work in a similar manner, including clopidogrel and dipyridamole, and recently it has been shown that being treated

with both dipyridamole and aspirin gives a greater risk reduction than aspirin alone (Halkes et al., 2008).

Anticoagulants, such as warfarin, are recommended for those having suffered an ischaemic stroke caused by a blood clot from the heart. The majority of these patients will have atrial fibrillation (AF), which is an abnormal heart rhythm. These patients are at a much higher risk of recurrent stroke than those without AF. The "Birmingham Atrial Fibrillation Treatment of the Aged Study" (BAFTA) showed that treatment with warfarin compared to aspirin significantly reduced the risk of recurrent events (1.8% per year compared to 3.8% per year, p=0.003) in patients aged over 75 (Mant et al., 2007).

Carotid surgery (endarterectomy) can be used for those whose stroke was caused by a blocked blood vessel on the side of the neck, in order to clear the blockage. The "North American Symptomatic Carotid Endarterectomy Trial" (NASCET) showed that carotid surgery reduced two year absolute risk of stroke by 17% (North American Symptomatic Carotid Endarterectomy Trial Collaborators, 1991).

1.3 OUTCOME

Outcome is defined as "a change in a patient's current and future health status that can be attributed to antecedent care" (Donabedian, 1980). Outcome after stroke is important for clinical research as it can be used to measure an individual's progress or to compare groups of patients. For example, in a clinical trial, outcome (i.e. number of recurrent strokes or level of disability) can be used to compare a new treatment to the standard treatment after a predefined length of time. There are many outcomes which can be used, from objective measures such as mortality to more complex subjective measures such as

quality of life. Functional outcome is regularly used as the primary outcome in clinical trials on stroke.

1.3.1 Functional outcome

After suffering a stroke approximately a third of patients will die, a third will return to full independence (although residual disability may be present) and a third will have some sort of lasting disability and therefore dependency on others.

In 1980 the WHO published the International Classification of Impairments, Disabilities and Handicaps (ICIDH) (World Health Organization, 1980). This was produced to give a framework against which information could be organised to clarify the consequences of disease (Kearney and Pryor, 2004). Impairment was defined as any loss or abnormality of psychological, physiological or anatomical structure or function, so for example, in stroke leg or arm weakness. Disability was classified as any restriction or lack of ability to perform an activity in a manner which is normal for a human being, i.e. the functional results of impairment. Whereas, handicap is a disadvantage for a given individual, normally resulting from a disability or impairment that limits or prevents the person fulfilling their normal role. This model states that both impairment and disability are pre-requisites of handicap, and therefore suitably implying that impairment and disability cause handicap. In this model impairment is the least important measure to the patient, with handicap being the most important (Roberts and Counsell, 1998).

The ICIDH has been updated and revised and in 2001 the WHO published the International Classification of Functioning, Disability and Health (ICF). Importantly this revision included the opinions of disabled people, which were

not represented in the ICIDH. This model is much more complex than the original and aims to give a unified language and framework for the description of health and health-related states. It is made up of two parts; the first part considers physiological impairments, limits to activities and involvement in life situations. The second part considers contextual factors such as the environment and personal characteristics (World Health Organisation, 2001).

The term "disability" is no longer included as a component within the ICF, but rather as an umbrella term for any impairment of body structure or function, limitation of activities or restriction in participation (Bowling, 1997). The ICF as a whole describes a person's level of functioning, with functioning now being a continuum rather than only focusing on the extreme points.

Although the more recent ICF is now the accepted way of defining disability, the scales used throughout this project were based on the previous definitions and therefore use the terms impairment, disability and handicap.

1.3.2 Outcome scales

An outcome scale normally takes the form of a number of predefined levels on an ordinal scale, normally ranging from the worst possible state to the best possible state. Scales can either have a set of questions which, when answered, give the patient a score, that relates to their place on the scale. Or conversely, each level on the scale has a clear definition and the person assessing the patient decides which level describes the patient best.

In stroke research the type of outcome used depends on whether the researcher wants to measure impairment, disability or handicap. Impairment is normally assessed using a scale for neurological deficit and handicap is gauged

by using a scale which assesses change in the patient's social role. Disability is routinely determined using a scale which assesses Activities of Daily Living (ADL). ADL scales generally include items on excretion (bowels, bladder, toileting), mobility (transfers, wheelchair/walking, stairs), hygiene (grooming, bathing), feeding and dressing. ADL scales can also be extended (EADL, also called instrumental ADL) to take into account housework, shopping and leisure activities (Barer and Nouri, 1989).

1.3.3 Choosing a scale

When choosing an outcome scale there are issues which need to be investigated, namely: reliability, validity, sensitivity and simplicity (Wade, 1992). Reliability simply assesses that the scale is measuring something that is reproducible. For example, do different assessors give the same patient the same score (inter-rater reliability) or do different methods of administration produce comparable results (inter-method reliability). Reliability also measures the extent to which the items within the scale are measuring the same characteristic (internal consistency) (Streiner and Norman, 1995, Hantson and De Keyser, 1994). Test-retest reliability is determined by administering the test on the same population on two occasions and comparing the results, usually with correlation (Bowling, 1997).

Validity assesses what the scale is actually measuring, and whether or not the scale is measuring what it claims to be. There are three aspects to validity: construct, criterion and content. Construct validity establishes whether the results obtained from the scale concur with the results predicted from the underlying theoretical model. Testing the scale against the gold standard measures criterion validity. Content validity is measured by the extent to which

the scale contains all relevant dimensions of what is being measured (Hantson and De Keyser, 1994, Wade, 1992).

The scale chosen needs to be able to detect clinically important changes in the patient's condition; this is referred to as the scale's sensitivity. The simplicity of the scale is also important; using a simple measure will improve compliance and reliability. Unfortunately, for a scale to be sensitive a complex measure is normally required and therefore this decreases the reliability and simplicity, leading to a trade off between the three (Wade, 1992).

Alongside these statistical factors, an outcome also needs to be able to detect clinically relevant differences in the effectiveness of various therapies for a given disease, with the smallest number of patients possible (Broderick et al., 2000).

1.3.4 Frequently used outcome scales

There are many outcome scales available for measuring disability, impairment, and handicap. In stroke research three scales are predominantly used in large multi centre randomised controlled clinical trials; these are the Barthel Index, modified Rankin Scale, and the Three Questions outcome, each is discussed in detail below.

Barthel Index (BI)

The Barthel Index (BI) was first published in 1965 as a simple and effective way of measuring a patient's level of independence (Mahoney and Barthel, 1965). It consists of ten weighted items which measure feeding, bathing, grooming, dressing, bowel control, bladder control, toileting, chair transfer, and stair climbing. A score of zero is given when the patient cannot meet any of the criteria, and 100 is the maximum score. Many have used the BI with a score out of 20 rather than 100, as it is thought that larger score gives a false impression of the scale's accuracy (Collin et al., 1988). Although not defined in the scale, patients who die are usually given the arbitrary score of minus five to distinguish them from those with the lowest level of dependence.

An example of the BI is given in Table 1.5 at the end of this chapter.

The reliability and validity of the BI are well established (Collin et al., 1988, Granger et al., 1979, Wade and Langton Hewer, 1987). In 1996 a study which re-evaluated the reliability and validity of stroke scales found that the BI was the most reliable disability scale (D'Olhaberriague et al., 1996). The BI has been shown not only to have high reliability and validity when used as an ordinal scale, but also when dichotomised at 90 to compare those who are independent (≥90) against those who are dependent. The BI can be administered reliably in a variety of ways, including face to face interview, telephone interview and by using a postal questionnaire (Yeo et al., 1995). This makes the BI especially useful in studies with a long follow up period or where a large population of highly dependent patients is being assessed. The BI can be used to predict outcome and has been shown to forecast survival, length of hospital stay and progress in stroke patients (Wilkin et al., 1993). The main disadvantages of the BI are the presence of profound floor and ceiling effects.

Floor and ceiling effects occur when many participants are scored at the highest or lowest point of the scale, although this is true for most measures of ADL. The BI is also insensitive to small changes in functional ability. Some modifications to the BI have been proposed to overcome some of these problems (Granger et al., 1979), but none have sufficiently improved the original in terms of reliability and validity to replace it (Wade, 1992). The BI is the most commonly used ADL scale (Wade, 1992, Roberts and Counsell, 1998).

Modified Rankin Scale (mRS)

The Rankin Scale was developed as a five level scale in 1957 from research on the prognosis of stroke (Rankin, 1957). The scale is a simple and relatively crude measure of handicap and is the stroke equivalent of the Glasgow Outcome Scale (GOS) for brain injury (Jennett and Bond, 1975). In 1991 the Rankin Scale was modified for use in the UK-TIA study to accommodate language disorders and cognitive defects (now referred to as the modified Rankin Scale, mRS) (Farrell et al., 1991).

An example of the mRS is given in Table 1.6.

The mRS is used regularly throughout stroke research. This is probably due to the ease of administration and time efficiency of the scale. When analysing the mRS, the scale has historically been dichotomised, comparing patients with a good outcome to those with a poor outcome. A review by Sulter of stroke research found that most studies defined a good outcome as either having a mRS of ≤1 or a mRS of ≤2 (Sulter et al., 1999). This raises concern, since de Hann found that a valid dichotomy was at mRS ≤3 (de Haan et al., 1995).

The scale is predominantly used to measure handicap, although many agree that the scale actually measures disability rather than handicap (Bloch, 1988).

A study by de Hann in 1995 found that results from the mRS were strongly associated with mobility, disability in daily and instrumental activities, and living arrangements. It found a low association with cognitive and social functioning. ADL were found to be the most important explanatory factor of mRS scores. This study concluded that the mRS should therefore be used as a measure of functional health and physical disability rather than a measure of handicap (de Haan et al., 1995).

The reliability of the mRS is well documented. A study looking at the inter rater agreement found that out of 100 pairs of raters, 65 agreed with the level of handicap (van Swieten et al., 1988). Giving raters a structured interview to follow has been shown to improve reliability further (Wilson et al., 2002). Little is known about the validity of the mRS (Bowling, 1995). The mRS has low sensitivity; this is probably due to the simplicity of the measure. Improvements have been suggested for the mRS, including reducing the number of grades and removing the assessment of walking, but these have not been implemented as this reduction would lead to an even more decreased level of sensitivity (van Swieten et al., 1988).

Three Questions outcome (3Q)

The International Stroke Trial (IST) was a large randomised controlled trial comparing treatment with aspirin, heparin or both in 19,435 patients with acute ischaemic stroke (International Stroke Trial Collaborative Group, 1997). The trialists wanted a simple method of assessing dependency, as the large sample size meant that standard methods such as the BI would be too costly in terms of both time and money.

In 1994 a pilot study was carried out to identify a few simple questions which could establish functional status in a valid and reliable way (Lindley et al., 1994). The questions chosen also needed to be reliable when administered in a variety of ways, including face to face interview, over the telephone or as a postal questionnaire. The questions selected were:

1. Is the patient alive? (Vital status question)
2. In the last two weeks did you require help from another person for everyday activities? (Dependency question)
3. Do you feel you have made a complete recovery from your stroke? (Recovery question)

By comparing the 3Q outcome with the BI and the Oxford Handicap Scale (OHS) (Bamford et al., 1989) (a variant of the mRS), the study found that asking these three simple questions was a valid way of distinguishing between patients who had good and bad functional outcomes after stroke. They found that even though the scale was crude, as the intention of the study was to look at overall functional outcome for a large group of people it was sufficient to do this. The study ascertained that there was no significant difference in the accuracy of the scale when administered by either a postal questionnaire or a telephone interview. When looking at the raters, it was found that patients were better at rating themselves than carers when they had a good functional outcome and, interestingly, that carers were better at rating the patients when the patient had a bad outcome.

When comparing the BI and the OHS, it was found that the second question could accurately identify a poor outcome, defined as BI<100, 75% of the time. Similarly, the third question could identify an OHS score of zero (equivalent to

mRS of zero) 90% of the time (Lindley et al., 1994). A study using data from the Italian centres in the IST trial found comparable results (Celani et al., 2002).

1.3.5 Issues with data from functional outcome scales

Data gained from outcome scales have particular properties which need to be appreciated when choosing the type of analysis to carry out. There are four main types of data that can be measured: nominal, ordinal, interval and ratio. Nominal data is considered the lowest level of data, where the data are categorical and no ordering can be applied. Examples of nominal data are gender, blood group, and marital status. This type of data is usually analysed using contingency tables and comparing frequencies using a chi square test (Jakobsson, 2004, Wade, 1992).

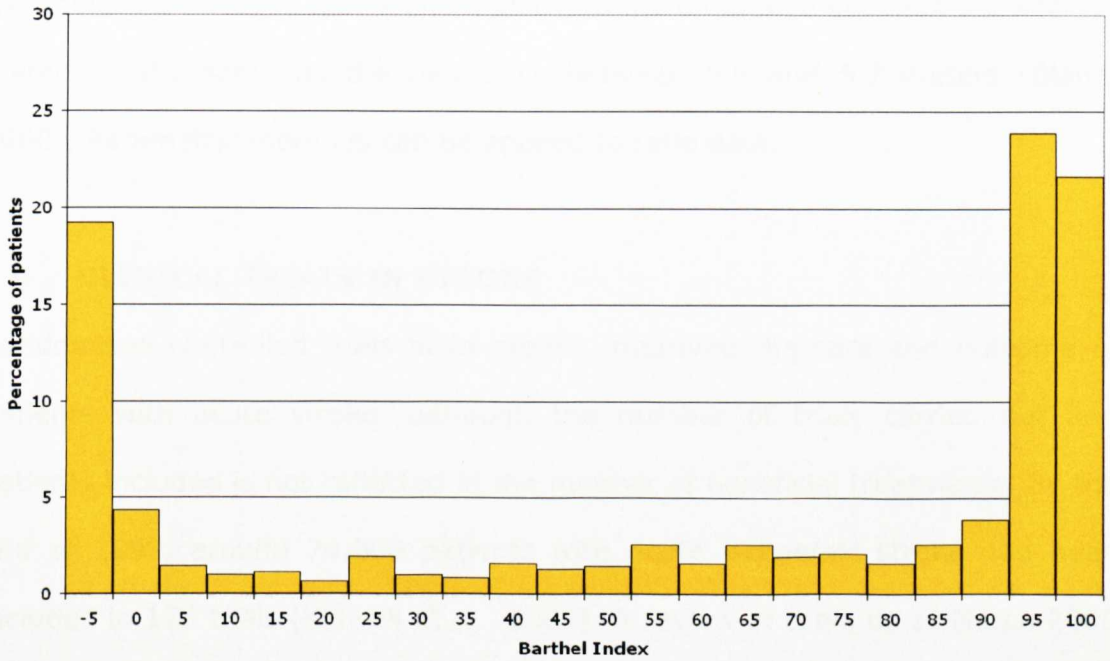
Outcome scales are usually ordinal in nature. The central feature of ordinal data is that it expresses increasing or decreasing order to the extent of some observable phenomenon. For example, education is ordinal when measured as "primary", "secondary", "college", "undergraduate" and "postgraduate" (Moses et al., 1984). A secondary feature is that although there is clear ordering to the categories the absolute distance between them is unknown (Agresti, 1984). Using the BI as an example, a patient who scores 20 on the BI is more disabled than someone scoring 40, but the patient scoring 40 does not necessarily have half the disability of the patient who scored 20. Data from scales such as the BI and mRS which produce numbered ordered categories are often mistaken for continuous data, but the values are just indicating the order and not actual numeric values. Historically across many disciplines, not only stroke, ordinal data is analysed incorrectly. In 1984, Moses carried out a review of articles from the New England Journal of Medicine over a six month period; this found

that 18/168 studies collected ordinal data. Of these he found that 30% dichotomised the data and 33% analysed the data in a contingency table that ignored the ordering (Moses et al., 1984). A study looking at ordinal data analysis in a rheumatology journal found similar results with only 39% of the articles surveyed having appropriate data presentation and 63% having appropriate analysis (Lavalley and Felson, 2002). A further study looking at nursing research found that out of 166 articles, 51 had used ordinal methods, with only 49% of these displaying this data appropriately and 57% using appropriate data analysis (Jakobsson, 2004).

Another feature of data from outcome scales is its distribution. Data from the BI, for example, has profound floor and ceiling effects. This is because around a third of the patients will have died (scoring -5), around another third of the patients will have recovered completely (scoring 100). The remaining patients spread across the rest of the scale (See Figure 1.5).

FIGURE 1.5

Distribution of BI at three months, data from the NINDS trial (The National Institute of Neurological disorders stroke rt_PA stroke study group, 1995).



This unusual distribution means that standard parametric methods such as comparing means may not be valid and a non-parametric approach should be taken.

Interval scale data is similar to ordinal data but the differences between the scores are identical. Therefore the unit difference between ten and 11 on a scale is the same as a difference between 50 and 51. An interesting point about interval scales is that there is no natural zero, which means that ratios of the data do not make sense. For example, like ordinal scales, a score of ten on an interval scale is not twice as good as a score of five. A good example of an interval scale is the Fahrenheit scale for temperature. Equal differences on this scale represent equal differences in temperature, but a temperature of 30 degrees is not twice as warm as one of 15 degrees. Interval scale data can be analysed using parametric methods (Wade, 1992).

The final type of scale data is ratio. Ratio data are continuous data where both the differences between units and ratios are interpretable. Unlike interval data, ratio data have a natural zero. Height and weight are examples of ratio data, two meters is twice as tall as one metre and the difference between 1.2 and 1.3 meters is the same as the difference between 5.6 and 5.7 meters (Bland, 2000). Parametric methods can be applied to ratio data.

1.4 CLINICAL TRIALS IN STROKE

Randomised controlled trials have greatly improved the care and outcome of patients with acute stroke, although the number of trials carried out and patients included is not reflected in the number of beneficial treatments. By the end of 1999, around 74,000 patients with acute ischaemic stroke had been included in 178 trials (Kidwell et al., 2001). A review of trials up to March 2006 gave much higher estimates with 9,409 completed stroke trials, 2,240 of these being in the acute setting (Bath et al., 2007). The majority of these trials have shown no treatment effect, with aspirin and thrombolysis with alteplase, the only agents now being used in acute stroke. There are many possible reasons for the failure of these trials, including the relevance of laboratory findings to clinical stroke, inadequate sample size, the choice of primary outcome and its statistical analysis.

1.4.1 Relevance of laboratory findings

Many potential stroke treatments have shown efficacy in animal models but there has been a difficulty in translating these results into humans. For example, NXY-059, a neuroprotection agent, significantly reduces infarct volume in mice, rats and marmosets, but when tested in a large clinical trial showed neutral results (Lees et al., 2006, Bath et al., 2008). Many reasons have been put forward for this failure including the quality of animal studies, the design of animal studies (randomisation to treatment, blinding of outcome measures, sample size calculation) and the applicability of animal studies to humans (Sena et al., 2007). The "Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Stroke" (CAMARADES) is a multidisciplinary collaboration addressing these problems (Macleod and Sandercock, 2005).

1.4.2 Inadequate sample size

In 2004 a review of sample size calculation in acute stroke trials was carried out (Weaver et al., 2004). This review included 189 fully reported randomised controlled trials and found that only 57 gave detail on their sample size calculation (30%). Most of these 57 were published after 1996 with the introduction of the "Consolidated Standards of Reporting Trials" (CONSORT) statement, which required trials to include their sample size calculation in the trial manuscript in order to be published in prestigious peer reviewed journals (The CONSORT Statement, 1996). Of these 57 the majority were underpowered, using unrealistic event rates and intervention effects or using inappropriate outcomes, such as death (Weaver et al., 2004). For example, 24 trials had a primary outcome of death or dependency, and had a median intended reduction of 12% (inter quartile range 10%-15%). Whereas on completion, the actual median reduction found was 1.9% (inter quartile range -

0.5%-5.4%), which shows a major overestimation of the desired clinically important difference used in these trials.

It is important therefore to consider how sample size is to be calculated when recommending any particular method of analysis to trialists.

1.4.3 Choice of primary outcome and its statistical analysis

In 1998, a study reviewed the outcomes used in stroke research and the appropriateness of these outcomes and the statistical analysis applied to them. All published acute stroke trials reported in English from 1955 to 1995 were included in the review. They found that the most common measures of disability were the BI (21%), trial specific outcomes (11%) and the mRS (9%). This is a concern as more trials were using an unvalidated measure, as opposed to the mRS which has been shown to be a reliable way of measuring disability. Several of the trials assessed had measured disability using more than one scale.

The review found that most trials had used a less than optimal method of analysis. They found that many trials had analysed the outcome scale data as if it were continuous, using parametric methods. Twelve trials using the BI analysed it as a dichotomous variable, but with no standardisation in the cut-off point used to define a good outcome. Five trials used a cut-off of ≥60, four used ≥70, one used ≥90 and another trial used a cut-off of ≥95 (Roberts and Counsell, 1998).

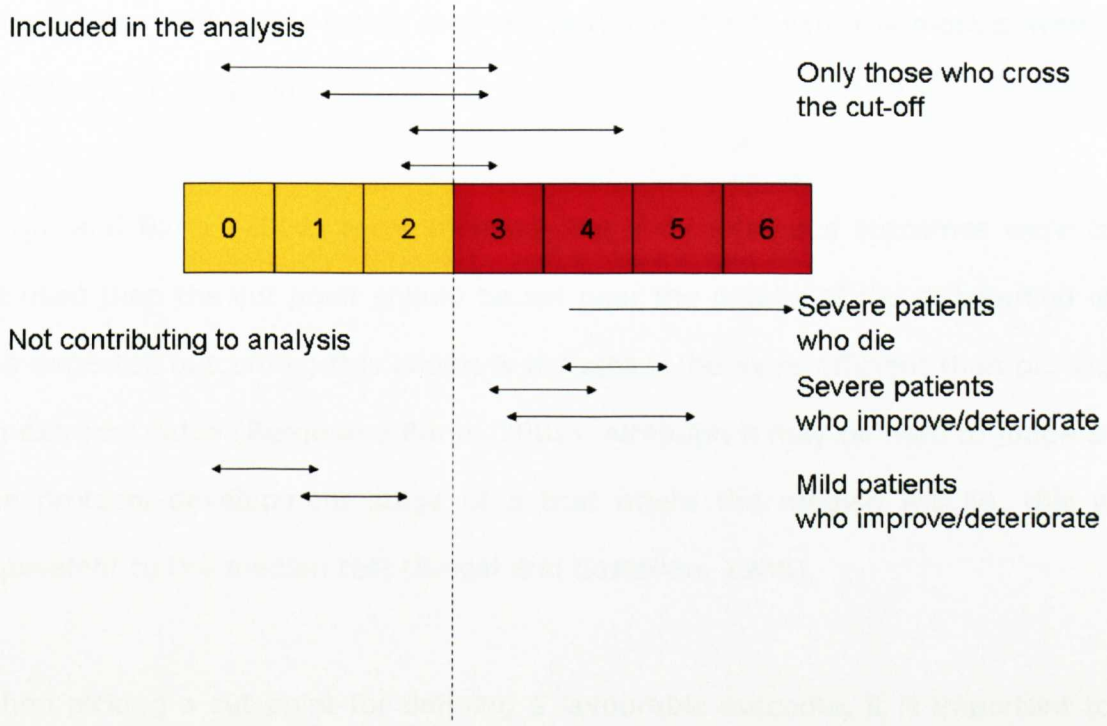
This review highlights a plethora of problems in the choice of outcome and analysis; these problems are assessed in the subsequent sections.

Dichotomisation

Dichotomisation involves collapsing data into two groups; dichotomous data is a type of nominal data. Dichotomous outcomes are perceived as clinically meaningful, as clinical definitions can be placed on the groups and therefore easily interpreted. For example, thrombolysis with alteplase reduced death or dependency (defined by a score of greater than one on the mRS) by 13% in the NINDS part two trial (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995). Whereas an analysis based on the actual ungrouped data would be presented as average improvements, e.g. alteplase improved the mRS by one of seven points and BI by 22.5 (of 100) points, which may be harder to explain to patients. However, using a dichotomous outcome inherently means that clinical meaning is only attributed to transitions in outcome that occur over the pre-specified cut-off point for a favourable outcome. This is demonstrated in Figure 1.6, which shows an artificial example of a trial which has chosen a cut-off for a good outcome of ≤2 on the mRS. Much data is lost, for example, very severe patients who improve a point on the mRS do not add anything as both their pre and post scores are higher than the threshold for a "good outcome".

FIGURE 1.6

The pitfalls of using a dichotomised outcome.



A review of the use of the BI and mRS in stroke trials found that a favourable outcome was defined variably on the BI as ≥50, ≥60, ≥75, ≥85 and ≥95, and on the mRS as ≤1, and ≤2. Other trials had compared median scores and three trials had used a combined BI/mRS scale. The review highlighted that most of these end points were arbitrarily chosen and there was no evidence of validation. The review concluded that it might be beneficial to use poor outcome as an end point and to define this if any of the following occur; death, institutionalisation, mRS>3, or BI<60 (Sulter et al., 1999).

In contrast, another study found that changing the outcome from mRS≥2 to mRS≥3 did not change the result of a meta analysis looking at the efficacy of thrombolytic therapy. The study concluded that if a treatment is beneficial it probably doesn't matter where the data is dichotomised (Wardlaw et al., 2000).

A post hoc study of the NINDS stroke trial data used classification and regression tree analysis to find the most powerful binary outcome. The results showed that end points which used the mRS cut at ≤1 were the most powerful (Broderick et al., 2000).

Berge and Barer (2002) recommended that if dichotomous outcomes were to be used then the cut point should be set near the middle of the distribution of the expected outcomes; this choice is thought to be more efficient than picking an extreme value (Berge and Barer, 2002). Although it may be hard to judge at the protocol development stage of a trial where the median will lie, this is equivalent to the median test (Siegel and Castellan, 1988).

When picking a cut point for defining a favourable outcome, it is important to take into account the population of patients to be recruited into the trial. A recent trend in stroke trials has been to copy the outcomes used in a previous trial which showed a statistically significant treatment effect, but this may lead to trials picking an unsuitable cut. For example, the "Surgery for the Treatment of Malignant Infarction of the Middle Cerebral Artery" (DESTINY) trial (Juttler et al., 2007) of decompressive surgery, recruited patients who were suffering from life-threatening brain swelling as a consequence of a massive ischaemic stroke. These patients have very severe strokes and therefore a cut between three and four on the mRS was chosen for the primary outcome. See Figure 1.7. In contrast, Figure 1.8 shows the distribution of the mRS from the NINDS trial which included much milder patients and therefore used a cut between one and two on the mRS. An advantage of dichotomy is that it negates the need to assign arbitrary values to dead patients, where no value for death automatically exists in the scale (Tilley et al., 1996).

FIGURE 1.7

Distribution of outcomes in the DESTINY Trial (Juttler et al., 2007).

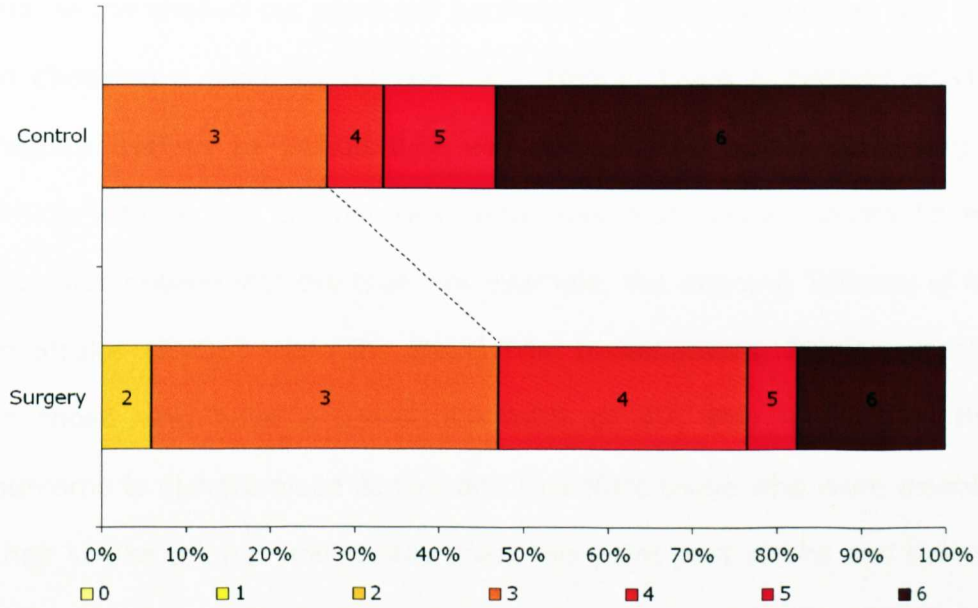
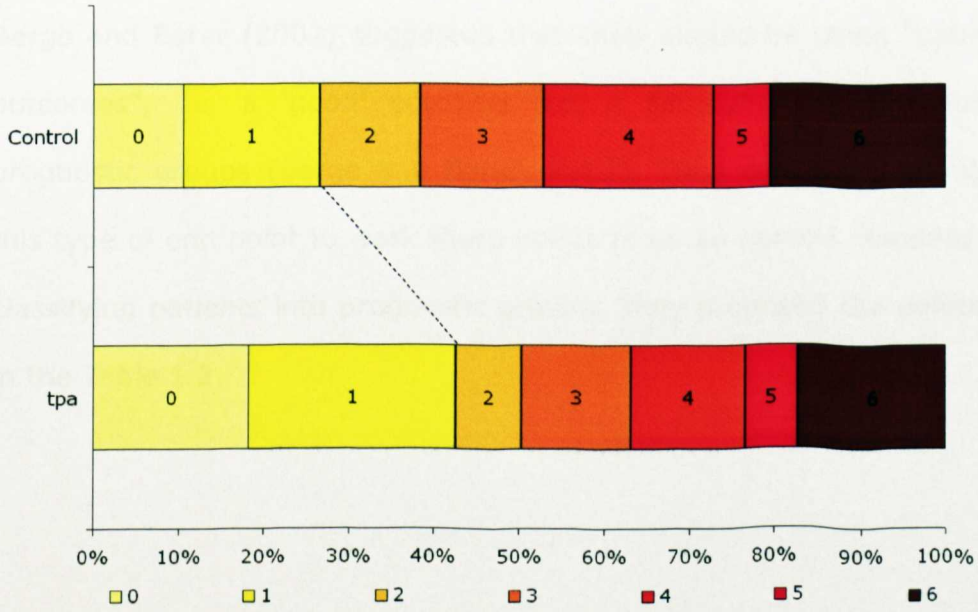


FIGURE 1.8

Distribution of outcomes in the NINDS Trial (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995).



Overall, there appears to be little consensus as to where trialists should dichotomise data. The main disadvantages of using outcomes which have been dichotomised are the loss of information, as only those patients who move across the chosen cut point will be included in the comparison and the difficulty in choosing a place to cut the data. Hence, using a method which does not require trialists to dichotomise will avoid these pitfalls. Choosing a method which retains the original raw data may also allow trialists to widen their inclusion criteria into the trial. For example, the ongoing 'Efficacy of Nitric Oxide in Stroke' (ENOS) trial (The ENOS Trial Investigators, 2006) restricts inclusion to those who have a pre-stroke mRS of ≤2; this is because the primary outcome is dichotomised at two and therefore those who were disabled prior to their stroke will not realistically cross this point post stroke and therefore would not add any information to the primary end point (The ENOS Trial Investigators, 2006).

Patient specific outcomes

Berge and Barer (2002) suggested that trials should be using "patient specific outcomes", i.e. a 'good' outcome has a separate definition for separate prognostic groups (Berge and Barer, 2002). They contended though that for this type of end point to work there needs to be an agreed standard method of classifying patients into prognostic groups. They proposed the definitions given in the Table 1.2.

TABLE 1.2

Proposed outcome by Berge and Barer (2002).

Prognostic group	Outcome group		
	Good (mRS score)	Intermediate (mRS score)	Bad (mRS score)
Severe	0-3	4	5/dead
Moderate/bad	0-2	3	4-5/dead
Moderate/good	0-1	2-3	4-5/dead
Good	0-1	2	3-5/dead

The patient specific outcome was assessed alongside those which dichotomised, in a study which aimed to find the most powerful end point for use in acute stroke trials (Young et al., 2003). This study used simulation to explore the patterns and magnitudes of treatment effects and the statistical power for a range of end points based on the BI and mRS. The study found that generally mRS end points were more powerful than those using the BI. It was also found that the most powerful end points were patient specific, those which were dichotomised towards the favourable extreme and those which combined the BI and mRS.

A more recent paper has also focused on the patient specific outcome, and aimed to find definitions of a good outcome on the mRS for various levels of baseline severity, measured by the NIHSS scale (Adams et al., 2004). The definitions used in this study are given in Table 1.3.

An example of the NIHSS is given in Table 1.7.

TABLE 1.3

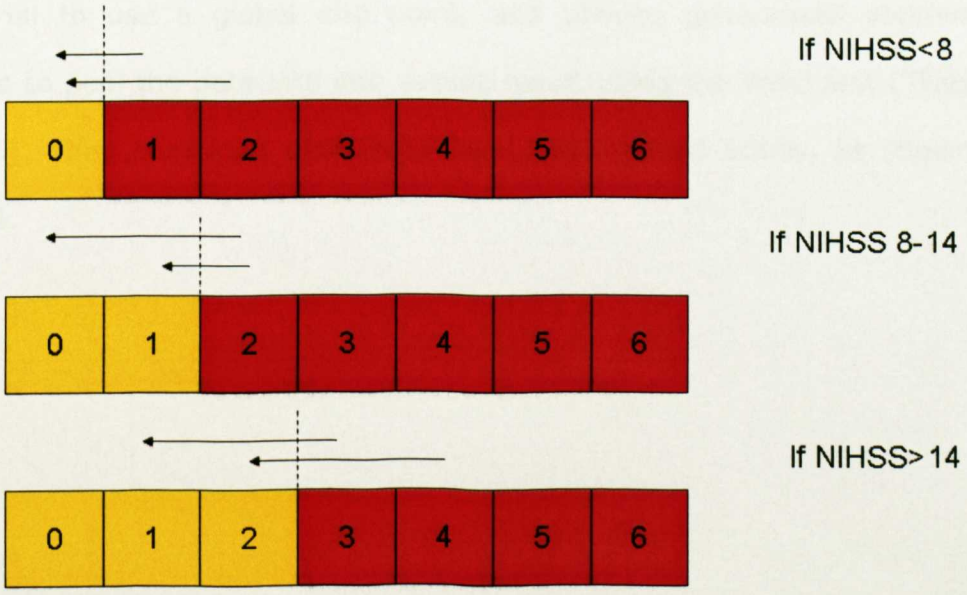
Definitions of a good outcome on the mRS for levels of baseline severity on the NIHSS scale (Adams et al., 2004).

Baseline NIHSS score	Outcome group mRS
<8	0
8-14	0-1
>14	0-2

This paper takes the earlier work of Berge and Barer (2002) one step further by giving actual levels of severity instead of just the subjective headings of mild, moderate and severe. The study carried out its proposed analysis on three completed and reported clinical trials. They found that although the patient specific analysis did not change the overall result of any of the trials it gave the opportunity to look at the effect of the treatment across the levels of baseline severity (see Figure 1.9) (Adams et al., 2004).

FIGURE 1.9

Diagram of patient-specific outcome.



The phase two trial "Emergency administration of abciximab for treatment of patients with acute ischemic stroke" (AbESTT) was one of the first stroke trials to include a patient specific outcome as a secondary end point. This end point along with the primary end point showed a beneficial effect of abciximab compared to placebo (Abciximab Emergent Stroke Treatment Trial (AbESTT) Investigators, 2005). A phase three trial was then initiated using the patient specific outcome as the primary end point (AbESTT II). Unfortunately this trial was terminated prematurely due to an excess of bleeding events in the abciximab group. The patient specific outcome also did not show efficacy of abciximab (Adams et al., 2008).

This type of outcome is perhaps more appealing than simple dichotomisation, but it is still based on a group of dichotomised end points.

Global outcomes

Some have argued that restricting an end point to one scale may be limiting, as no scale describes all dimensions of recovery from stroke. Global outcomes can be used to combine data from two or more scales. The NINDS trial was the first stroke trial to use a global end point, and utilised generalised estimating equations to pool the data into one overall result using the Wald test (Tilley et al., 1996). They combined data from four dichotomised scales, as shown in Table 1.4.

TABLE 1.4

NINDS global outcome definitions of a favourable outcome (Tilley et al., 1996).

Scale	Dichotomy used
mRS	≤ 1
BI	≥ 95
NIHSS	≤ 1
Glasgow Outcome Scale (GOS)	1

The NINDS trial showed a beneficial treatment effect for thrombolysis with alteplase using both the global outcome and additionally testing each scale separately. It does require data to be collected on four scales at the follow up point which could increase the length and costs of follow ups and, as with the patient specific outcome, is still based on dichotomised data and therefore has all the disadvantages of these. The European Medicines Evaluation Authority is also reluctant to consider global end points as they may combine very diverse data (Committee for Proprietary Medicinal Products (CPMP), 2001). Although a study comparing a global outcome to those based on a single scale found that combining the mRS and BI gave a more statistically powerful outcome than analysing either scale on its own. They also found the global outcome to be more powerful than a patient specific outcome (Young et al., 2003).

Type of analysis

Little work has been done looking at outcomes which maintain the raw data from the outcome scales. By dichotomising ordinal scales information is lost and it might be expected that types of statistical analysis that preserve and utilise the data in this ordinal form may be more powerful.

A study carried out in 2006 reviewed 100 trials where the BI had been used as the outcome (Song et al., 2006). They recommended that trialists reported

mean BI scores to facilitate meta analyses and that the Wilcoxon test appeared to have the greatest power to detect differences between treatment groups compared to dichotomised end points and using ordinal logistic regression analysis. However, this is a confusing message as the paper advocates a non-parametric method of analysis but also suggests giving parametric summary statistics.

Summary

This literature review has shown the general lack of agreement on a standard effective end point in stroke research. Most of the research in this area has focused on reviewing the methods which have been used previously in published clinical trials. Only the Young and NINDS studies (Young et al., 2003, Broderick et al., 2000) tested to see which end points were the most powerful and therefore should be recommended for use. Although the NINDS study only considered binary end points and the Young study did not consider methods for non parametric ordinal data, such as the Wilcoxon test. All of the studies discussed above are based in the acute setting and although outcome scales are frequently used in rehabilitation studies as well, no studies have focused on this area.

1.4.4 Published alternative statistical analysis of clinical trials

There have been several clinical trials where an alternative statistical analysis has been performed and the results have been published. These give weight to the argument that sub optimal end points and statistical analyses are being used in clinical trials in stroke.

The first "European Cooperative Acute Stroke Study" (ECASS I) tested the efficacy and safety of alteplase given within six hours of ischaemic stroke onset (Hacke et al., 1995). The primary end points were the median BI and mRS at 90 days post randomisation. The study was powered to detect a 15% improvement of the median of each primary end point. The results showed no statistically significant benefit for alteplase. The NINDS trial also tested alteplase but with different time windows, 90 and 180 minutes. The NINDS trial used a global end point analysis as their primary outcome (Tilley et al., 1996) and showed a beneficial result and the Food and Drug Administration therefore licensed thrombolysis with alteplase for use in acute ischaemic stroke.

The ECASS I investigators undertook a post hoc analysis of the trial data to see whether using a different statistical design would have given them a statistically significant result. Global end point analysis was carried out on the ECASS I data using three outcomes, ≤1 on the mRS, ≥95 on the BI and ≤1 on the NIHSS. The global outcome analysis showed a statistically significant increase of favourable outcome in the alteplase group ($p=0.008$, OR=1.5, 95% CI 1.1 to 2.0). It was concluded that this post hoc analysis may indicate that the time window for alteplase may be as long as six hours, and that the initial choice of end point was sub optimal. However, as this was a retrospective analysis this result could only be used to support data from the NINDS trial and not to show efficacy in using alteplase in a six hour time window (Hacke et al., 1998).

The second "European Cooperative Acute Stroke Study" (ECASS II) was similar in design to ECASS I but used a lower dose of alteplase. Akin to ECASS I, no significant benefit of alteplase was found. Again the researchers felt that had a different outcome been used a statistically significant result may have been found. The primary outcome used was the mRS dichotomised between one and two. (See Figure 1.10). It was decided to re-analyse the data using bootstrapping. Bootstrapping is a computer intensive method that involves choosing random samples with replacement from a data set and analysing each sample the same way, and then using these samples to make inferences (bootstrapping will be described in further detail in Chapter 3) (Efron and Tibshirani, 1993). Bootstrapping was chosen as its does not require the researcher to make any assumptions about the distribution of the data. For example, changing the cut on the mRS could be perceived as data driven, i.e. picking a cut point which gives the lowest p value. The post hoc bootstrap analysis showed a statistically significant beneficial treatment effect. The ECASS II investigators concluded that further clinical trials would need to be carried out to confirm this result (Stingele et al., 2001).

FIGURE 1.10

Re-analysis of the ECASS II data.

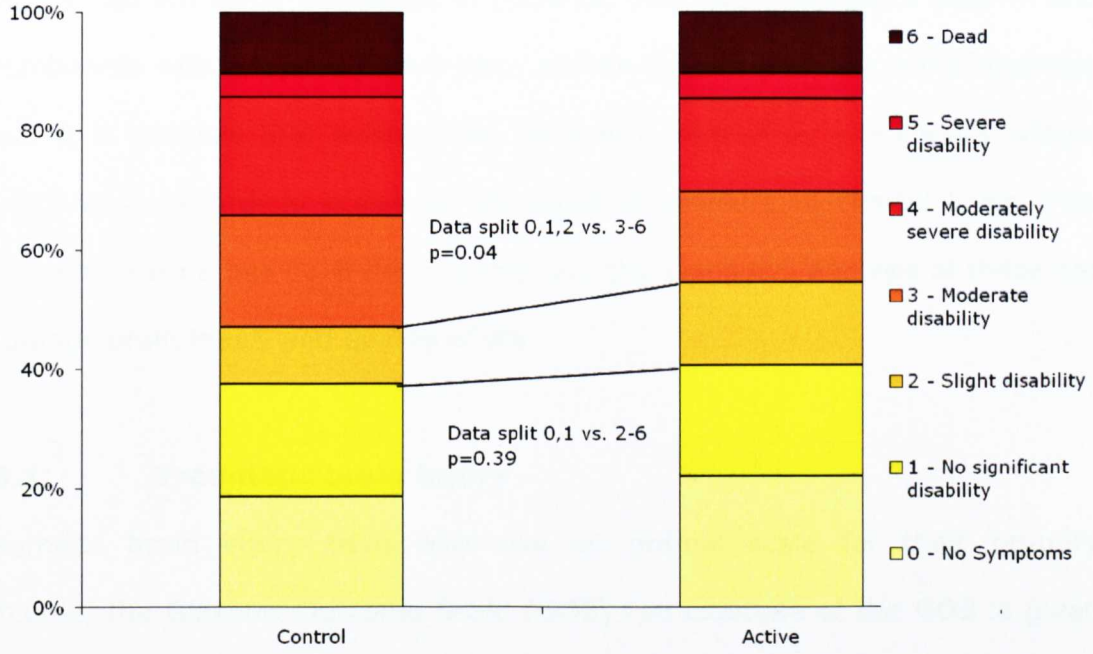


Figure 1.10 shows the arbitrary nature of dichotomous end points and the effect the choice of cut can have on the result found. The cut point chosen in the original trial shows no difference between the two groups (p=0.4). If the trialists had chosen the next cut up when setting up the trial, a statistically significant result would have been found in the favour of alteplase (p=0.04).

More recently, both the NINDS and ECASS II trials have been re-analysed using the Cochran Mantel-Haenszel test. This test compares two groups adjusting for one or more variables (Savitz et al., 2007). Parallel to the previous paper, when re-analysed a statistically significant result was found for the ECASS II trial.

These examples have demonstrated how important the choice of primary outcome and the subsequent analysis is. If dichotomising an outcome scale, the choice of cut seems to be particularly important.

1.5 OTHER AREAS USING ORDINAL SCALES

Stroke trials have come to a crisis point. Although a plethora of research has been carried out using thousands of patients, only two treatments (aspirin and thrombolysis with alteplase) have been proven to work and are being routinely used. It is possible that lessons can be learnt from other therapeutic areas. Many areas use ordinal scales as an outcome measure in clinical trials. Two areas where work has been done to improve the statistical analysis of these are traumatic brain injury and quality of life.

1.5.1 Traumatic brain injury

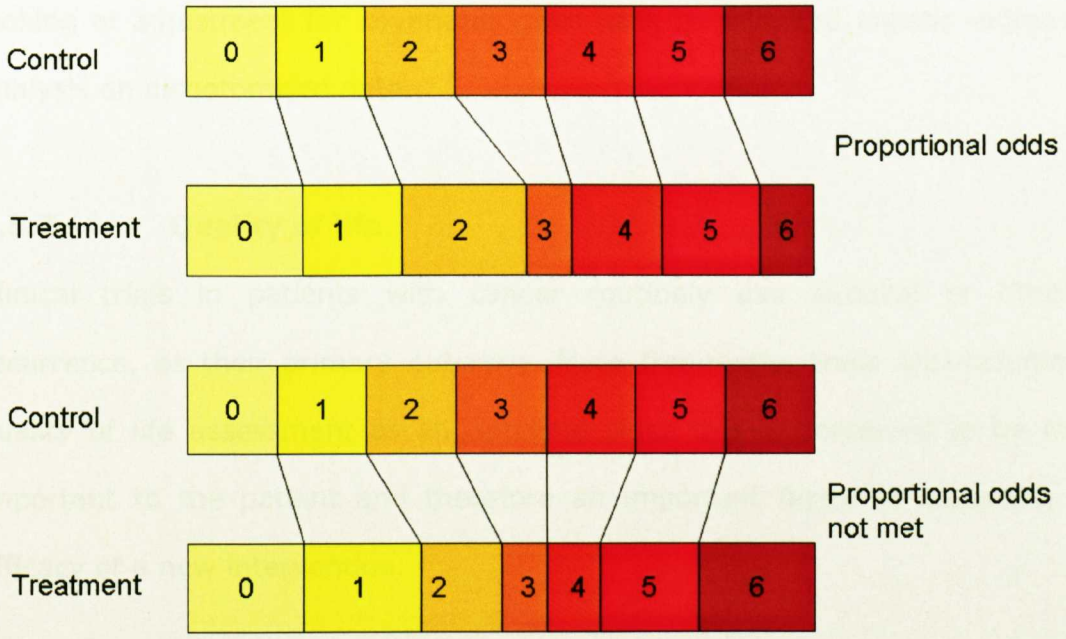
Traumatic brain injury trials also use an ordinal scale for their primary outcome, the Glasgow Outcome Scale (GOS) (an example of the GOS is given in Table 1.8), is a five level scale ranging from one (dead) to five (fully recovered) (Jennett and Bond, 1975). The "International Mission for Prognosis and Clinical Trial" (IMPACT) study is looking at ways to improve the design and analysis of traumatic brain injury studies; they are interested in both clinical trials and epidemiological studies (Marmarou et al., 2007). Part of this project has looked at improving the analysis of data from the GOS. Historically the majority of traumatic brain injury trials have, akin to stroke trials, dichotomised outcome scales into favourable and unfavourable groups. The IMPACT study has condemned this type of analysis as those with severe injury do not contribute to the final outcome as their improvement will be limited and almost certainly will not cross the pre-specified cut-off for a favourable outcome (Murray et al., 2005).

IMPACT proposed the use of a patient specific outcome based on the work of Berge and Barer (previously discussed in section 1.4.3) (Berge and Barer, 2002), although they termed this type of outcome as a "sliding dichotomy".

They assessed this by comparing it with ordinal logistic regression analysis using data from two completed clinical trials. Ordinal logistic regression compares data across the whole scale and is similar to logistic regression but does not require dichotomisation. Ordinal logistic regression makes the assumption that odds ratio for the treatment covariate is the same for each transition of the scale (termed 'proportional odds assumption') (see Figure 1.11).

FIGURE 1.11

Diagram of proportional odds assumption.



They extended the previous work by using a prognostic model based on age, baseline motor score, and baseline CT to divide patients into tertiles of risk, with each risk group being given a different definition of a favourable outcome. They found that the sliding dichotomy analysis was more sensitive than the ordinal logistic regression. This type of analysis has since been used in two trials of traumatic brain injury (Maas et al., 2006, Mendelow et al., 2005).

The next part of the IMPACT project dealing with the analysis of ordinal outcomes looked at taking into account covariates. They found that by adjusting logistic regression analysis for seven important prognostic factors sample size could be reduced by 25% (Hernandez et al., 2006).

This area of research is helpful for stroke trials, although the IMPACT study is looking at a wide range of questions and therefore the focus is not just on improving the statistical analysis of trials. The initial part of the project advocates using a method of analysis that allows for shifts in outcome across the whole scale, either by using ordinal logistic regression or a dichotomous analysis with differing cuts for differing levels of risk, although the second part looking at adjustment for covariates goes back to a limited logistic regression analysis on dichotomised data.

1.5.2 Quality of life

Clinical trials in patients with cancer routinely use survival or time to recurrence, as their primary outcome. More frequently, trials are including a quality of life assessment as an outcome since this is perceived to be more important to the patient and therefore an important factor in assessing the efficacy of a new intervention.

Quality of life scales are similar to functional outcome scales as they are both ordinal in nature. Therefore studies which have looked at the analysis of data from quality of life scales maybe useful in improving the analysis of data from functional outcome scales. The studies discussed here all use the Short-Form 36 health survey to measure quality of life, this measure contains 36 questions on eight domains of quality of life resulting in a score of zero to 100, where 100 is indicative of "good health" (Ware et al., 1993).

The majority of work carried out in this area has looked at comparing the given sample sizes needed if different methods of analysis are applied to trial data. This type of analysis is analogous to looking at the power of that test; more powerful tests will require smaller samples to find the same result as a lower powered test. A study by Julious et al showed that methods of analysis which rely on the assumptions of normality may not be suitable for quality of life data and can lead to either over or under estimated sample sizes. Therefore methods which do not make assumptions about distribution should be employed (Julious et al., 2000). In contrast, others have suggested that where scales have seven or more categories, methods such as the t-test which assume normality may be reliably used, with the analysis of scales with fewer categories using ordinal logistic regression (Walters et al., 2001, Walters, 2004).

Parallel to the ECASS II trial, bootstrapping (Efron and Tibshirani, 1993) has also been assessed as an option for the analysis of quality of life data. Here it was found that bootstrapping was no more powerful than other standard methods and therefore given its complexity to carry out should not be promoted for analysing quality of life data (Walters and Campbell, 2005).

The work of Walters and Campbell is interesting as quality of life data suffer the same problems as functional outcome scale data (floor and ceiling effects, nonlinearity), but this work is only based on Short-Form 36 health survey and therefore may not be generalisable to other scales.

1.6 AIM

The main aim of this project is to identify the most statistically efficient techniques for analysing functional outcome data from stroke clinical trials. This project intends to improve and extend on the research which has already been carried out in five ways:

1. **Using real trial data**; functional outcome data follows an unusual distribution which can be difficult to model with artificial data.
2. **Using data from three stroke settings**; all previous work has been based on acute stroke trials, this project will include data from acute stroke, rehabilitation, and stroke unit trials.
3. **Assessing all methods of analysis**; this project will not only assess traditional nominal methods of analysis but also methods for ordinal data, bootstrapping and modelling. The project will also look at other outcomes which have been used in stroke trials, such as patient specific outcomes and global outcomes which combine data from two or more scales.
4. **Adjustment for covariates**; so far research has only considered univariate methods. In some cases it may be beneficial to adjust for imbalances in baseline characteristics or take into account prognostic variables. After the assessment of univariate methods this project will also assess models available for ordinal outcome scales.
5. **To consider the analysis of stroke prevention trials.**

TABLE 1.5

Barthel Index (BI) (Mahoney and Barthel, 1965).

Scored out of 100 with those who have died coded as -5

Domain	Item	Points
Bowels	Incontinent (or needs to be given enemata)	0
	Occasional accident (once a week)	5
	Continent	10
Bladder	Incontinent, or catheterised and unable to manage alone	0
	Occasional accident (maximum once per 24 hours)	5
	Continent	10
Toilet use	Dependent	0
	Needs some help, but can do something alone	5
	Independent	10
Grooming	Needs help with personal care	0
	Independent face/hair/teeth/shaving (implements provided)	5
Feeding	Unable	0
	Needs help cutting, spreading butter, etc.	5
	Independent	10

Transfer (from bed to chair and back)	Unable, no sitting balance	0
	Major help (one or two people, physical), can sit	5
	Minor help (verbal or physical)	10
	independent	15
Mobility	Immobile	0
	Wheelchair independent, including corners	5
	Walks with help of one person (verbal or physical)	10
	Independent (but may use any aid; for example, stick)	15
Dressing	Dependent	0
	Needs help but can do about half unaided	5
	Independent (including buttons, zips, laces, etc.)	10
Stairs	Unable	0
	Needs help (verbal, physical, carrying aid)	5
	Independent	10
Bathing	Dependent	0
	Independent (or in shower)	5

TABLE 1.6

Modified Rankin Scale (mRS) (Rankin, 1957).

Level	Description
0	No symptoms
1	No significant disability, despite symptoms; able to perform all usual duties and activities
2	Slight disability; unable to perform all previous activities but able to look after own affairs without assistance
3	Moderate disability; requires 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 requires constant nursing care and attention
6	Dead

TABLE 1.7

National Institute of Health Stroke Scale (NIHSS) (Brott et al., 1989).

Domain	Item	Points
Level of Consciousness	Alert, keenly responsive	0
	Obeys, answers or responds to minor stimulation	1
	Responds only to repeated stimulation or painful stimulation (excludes reflex response	2
	Responds only with reflex motor or totally unresponsive	3
LOC questions (month and age)	Answers both correctly	0
	Answers one correctly or patient unable to speak due to any reason other than aphasia or coma	1
	Answers neither correctly, or too stuporous or aphasic	2
LOC commands (open and close eyes and then grip and release non-paretic hand)	Performs both tasks correctly	0
	Performs 1 task correctly	1
	Performs neither task correctly	2
Best gaze	Normal	0
	Partial gaze palse	1
	Forced deviation or total gaze paresis not overcome by oculocephalic maneuver	2

Visual fields	No visual loss	0
	Partial hemianopia	1
	Complete hemianopia	2
	Bilateral hemianopia (blind from any cause including cortical blindness)	3
Facial palsy	Normal symmetrical movement	0
	Minor paralysis (flattened nasolabial fold, asymmetry on smiling)	1
	Partial paralysis (total or near total lower face paralysis)	2
	Complete paralysis (absence of facial movement upper/lower face)	3
Arm Motor (score both right and left arm)	No drift - holds for full 10 seconds	0
	Drifts down before ten seconds but does not hit bed/support	1
	Some effort against gravity, but cannot get up to 90 (or 45 if supine) degrees	2
	No effort against gravity, limb falls	3
	No movement	4

Leg motor (score right and left leg)	No drift - holds for full five seconds	0
	Drifts down before five seconds but does not hit bed/support	1
	Some effort against gravity	2
	No effort against gravity, limb falls	3
	No movement	4
Limb ataxia	Absent	0
	Present in one limb	1
	Present in two limbs	2
Best language	No aphasia	0
	Some loss of fluency or comprehension	1
	Severe aphasia - fragmentary communication, listener carries burden of communication	2
	Mute, global aphasia. NO usable speech OR auditory comprehension	3
Dysarthria	Normal	0
	Slurs some words	1
	So slurred as to be unintelligible, or mute	2

Extinction/Inattention	No abnormality	0
	Inattention to any sensory modality or extinction to bilateral simultaneous stimulation in one sensory modality	1
	Profound hemi-inattention or hemi-inattention to more than one modality. Does not recognize own hand	2

TABLE 1.8

Glasgow Outcome Scale (GOS) (Jennett and Bond, 1975).

Level	Description
1	Dead
2	Vegetative state: Unable to interact with environment; unresponsive
3	Severe disability: Able to follow commands; unable to live independently
4	Moderate disability: Able to live independently; unable to return to work or school
5	Good recovery: Able to return to work or school

CHAPTER 2

GENERAL METHODS

PUBLICATIONS/PRESENTATIONS CONTRIBUTING TO THIS CHAPTER

Gray L.J, Bath P.M.W, OAST Collaborators (2005) Statistical analysis of ordered outcome data. A review of methods used in the trials included in 'Optimising the Analysis of Stroke Trials' project. (OAST) *Poster presentation at the Research Students Conference, Cambridge, April 2005.*

2.1 INTRODUCTION

This chapter details the general methods which apply to subsequent chapters. In particular this chapter will discuss in detail the setting up of the trial database which was used throughout this project. Chapter specific methods will be discussed in the relevant chapter.

2.2 OPTIMISING ANALYSIS OF STROKE TRIALS DATA

This section will discuss the setting up of the 'Optimising Analysis of Stroke Trials' (OAST) database.

2.2.1 Identification of trials

Individual patient data from randomised controlled trials assessing functional outcome after stroke were sought from four groups of studies:

1. Trials showing significant benefit on their primary outcome
2. Trials showing significant harm on their primary outcome
3. Trials showing no significant effect but within a meta analysis showing significant benefit
4. Trials showing no significant effect but within a meta analysis showing significant harm

Trials relating to ineffective interventions as determined from published meta analyses were excluded. Trials were identified using a number of search strategies. Firstly, meta analyses of beneficial or harmful treatments were identified from the Cochrane Library. Secondly, electronic searches of Medline, Embase and PubMed were carried out using the terms "Stroke" and "Trial", if any trials of interventions not identified during the search of the Cochrane Library were found, another search of the Cochrane Library was carried out looking for reviews of that intervention. Thirdly, hand searches of the journals Stroke and Cerebrovascular Diseases were carried out and all trials listed on

the online directory of stroke clinical trials were assessed (see: http://www.strokecenter.org /trials/). Finally, new trials of beneficial or harmful interventions were sought from the announcement of the trial's results at international conferences.

The Chief Investigator (CI) of each trial identified was contacted and asked if they would share their data with the 'Optimising Analysis of Stroke Trials' (OAST) Collaboration. All contacted investigators were informed that all data shared with the collaboration would be kept confidential and would not be used for any other purposes. CIs were asked to share the following data (where available):

- Randomised treatment (compulsory)
- Functional outcome data (compulsory)
- Age
- Sex
- Baseline severity

Individual patient data from several studies had already been gathered in four data pooling projects ('Blood pressure in Acute Stroke Collaboration' (BASC) (Blood pressure in Acute Stroke Collaboration (BASC), 2001), community occupational therapy (Walker et al., 2004), low molecular weight heparin, tirilazad (The Tirilazad International Steering Committee, 2002)) and was used where relevant following agreement with the CI.

Table 2.1 summarises the trials which were selected for inclusion and where permission was granted and data were supplied. For some trials permission to use data was not given, but it was possible to extract the data from the original publication (MAST-E (Multicentre Acute Stroke Trial-Italy (MAST-I) Group, 1995), EAST (Enlimomab acute stroke trial investigators, 2001), Edaravone

(The Edaravone Acute Brain Infarction Study Group (Chair: Eiichi Otomo MD), 2003), Helsinki (Kaste et al., 1995), PROACT II (Furlan et al., 1999), Orpington 2000 (Kalra et al., 2000), Streptokinase pilot (Morris et al., 1995), and FISS TRIS (Wong et al., 2005)).

2.2.2 Setting up the database

SAS statistical software version 8 (SAS Institute, Cary NC) was used for all data manipulation. Initially each data set was reformatted into a SAS data set and saved into a library. All data was reformatted so that variable names, labels and formats were consistent across all data sets. The variables needed for this project were then copied from the original data set and saved. These new data sets were then merged together to create the final database to be used in this project.

Trials where two or more effective or detrimental active treatments had been compared to a control were treated as two or more separate trials; therefore the control patients were duplicated in the database. From here on, these will be treated as separate trials. For example, the INWEST (Wahlgren et al., 1994) trial is included twice: low dose nimodipine versus control; and high dose nimodipine versus control. This is because both high and low dose nimodipine showed a detrimental treatment effect when compared to control. Whereas, the IST trial compared, in a factorial manner, aspirin and heparin with no treatment (International Stroke Trial Collaborative Group, 1997). Here only aspirin versus no aspirin is included as there was no effect seen for heparin versus no heparin. From here on, the actual trials included in the OAST project will be named 'trials' whereas the total number of treatment comparisons will be referred to as 'data sets'. Table 2.2 lists the trials where multiple data sets have been included.

2.2.3 Data checking

All data were checked against the original trial publication to ensure consistency. Where discrepancies were found the authors were contacted and changes were made based on their recommendations. In some cases, especially in older trials, it was not possible to reflect the findings exactly as they were reported in the original trial publication.

2.2.4 Data manipulation

To ensure consistency across the OAST database many decisions about the format of the data had to be made. This section details the changes and judgments made while compiling the OAST database.

2.2.5 Primary functional outcome scale

Many of the trials included collected data on more than one functional outcome scale. For example, the NINDS trial (The National Institute Of Neurological Disorders And Stroke rt-Pa Stroke Study Group, 1995) used a global outcome which merged data from four scales (mRS, BI, NIHSS, GOS) for its main outcome. In cases such as this, a decision had to be made about which scale to use in the functional outcome analysis. Since the mRS is used in modern trials, the mRS has been taken as the main outcome in the OAST project. In older clinical trials, the BI was routinely used so it was hoped that this would give large equal numbers of trials to be analysed using the BI and mRS. No trials were included with multiple scales which did not include the mRS.

For consistency within and across functional outcome scales decisions had to be made about the coding of each scale. It was decided that patients who were dead at the time of follow up should be coded as a unit lower/higher, depending on the direction of the scale, than the worse level on each scale. For example

the worse level of the BI is zero and the scale increases in units of five, so those patients who had died were recoded as minus five. It was decided to recode those who had died and not simply assign them to the worse level, as some of the scales being assessed have been designed to have a level for death, for example mRS and the 3Q scale, and it was felt that it would be more consistent to therefore have a separate level for all scales. A unit below the most severe category was chosen as this was straightforward to put into practice for any scale and the actual value chosen does not affect most of the statistical methods which are being compared. For example, those methods which are based on ranks such as the Wilcoxon test, and those which collapse the data, such as the chi square test are not affected by the actual value assigned to dead patients. There is also evidence that assigning a lower score on the BI to those who have died increases statistical power (Song et al., 2006). Table 2.3 describes the scales included and the levels ascribed to those patients who had died.

Only four of the trials included did not use the BI, mRS or 3Q scale to measure functional outcome. Two related trials used the Nottingham ADL scale (Barer et al., 1988), another trial used the Rivermead scale (Walker et al., 1996) and the remaining trial used a thirty point ADL scale (Sivenius et al., 1985) (See Table 2.3 for details of other scales).

Where no data on death at the time of follow up was given a number of strategies were set in place to recode those patients who had died. If the mRS had been assessed as well as the primary outcome scale then this was used to recode the primary outcome scale, as the value six is routinely used to denote death. In a similar manner to the mRS, the GOS or NIHSS scale were also used. If the dates of death and randomisation were given then the time of

death was calculated and where this fell before or at the time of follow up the patients were recoded as dead.

2.2.6 Length of follow up

Some trials also had two follow up assessment times reported, for example, ATLANTIS A (Clark et al., 2000), EAST (Enlimomab Acute Stroke Trial Investigators, 2001), Ebselen (Yamaguchi et al., 1998) and Corr (Corr and Bayer, 1995). If one of these times was reported as the primary outcome, data from this time point was used, i.e. in the EAST trial (Enlimomab Acute Stroke Trial Investigators, 2001) the mRS was measured at day five, 30 and 90, but day 90 was quoted as the primary outcome time. If equal emphasis was given to two time points then the time point closest to three months was used, i.e. in the Ebselen trial (Yamaguchi et al., 1998) no primary outcome time was listed, instead both day 30 and day 90 were given as major end points; therefore the data from day 90 was used here as the primary outcome.

2.3 STATISTICAL METHODS

All main analyses were carried out in SAS (version 8) or STATA (version 7 or version 8). All p values quoted are two sided, with statistical significance relating to a p value of less than or equal to 0.05.

The specific statistical methods relating to each results chapter are discussed within that chapter. Throughout this project there are two broad approaches used to compare the various methods of analyses; firstly re-analysing data from completed clinical trials and secondly using simulation to create data sets with known treatment effects. Re-analysing data from completed trials is the preferred method of comparison in this project. Stroke trial data has complicated covariate structures which are difficult to replicate using

simulation. Different interventions also have differing types of treatment effect, i.e. some treatments may work well in all patients whereas others may have the greatest effect in those with mild impairments, and again this type of intricacy is difficult to reproduce artificially. Simulation is used when the OAST data set does not have sufficient data to answer a particular question, but importantly where simulation is used it is based on actual trial data and therefore the covariate structures are retained and only the treatment effect is altered. The two approaches also look at two slightly different questions, re-analysing completed trial data is based on the treatment effect found in that trial, whereas simulation looks at the efficiency to detect specific known effect sizes. Simulation methods are also difficult to explain to clinicians and trialists.

Both approaches have been used to compare statistical methods in the previous literature, but to variable extents. Where actual completed trial data has been used, usually only one or two trials have been included and only comparing a limited number of statistical methods. For example, the two papers which re-analyse data from the two ECASS trials only re-analysed one data set and with only one additional method compared to the original endpoint (Hacke et al., 1998, Stingele et al., 2001). Two types of simulation have been used previously, simulating data from one completed clinical trial (Young et al., 2003) and simulating hypothetical data with known distributions (Song et al., 2006). None of the literature has simulated data from a vast variety of trial types, as used in this project.

When evaluating various methods of statistical analysis there are three comparators which can be used – the level of statistical significance (p value), the sample size needed for a given power, or the level of statistical power for a given sample size. Here I have used the level of statistical significance and the

reduction in sample size gained from using specific methods. These are the two comparators which are most frequently used in the previous literature. The level of significance is important to trialists as a statistically significant result showing benefit can be used to gain approval for new products, a comparison of p values was used in the re-analysis of the ECASS and NINDS trials (Hacke et al., 1998, Stingele et al., 2001, Savitz et al., 2007). The reduction in sample size is meaningful to trialists and can be directly translated into savings in terms of the cost and duration of clinical trials. Reductions in sample size have been used as the comparator in work comparing methods of analysis in trials of brain injury (Murray et al., 2005, Hernandez et al., 2006).

Throughout this project the term "statistical efficiency" refers to the level of statistical significance found, i.e. the most statistically efficient test will report the smallest p value in comparison to the other tests being compared.

2.4 DESCRIPTION OF OAST DATA SET

A total of 55 data sets (54,173 patients) were included, these comprised individual patient data from 38 trials and summary data extracted from the publications of a further nine studies; six trials had two active treatment groups, and one had three active groups so a further eight data sets were available. This section will further describe the data included in the OAST data set.

2.4.1 Baseline data

Table 2.4 shows the baseline characteristics of the trials included. The data related to 34 acute stroke trials, seven trials of rehabilitation (1,164 patients) and six trials of stroke units (1,399 patients).

Trial characteristics

There was great variation in the size of the trials included, ranging from 20 to 20,655 (mean 1153, median 302). The majority of the trials had recruited less than 1,000 patients (96%), with only the mega trials, IST and the Chinese Acute Stroke Trial (CAST), including approximately 20,000 each.

The included trials covered a wide range of interventions:

- Abciximab (AbESTT)

Inhibits clot formation by preventing fibrinogen binding between platelets. The AbESTT phase II trial included here (400 patients) showed a non significant improvement in functional outcome at three months. A subsequent phase three trial was stopped prematurely due to increased bleeding events in the treated group (Abciximab Emergent Stroke Treatment Trial (AbESTT) Investigators, 2005).

- Alteplase (ATLANTIS A & B, ECASS II, NINDS)

Is a powerful thrombolytic agent (clot busting), which is now licensed for use in acute ischaemic stroke within three hours of onset. Four trials included (2,179 patients) (Clark et al., 2000, Clark et al., 1999, Hacke et al., 1998, The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995).

- Aspirin (CAST, IST)

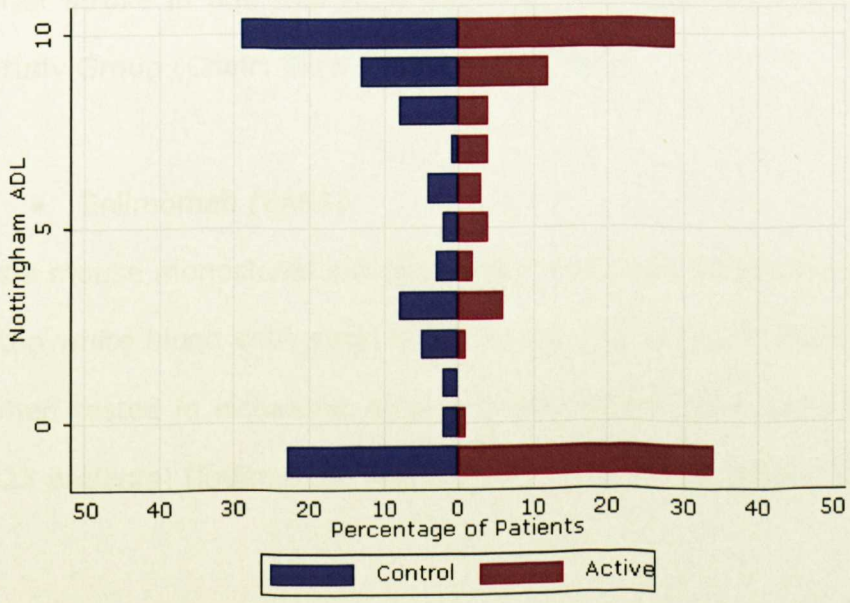
Is an antiplatelet, i.e. it stops platelets sticking together. It reduces death or dependency at six months and this reduction is probably caused by preventing early recurrent strokes. Two trials included (40,090 patients) (CAST (Chinese Acute Stroke Trial) Collaborative Group, 1997, International Stroke Trial Collaborative Group, 1997).

- Atenolol - propranolol (BEST)

Belong to a class of drugs called beta-blockers. A data pooling project showed that this class of drug increased death and dependency after stroke. Four trials included (367 patients) (Blood pressure in Acute Stroke Collaboration (BASC), 2001, Barer et al., 1988). Figure 2.1 shows the data from the BEST main trial. This shows the distribution of the Nottingham ADL scale and demonstrates, akin to the BI, a 'U' shaped distribution is found.

FIGURE 2.1

Distribution of functional outcome by treatment group for the BEST main trial of atenolol versus control (Barer et al., 1988).



- Citicoline (Citicoline 1, 7, 10 and 18)

Is thought to have neuroprotective benefits and appears to improve outcome after stroke. Four trials included (1,652 patients) (Clark et al., 2001, Clark et al., 1997, Clark et al., 1999, Warach et al., 2000).

- DCLHb

Diaspirin cross-linked haemoglobin (DCLHb) induces hypertension and had shown promise in animal models of stroke. A trial of DCLHb showed it significantly worsened outcome after stroke (85 patients) (Saxena et al., 1999).

- Ebselen

Is being investigated as a possible neuroprotectant in acute stroke. One small trial (298 patients) to date has shown that ebselen given with 24 hours of stroke onset improves outcome at one month (Yamaguchi et al., 1998).

- Edaravone

Is a novel free radical scavenger and has been shown to be neuroprotective after stroke in one trial (250 patients) (The Edaravone Acute Brain Infarction Study Group (Chair: Eiichi Otomo MD), 2003).

- Enlimomab (EAST)

Is a mouse monoclonal antibody which has been shown in laboratory studies to stop white blood cells sticking to the internal lining of blood vessels. However, when tested in ischaemic stroke, it was shown to worsen outcome (one trial, 623 patients) (Enlimomab Acute Stroke Trial Investigators, 2001).

- Factor VIIa

Treatment with factor VIIa within four hours after the onset of intracerebral haemorrhage has been shown to limit the growth of the hematoma, reduce mortality, and improves functional outcomes at three month post stroke (399 patients) (Mayer et al., 2005). However, the follow up phase three trial showed no effect of factor VIIa on functional outcome (Mayer et al., 2007).

- Feeding (FOOD 3)

The FOOD 3 trial compared feeding with a nasogastric tube versus percutaneous endoscopic gastrostomy (PEG) tube. Fatality and poor outcome was significantly higher for patients who were fed via PEG tube (321 patients) (The FOOD Trial Collaboration, 2005).

- Low molecular heparin – Nadroparin/ fraxiparine (FISS, FISS-TRIS)

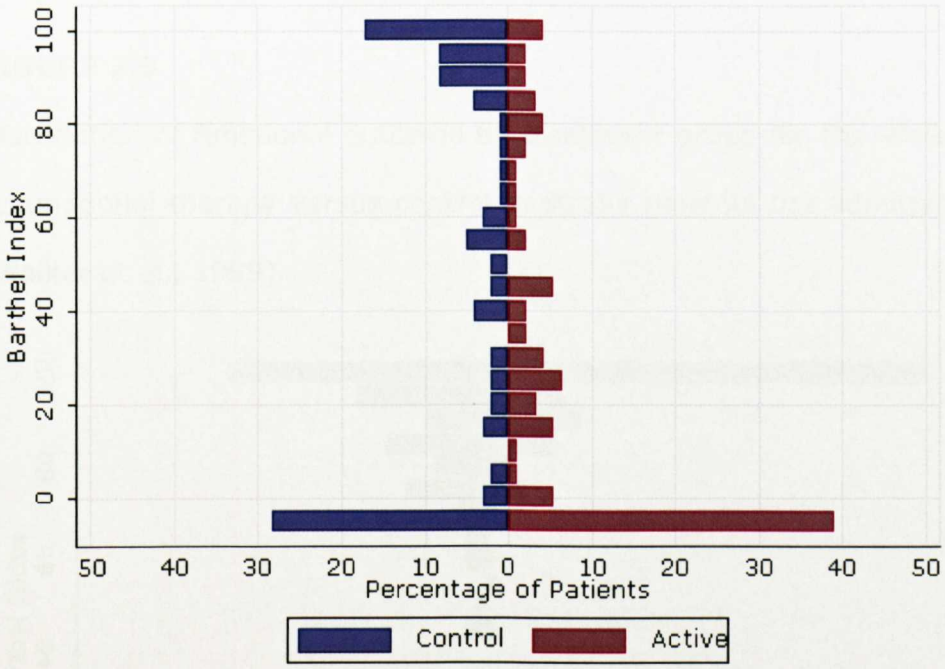
Works by thinning the blood and has been shown to improve outcome at six months in acute ischaemic stroke in these two trials. Two trials included (907 patients) (Kay et al., 1995, Wong et al., 2005).

- Nimodipine (INWEST)

Is a calcium channel blocker and was originally used to treat hypertension. A trial of nimodipine in acute ischemic stroke (295 patients) was stopped early due to a poor outcome in the treated patients (Wahlgren et al., 1994). Figure 2.2 gives the distribution of the BI in the high dose group compared to control, it shows an excess of deaths (-5) in the treated group and a greater percentage of good outcomes in the control group. This plot also again demonstrates the 'U' shaped data distribution associated with the BI in acute stroke trials.

FIGURE 2.2

Distribution of functional outcome by treatment group for the INWEST trial of high dose nimodipine versus control (Wahlgren et al., 1994).

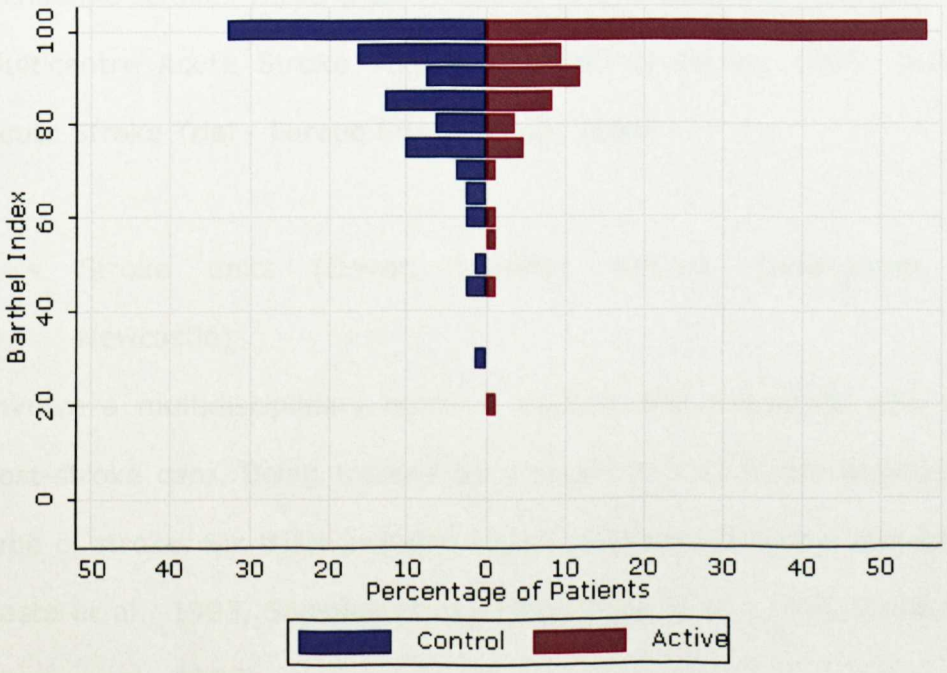


- Occupational therapy (Corr, Gilbertson, Logan, TOTAL, Walker I and Walker II)

Is the assessment and treatment of physical and psychiatric conditions using activities to prevent disability and promote independent function in all aspects of daily life. A data pooling project showed that occupational therapy improves outcome after stroke. Six trials included (1,040 patients) (Walker et al., 2004, Corr and Bayer, 1995, Gilbertson et al., 2000, Logan et al., 1997, Parker et al., 2001, Walker et al., 1999, Walker et al., 1996). Figure 2.3 shows the results from a trial of occupational therapy in stroke patients not admitted to hospital. This shows a non-typical distribution of the BI with patients bunched at the top end of the scale, this is because the patients enrolled had suffered from mild strokes and therefore had not scored badly on the BI at follow up.

FIGURE 2.3

Distribution of functional outcome by treatment group for the Walker II trial of occupational therapy versus control in stroke patients not admitted to hospital (Walker et al., 1999).



- Physiotherapy (Young)

Is concerned with maximising function and movement and is beneficial after stroke. One trial included (24 patients) (Young and Forster, 1992).

- Pro-urokinase (PROACT II)

Is another type of thrombolytic therapy and been shown to improve outcome after stroke in one trial (180 patients), although it increases early haemorrhage (Furlan et al., 1999).

- Selfotel (ASSIST)

Is an N-methyl-D-aspartate antagonist which blocks a receptor that can lead to neuronal damage. However, selfotel increased mortality in two trials (570 patients) in patients with acute ischaemic stroke (Davis et al., 2000).

- Streptokinase (ASK, MAST-E, MAST-I)

Is a thrombolytic therapy which has efficacy in breaking down clots in heart attack patients, but was shown to increase death and poor outcome in acute ischaemic stroke. Three trials included (1,272 patients) (Donnan et al., 1996, Multicentre Acute Stroke Trial-Italy (MAST-I) Group, 1995, The Multicenter Acute Stroke Trial - Europe Study Group, 1996).

- Stroke units (Dover, Helsinki, Kuopio, Nottingham, Orpington, Newcastle)

Involve a multidisciplinary team of doctors and therapists who specialise in post-stroke care. Being treated on a stroke unit is highly beneficial after any type of stroke. Six trials included (1,399 patients) (Stevens and Ambler, 1982, Kaste et al., 1995, Sivenius et al., 1985, Juby et al., 1996, Kalra et al., 2000, Aitken et al., 1993).

- Tirilazad (RANTTAS I & II, STIPAS, TESS I & II)

Is a nonglucocorticoid, 21-aminosteroid which had been shown to work well in animal models of stroke but was shown to worsen outcome in humans (five trials, 1,702 patients) (The RANTTAS Investigators, 1996, Haley, 1998, The STIPAS Investigators, 1994, Peters et al., 1996, Orgogozo, 1995).

Data from such a wide range of interventions means the results from this project will be generalisable to many different types of trial. The different types of intervention are also reflected in the timing of the treatments, this ranging from stroke onset to three hours post stroke for the alteplase trials and up to one month for the occupational therapy trials.

The majority of trials followed patients up at three months (66%), with other follow ups occurring at six months (23%), one year (9%) and one month (2%).

Patient characteristics

A total of 54,173 patients are included in this project. Where possible, data on age, sex and severity were also collated on these patients.

Similarly aged patients were recruited into the majority of trials (mean average age 71 (range 66 to 78)), reflecting the average age group of those suffering from a stroke. Almost all of the included trials recruited slightly more males than females (mean percentage males 53%).

Only 14 (all acute) of the included trials measured baseline severity using the NIHSS. The average NIHSS scores varied from eight up to 14, with a mean of 12. This reflects a moderate stroke severity.

2.4.2 Primary outcome

Table 2.5 shows details on the types of analysis used in each trial. The BI was used to measure functional outcome in 22 trials (47%), 18 used the mRS (38%), three used the 3Q scale (6%), one used the Rivermead scale (2%), two related trials used the Nottingham ADL scale (4%), and one trial used its own ordinal measure of ADL (2%).

The method of analysing functional outcome used in the original trial publications varied considerably. Twenty three (48.9%) trials assessed the treatment effect using a method which required the data to be collapsed into groups, e.g. chi-square test; 17 (36.2%) used a test based on comparing medians and four (8.5%) used a test which compared means; the remaining trials were unpublished so the method of analysis is not known.

Where data had been collapsed into two or more groups the cut points chosen varied significantly. Cuts points used on the BI were: <100, <95 and <60. For the mRS >1, >2 and >3 were used, with >2 being used in the most trials (five).

Thirty (65%) of the included trials were individually neutral, therefore they were part of a meta analysis showing a treatment effect. Fourteen trials (30%) showed a beneficial treatment effect and only two (5%) showed a harmful treatment effect.

TABLE 2.1

Trials selected for inclusion into the OAST project.

✓ data supplied by PI, * no data supplied but able to extract summary data from manuscript, × data not supplied.

Trial	Year	Intervention	Individual data supplied
AbESTT (Abciximab Emergent Stroke Treatment Trial (AbESTT) Investigators, 2005)	2005	Abcixmab	✓
APTIGANEL (Albers et al., 2001)	2001	Aptiganel Hydrochloride	×
ASK (Donnan et al., 1996)	1996	Streptokinase	✓
ASSIST 07 (Davis et al., 2000)	2000	Selfotel	✓
ASSIST 10 (Davis et al., 2000)	2000	Selfotel	✓
ATLANTIS A (Clark et al., 2000)	2000	Alteplase	✓
ATLANTIS B (Clark et al., 1999)	1999	Alteplase	✓
BEST (Barer et al., 1988)	1988	Low dose β blockade	✓
BEST pilot (Barer et al., 1988)	1988	Low dose β blockade	✓
Young (Young and Forster, 1992)	1992	Community physiotherapy	✓
CAST (CAST (Chinese Acute Stroke Trial) Collaborative Group, 1997)	1997	Aspirin	*
Citicoline 1 (Clark et al., 1997)	1997	Citicoline	✓
Citicoline 7 (Clark et al., 1999)	1999	Citicoline	✓
Citicoline 10 (Warach et al., 2000)	2000	Citicoline	✓
Citicoline 18 (Clark et al., 2001)	2001	Citicoline	✓
Corr (Corr and Bayer, 1995)	1995	Occupational therapy	✓

Day hospital trial (Hui et al., 1995)	1995	Day hospital	✗
DCLHb (Saxena et al., 1999)	1999	DCLHb	✓
DIAS (Hacke et al., 2005)	2005	Desmoteplase	✗
Dover stroke unit trial (Stevens and Ambler, 1982)	1982	Stroke Unit	✓
EAST (Enlimomab Acute Stroke Trial Investigators, 2001)	2001	Enlimomab	*
EBSELEN (Yamaguchi et al., 1998)	1998	Ebselen	✓
ECASS I (Hacke et al., 1995)	1995	Alteplase	✗
ECASS II (Hacke et al., 1998)	1998	Alteplase	✓
EDARAVONE (The Edaravone Acute Brain Infarction Study Group, 2003)	2003	Edaravone	*
Factor VII (Mayer et al., 2005)	2005	Recominant activated factor VII	✓
FISS (Kay et al., 1995)	1995	Low-molecular weight heparin	✓
FISS TRIS (Wong et al., 2005)	2005	Low-molecular weight heparin	*
FOOD 3 (Dennis, 2004)	2004	Percutaneous endoscopic gastrostomy	✓
Gilbertson (Gilbertson et al., 2000)	2000	Occupational therapy	✓
GLYCINE (Gusev et al., 2000)	2000	Glycine	✗
Goteberg stroke study (Fagerberg et al., 2000)	2000	Stroke unit	✗
Helsinki stroke unit trial (Kaste et al., 1995)	1995	Neurology ward	*
Hyperbaric oxygen (Rusyniak et al., 2003)	2003	Hyperbaric oxygen	✗
INWEST (Wahlgren et al., 1994)	1994	Nimodipine	✓
IST (International Stroke Trial Collaborative Group, 1997)	1997	Aspirin	✓
Indredavik (Indredavik et al., 1991)	1991	Stroke unit	✗

Kuopio stroke unit trial (Sivenius et al., 1985)	1985	Intensive treatment	✓
Lincoln (Juby et al., 1996)	1996	Rehabilitation unit	✓
Logan (Logan et al., 1997)	1997	Occupational therapy	✓
Lubeluzole (Grotta and The US and Canadian Lubeluzole Ischemic Stroke Study Group, 1997)	1997	Lubeluzole	✗
MAST-I (Candelise et al., 1995)	1995	Streptokinase / aspirin	✓
MAST-E (Multicenter Acute Stroke Trial - Europe Study Group, 1996)	1996	Streptokinase	*
NINDS (The National Institute Of Neurological Disorders And Stroke rt-Pa Stroke Study Group, 1995)	1995	Alteplase	✓
Orpington stroke unit trial (Kalra et al., 1993)	1993	Stroke unit	✗
Orpington stroke unit trial (Kalra and Eade, 1995)	1995	Stroke unit	✗
Orpington stroke unit trial (Kalra et al., 2000)	2000	Stroke unit / Stroke team	*
Parker (Parker et al., 2001)	2001	Occupational therapy	✓
PROACT I (del Zoppo et al., 1998)	1998	Recombinant Pro-Urokinase	✗
PROACT II (Furlan et al., 1999)	1999	Recombinant Pro-Urokinase	*
RANTTAS (The RANTTAS Investigators, 1996)	1996	Tirilazad	✓
RANTTAS II (Haley, 1998)	1998	Tirilazad	✓
Rodgers (Aitken et al., 1993)	1993	Geriatric unit	✓
STIPAS (The STIPAS Investigators, 1993)	1993	Tirilazad	✓
Ronning stroke unit trial (Ronning and Guldvog, 1998)	1998	Stroke unit	✗

STAT (Sherman et al., 2000)	2000	Ancrod	×
Streptokinase pilot (Morris et al., 1995)	1995	Streptokinase	*
Sulter stroke unit trial	2003	Stroke unit	×
TESS (Peters et al., 1996)	1996	Tirilazad	✓
TESS II (Orgogozo, 1995)	1995	Tirilazad	✓
WALKER1 (Walker et al., 1996)	1996	Dressing practice	✓
WALKER2 (Walker et al., 1999)	1999	Occupational therapy	✓
ZK200775 (Elting et al., 2002)	2002	AMPA Antagonist ZK200775	×

TABLE 2.2

Trials included in the OAST project with multiple treatment comparisons.

Trial	Comparison
BEST pilot (Barer et al., 1988)	Atenolol vs. placebo
	Propranolol vs. placebo
BEST (Barer et al., 1988)	Atenolol vs. placebo
	Propranolol vs. placebo
FISS (Kay et al., 1995)	High dose Nadroparin vs. placebo
	Low dose Nadroparin vs. placebo
INWEST (Wahlgren et al., 1994)	High dose nimodipine vs. placebo
	Low dose nimodipine vs. placebo
MAST-I (Candelise et al., 1995)	Aspirin vs. control
	Streptokinase vs. control
	Aspirin & Streptokinase vs. control
Parker rehabilitation trial (Parker et al., 2001)	Leisure therapy vs. control
	Activities of daily living therapy vs. control
Orpington stroke unit trial (Kalra et al., 2000)	Stroke team vs. domiciliary care
	Stroke unit vs. domiciliary care

TABLE 2.3

Description of scales used as the primary measure of functional outcome in the OAST project.

Scale	Range Dependent-Independent	Interval Size	Coding For Death
Barthel Index (Mahoney and Barthel, 1965)	0 - 100	5	-5
Rankin Scale (Rankin, 1957)	5 - 0	1	6
Q3 Scale (Lindley et al., 1994)	2 - 4	1	1
Nottingham ADL (Ebrahim et al., 1985)	0 - 10	1	-1
Rivermead ADL (Lincoln and Edmans, 1990)	0 - 16	1	-1
ADL Scale (Sivenius et al., 1985)	0 - 30	1	-1

TABLE 2.4

Baseline data for included trials.

Trial	Trial characteristics					Baseline		
	Sample size	Intervention	Time (hr)	Active groups	Follow up (mo)	Age (median [IQR])	Male (%)	Baseline severity (NIHSS) (median [IQR])
Acute								
AbESTT	400	Abciximab	6	1	3	69 [58-78]	56	9 [6-14]
ASK	340	Streptokinase	4	1	3	71 [64-77]	61	-
ASSIST 07	138	Selfotel	6	1	3	70 [63-76]	62	14 [8-18]
ASSIST 10	432	Selfotel	6	1	3	72 [63-77]	55	14 [9-20]
ATLANTIS A	142	Alteplase	0-6	1	3	70 [60-76]	68	11 [7-17]
ATLANTIS B	613	Alteplase	3-5	1	3	67 [59-75]	59	10 [6-15]
BEST pilot	65	Atenolol-propranolol	48	2	6	71 [62-81]	54	-
BEST	302	Atenolol-propranolol	48	2	6	71 [62-77]	52	-
CAST	20,655	Aspirin	48	1	1	-	-	-
Citicoline 1	259	Citicoline	24	1	3	70 [60-76]	47	11 [7-18]
Citicoline 7	394	Citicoline	24	1	3	73 [65-79]	47	11 [7-18]
Citicoline 10	100	Citicoline	24	1	3	74 [62-79]	49	12 [9-16]
Citicoline 18	899	Citicoline	24	1	3	71 [60-77]	52	13 [10-17]
DCLHb	85	DCLHb	18	1	3	-	-	-

EAST	623	Enlimomab	6	1	3	-	-	-
Ebselen	298	Ebselen	48	1	3	67 [59-74]	63	-
ECASS II	800	Alteplase	6	1	3	68 [59-74]	59	12 [8-16]
Edaravone	250	Edaravone	72	1	3	-	-	-
Factor VIIa	399	Factor VII	3	1	3	-	-	-
FISS	308	Nadroparin	48	2	3	68 [62-73]	58	-
FISS-TRIS	599	Heparin	48	1	6	-	-	-
FOOD 3	321	NG tube	-	1	6	78 [71-84]	45	-
INWEST	295	Nimodipine	24	2	3	73 [65-79]	46	-
IST	19,435	Aspirin	48	1	6	73 [65-80]	54	-
MAST-E	310	Streptokinase	6	1	6	-	-	-
MAST-I	622	Streptokinase-aspirin	6	3	6	71 [62-78]	54	-
NINDS	624	Alteplase	3	1	3	69 [60-75]	58	14 [9-20]
PROACT II	180	Prourokinase	6	1	3	-	-	-
RANTTAS	660	Tirilazad	6	1	3	74 [62-78]	55	9 [5-17]
RANTTAS II	126	Tirilazad	6	1	3	73 [65-81]	58	13 [9-18]
STIPAS	111	Tirilazad	12	1	3	68 [58-75]	56	8 [5-16]
Streptokinase pilot	20	Streptokinase	6	1	3	66 [62-75]	-	-
TESS	450	Tirilazad	6	1	3	72 [64-78]	56	-
TESS II	355	Tirilazad	?	1	3	70 [62-76]	60	-
Subtotal	51,610			40		72 [64-79]	54	12 [8-17]

Rehabilitation								
Corr	110	OT	-	1	12	75 [70-81]	37	-
Gilbertson	138	OT	-	1	6	71 [64-78]	45	-
Logan	111	OT	-	1	3	72 [66-79]	51	-
Parker	466	OT	-	2	12	72 [65-79]	58	-
Walker I	30	OT	-	1	3	70 [62-74]	53	-
Walker 2	185	OT	1 mo	1	6	75 [70-80]	51	-
Young	124	PT	-	1	6	70 [65-76]	56	-
Subtotal	1,164			8		72 [66-79]	52	-
Stroke unit:								
Dover	235	SU	1-2 w	1	3	74 [67-79]	41	-
Helsinki	232	SU	-	1	12	-	-	-
Kuopio	94	SU	1 w	1	3	72 [67-78]	38	-
Nottingham	315	SU	5 w	1	3	69 [62-75]	59	-
Orpington	457	SU	-	2	12	-	-	-
Newcastle	66	SU	72	1	6	77 [73-82]	50	-
Subtotal	1,399			7		72 [65-78]	49	-
Total	54,173			55		72 [64-79]	54	12 [8-17]

IQR: Inter quartile range; NIHSS: National Institute of Health Stroke Scale; hr: hours; w: weeks; mo: months; SU: stroke unit; OT: occupational therapy; PT: physiotherapy

TABLE 2.5

Primary outcome for included trials.

	Barthel Index (median [IQR])	Rankin Scale (median [IQR])	3Q (median [IQR])	Death rate (%) per month (control group)	Outcome scale	Type of analysis	Analysis approach used in the primary publication	Trial result (+/0/-)
Acute								
AbESTT	-	2 [1-4]	-	4.2	mRS	O	Ordinal logistic regression	0
ASK	62.5 [-5-100]	-	-	6.8	BI	D	Chi square test (BI <60)	0
ASSIST 07	70 [15-100]	-	-	5.0	BI	D	Cochran Mantel Haenszel test	0
ASSIST 10	55 [5-100]	-	-	6.1	BI	D	Cochran Mantel Haenszel test	0
ATLANTIS A	90 [20-100]	-	-	2.3	BI	D	Binomial test (BI <100)	0
ATLANTIS B	95 [50-100]	2 [1-4]	-	2.3	mRS	D	Binomial test (mRS >1)	0

BEST pilot	-	-	-	5.8	Nottingham ADL	?	Unpublished	0
BEST	-	-	-	3.8	Nottingham ADL	O	Kruskal-Wallis test	0
CAST	-	-	3 [2-4]	3.9	3Q Scale	D	Chi square test	+
Citicoline 01	80 [20-100]	4 [2-5]	-	5.1	BI	D	Logistic regression (0, 5-40, 45-60, 60-80, 85-100)	+
Citicoline 07	75 [5-100]	2 [1-4]	-	6.4	BI	D	Logistic regression (0, 5-40, 45-60, 60-80, 85-100)	0
Citicoline 10	70 [15-100]	3 [2-4]	-	2.8	mRS	O	Wilcoxon test	0
Citicoline 18	75 [10-100]	3 [1-4]	-	5.9	BI	D	Cochran Mantel Haenszel test	0
DCLHb	-	3 [2-4]	-	3.0	mRS	D	Chi square test (mRS >2)	-
EAST	-	3 [1-5]	-	5.4	mRS	O	Wilcoxon test	-
Ebselen				2.8	BI	O	Wilcoxon test	0
ECASS II	-	2 [1-4]	-	3.4	mRS	D	Fishers exact test (mRS >1)	0
Edaravone	-	2 [1-3]	-	1.5	mRS	O	Wilcoxon test	+
Factor VIIa	-	4 [2-5]	-	9.7	mRS	O	Adjusted cumulative	+

							logit model	
FISS	-	-	2 [2-3]	4.8	3Q Scale	O	Chi square test for trend (dichotomised >2)	+
FISS-TRIS	-	2 [1-3]	-	0.9	mRS	?	Unpublished	0
FOOD 3	-	5 [5-6]	-	8.1	mRS	D	Logistic regression (dichotomised >3)	0
INWEST	25 [-5-85]	-	-	9.3	BI	O	Wilcoxon test	-
IST	-	-	2 [2-3]	3.7	3Q Scale	D	Chi square test	0
MAST-E	-	5 [3-6]	-	6.4	mRS	D	Chi square test (mRS >2)	0
MAST-I	-	4 [1-6]	-	4.8	mRS	D	Chi square test (mRS >2)	0
NINDS	85 [7.5-95]	3 [1-5]	-	6.8	mRS	D	GEE global outcome (BI <95, RS >1, GOS >1 NIHSS >1)	+

PROACT II	-	3 [2-6]	-	9.0	mRS	D	Cochran Mantel Haenszel test	+
RANTTAS	95 [25-100]	-	-	4.9	BI	O	Kruskal-Wallis test	0
RANTTAS II	65 [-5-100]	-	-	10.8	BI	O	Kruskal-Wallis test	0
STIPAS	95 [55-100]	-	-	1.2	BI	D	Chi square test (BI <60)	0
Streptokinase pilot	48.5 [-5-100]	-	-	10	BI	O	Kruskal-Wallis test	0
TESS	65 [0-100]	-	-	7.2	BI	D	Chi square test	0
TESS II	75 [10-100]	-	-	5.6	BI	?	Unpublished	0
Rehabilitation								
Corr	55 [15-75]	-	-	1.8	BI	O	Mann Whitney U test	0
Gilbertson	80 [60-90]	3 [2-4]	-	1.2	BI	C	t-test	0
Logan	80 [55-90]	-	-	4.5	BI	O	Wilcoxon test	0
Parker	80 [60-95]	3 [1-4]	-	0.8	mRS	C	Multiple linear regression	0
Walker I	-	-	-	0	Rivermead	O	Wilcoxon test	+
Walker II	95 [85-100]	-	-	0	BI	O	Wilcoxon test	+
Young	85 [67.5-95]	-	-	0	BI	O	Mann Whitney U test	+
Stroke unit								

Dover	-	4 [2-6]	-	14.3	mRS	C	Comparison of average score	0
Helsinki	-	2 [1-5]		1.8	mRS	O	Mann Whitney U test	+
Kuopio	-	-	-	1.6	Trial specific ADL	C	ANCOVA	+
Nottingham	75 [45-90]	-	-	2.9	BI	O	Wilcoxon test	+
Orpington	-	2 [1-4]	-	1.3	mRS	D	Chi square test (mRS >3)	+
Newcastle	30 [15-50]	-	-	6	BI	O	Wilcoxon test	0

D: Dichotomised or data collapsed into multiple groups; O: Ordinal method; C: Continuous method +: Beneficial intervention effect; -: Harmful intervention effect; 0: No intervention effect but part of meta analysis showing a beneficial or harmful treatment effect

CHAPTER 3

RESULTS

COMPARISON OF UNIVARIATE STATISTICAL METHODS

PUBLICATIONS/PRESENTATIONS CONTRIBUTING TO THIS CHAPTER

The Optimising Analysis of Stroke Trials (OAST) Collaboration, Bath P.M.W, **Gray L.J**, Collier T, Pocock S, Carpenter J. (2007) Can we improve the statistical analysis of stroke trials? Statistical reanalysis of functional outcomes in stroke trials. *Stroke. 38: 1911-1915.*

Bath P.M.W, **Gray L.J**. (2008) Response to Letter by Miller and Palesch. *Stroke: 39:e15*

Gray L.J, Bath P.M.W, Collier T, OAST Collaborators (2005) Optimising the analysis of functional outcome in stroke clinical trials. *Oral presentation at the European Stroke Conference, Italy. May 2005. Cerebrovascular Diseases 19(suppl 2): 16.*

Gray L.J, Bath P.M.W, Collier T, OAST Collaborators (2006) Optimising the analysis of functional outcome in stroke clinical trials. *Poster presentation at Institute of Neuroscience Poster Day, University of Nottingham, UK, November 2006.*
Student poster prize winner

Gray L.J, Bath P.M.W, Collier T, OAST Collaborators (2006) Optimising the analysis of functional outcome in stroke clinical trials. *Poster presentation at GSK 2006 – UK – Statistics & Programming Practice Annual Conference, Ware, UK, October 2006.*
Student poster prize winner

3.1 INTRODUCTION

As discussed previously, there is little agreement as to the 'best' way of analysing data from functional outcome scales. Many trialists advocate dichotomising scales into two groups and comparing those with a 'good' to those with a 'bad' outcome, as this is thought to be clinically meaningful and easier to interpret. However, there is little consensus on where scales should be cut to create these groups or whether this actually matters (Wardlaw et al., 2000).

Song *et al* have encouraged the use of parametric methods, such as the t-test (Song et al., 2006), while others have categorised these as inappropriate for ordinal data (Roberts et al., 1998). To date no research has considered standard non-parametric methods, such as the Wilcoxon test, although bootstrapping has been considered as a viable option (Stingele et al., 2001).

This chapter aims to identify which statistical methods might optimise the analysis of data from functional outcome scales in stroke trials. This work focuses on univariate methods which do not take account of potentially confounding covariates. Methods such as the 'patient specific analysis' or those which adjust for covariates will be addressed in subsequent chapters.

3.2 METHODS

3.2.1 Trial data

All 55 data sets in the OAST project were included in this analysis as only data on functional outcome and treatment assignment were required. See Chapter 2 for details on the OAST data set.

3.2.2 Statistical tests

Sixteen different statistical tests for assessing treatment effect were compared. Some of these required the data to be collapsed into groups (such as the 2x2 chi square test) while others used the original ordinal data (such as Wilcoxon test and t-test). Statistical tests which dichotomised data were assessed with data collapsed at different places, e.g. mRS 0,1 versus 2-6, 0-2 versus 3-6 and 0-5 versus 6; see Table 3.1 for a complete listing of all dichotomisations for all the tests compared. The tests compared are discussed in the subsequent sections with technical detail being given for the less well known tests.

- Chi-square test

The chi-square test is currently the most common method of analysis used in stroke trials. Here the chi square test is used under five different conditions.

(i) 2x2 test - dead or poor outcome versus good outcome

(ii) 2x2 test - dead or poor outcome versus excellent outcome

(iii) 2x2 test - dead versus alive

(iv) 2x3 test (unordered data) - dead versus poor outcome versus good outcome

(v) 2x4 test (unordered data) - dead versus poor outcome versus good outcome versus excellent outcome

Not all of these comparisons could be carried out for all trials. For example, in some of the rehabilitation trials no patients died and therefore these trials are not included in the conditions where vital status is assessed. The same will apply to the other statistical tests being compared where vital status is assessed. Chi-square tests were performed without continuity correction (Hollander and Wolfe, 1999) since most trials enrolled more than 100 patients. Figures 3.1 and 3.2 show pictorially the cuts used on the mRS and BI and Table 3.1 defines the cuts used.

FIGURE 3.1

A diagram of the various cut points used on the modified Rankin Scale.

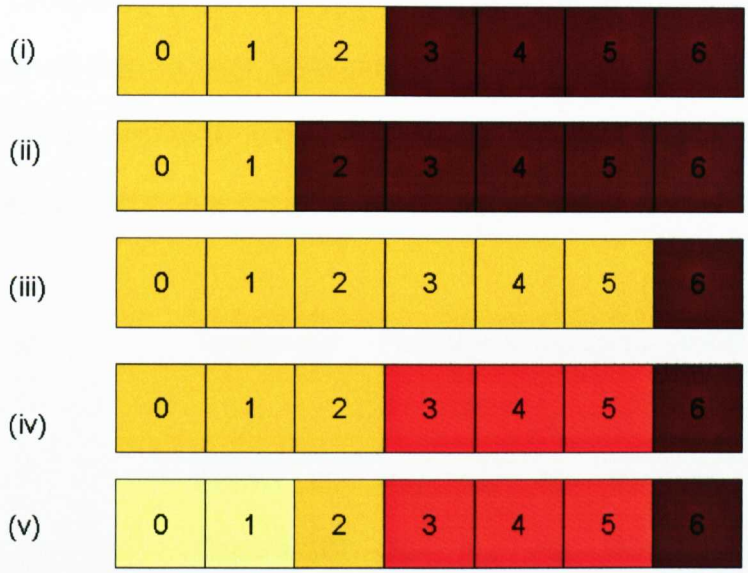
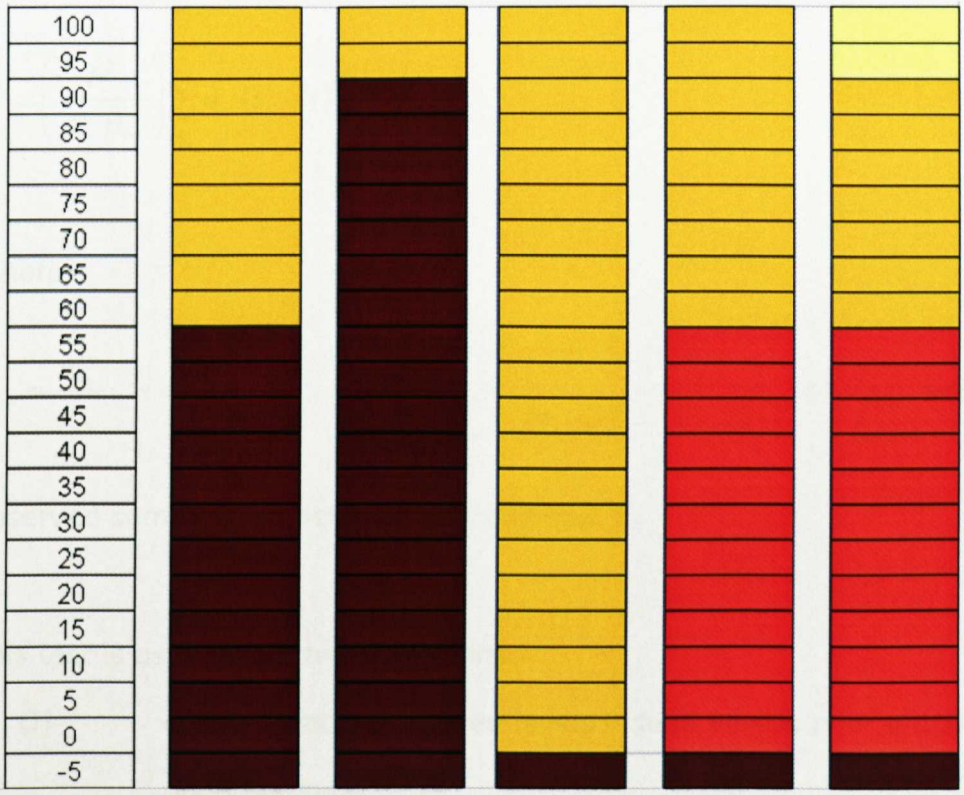


FIGURE 3.2

A diagram of the various cut points used on the Barthel Index.



- Cochran-Armitage trend test

This test is similar to the chi square test but takes into account the ordering across categories (Agresti, 2002). The 2x*l* table below shows a summary of the data gained from a trial assessing two treatments using an ordinal scale. Here the Cochran-Armitage trend test tests whether there is a linear trend in binomial proportions across the levels of functional outcome.

Treatment	Functional outcome (mRS)				
	0	1	*i*	*l*	total
Active (1)	n_{10}	n_{11}	n_{1i}	n_{1l}	n_{1+}
Control (0)	n_{00}	n_{01}	n_{0i}	n_{0l}	n_{0+}
Total	n_{+0}	n_{+1}	n_{+i}	n_{+1}	n_{++}

The test statistic for the Cochran-Armitage trend test is given below, where s denotes the functional outcome score:

$$z^2 = \left(\frac{b^2}{p_{1+}p_{0+}}\right)\sum_i n_{+i}(s_i - \bar{s})^2$$

Where

$p_{1+} = \frac{n_{1+}}{n_{++}}$, $\bar{s} = \frac{\sum_i n_{+i}s_i}{n}$, $b = \frac{\sum_i n+i(p_{1|i} - p_{1+})(s_i - \bar{s})}{\sum_i n_{+i}(s_i - \bar{s})^2}$, $p_{0+} = 1 - p_{1+}$ and $p_{1|i}$ is the observed sample proportion of the response 1.

This test is used under two conditions:

(i) ordered data with three levels - dead versus poor outcome versus good outcome

(ii) ordered data with four levels - dead versus poor outcome versus good outcome versus excellent outcome

- Ordinal logistic regression

Ordinal logistic regression can be used when the dependent variable is categorical and ordered. This model is also referred to as the proportional odds model and the cumulative logit model. It is similar to logistic regression but simultaneously estimates multiple end points instead of just one. The number of end points estimated is equivalent to the number of ordered categories minus one. For example, if the mRS was the dependent variable of interest it would compare the following j categories: 0 versus 1,2,3,4,5,6; 0,1 versus 2,3,4,5,6; 0,1,2 versus 3,4,5,6; 0,1,2,3 versus 4,5,6; 0,1,2,3,4 versus 5,6; 0,1,2,3,4,5 versus 6.

Ordinal logistic regression provides one overall estimate for each covariate in the model and not one for each cut point. This assumes that the overall odds ratio is constant no matter which cut is taken. So, for example, the odds ratio for the treatment effect would be interpreted as the odds of being in category j or above for all choices of j comparing treatment 1 to treatment 0 (Agresti, 1999).

The ordinal logistic regression model has the following form:

$$Pr(Y \leq y_j \mid x) = \frac{\exp(\alpha_j - x'\beta)}{1 + \exp(\alpha_j - x'\beta)},$$

$$j = 1,2.....,k$$

Here the regression coefficient β is not dependent on the level of the response variable j. This implies that the relationship between x and Y is independent of j. This independence is called the 'proportional odds assumption'.

This method is used in three different ways with the OAST data:

(i) Raw data

(ii) Three levels - dead versus poor outcome versus good outcome

(iii) Four levels - dead versus poor outcome versus good versus excellent outcome

- t-test

The t-test assesses whether the means of two independent samples are equal. This is a parametric test and makes the assumption that the samples are normally distributed. The t-test can be used under two different conditions, either assuming equal variance (pooled) or not (unpooled). Here the version of the test which does not assume equal variances was used (unpooled t-test).

- Median test

The median test assesses whether two independent groups have been drawn from a population with the same median. Although the median test is thought of as a non-parametric test it is basically a chi-square test which uses the combined median to determine where the data are collapsed into two groups (Siegel and Castellan, 1988).

- Wilcoxon test

The Wilcoxon test (also referred to as the Mann-Whitney U test) is the non-parametric equivalent of the t-test and tests whether two independent groups have been drawn from the same population. The method allowing adjustment for ties (using the average value) was used, as many patients will share the same outcome score (Siegel and Castellan, 1988, Wilcoxon, 1945).

- Robust rank test

The robust rank test is an alternative to the Wilcoxon test, testing whether the median of one group is equal to another. However, unlike the Wilcoxon test, it does not assume that the distributions of the two groups are equal, i.e. it makes no assumptions about the variance of the two groups (Fligner and Policello, 1981, Siegel and Castellan, 1988). The test statistic for the robust rank test is given below.

$$U = \frac{mU(YX) - nU(XY)}{2\sqrt{V_x + V_y + U(XY)U(YX)}}$$

Where m is the number of patients in group X and n is the number in group Y. $U(XY)$ and $U(YX)$ are based on the mean placements of the data, the following example shows how they are calculated.

In this example m =3 and n =4.

Treatment X: 2 4 6

Control Y: 0 1 3 5

Which has rank order

mRS	0	1	2	3	4	5	6
Group	Y	Y	X	Y	X	Y	X

$U(YX)$ is calculated from the mean number of Y values which rank lower than each X value, as shown below:

X_i	$U(YX_i)$
2	2
4	3
6	4

$$U(XY) = \sum_{i=1}^{m} \frac{U(YX_i)}{m}$$

A similar calculation yields $U(XY)$.

V_x and V_y are indices of variability for $U(YX_i)$ and $U(XY_j)$, and are calculated:

$$V_x = \sum_{i=1}^{m} [U(YX_i) - U(YX)]^2 \text{ and } V_y = \sum_{j=1}^{n} [U(XY_j) - U(XY)]^2$$

- Kolmogorov-Smirnov test

This is a test of whether two independent samples have been drawn from a population with the same distribution. It has the advantage of making no assumption about the distribution of data (Siegel and Castellan, 1988). The Kolmogorov-Smirnov test compares the cumulative frequency distributions of each group and looks for the largest difference between these. The test statistic for a two sided Kolmogorov-Smirnov test is as follows:

$$D_{m,n} = \max[S_m(X) - S_n(X)]$$

where the observed cumulative distribution for one sample (of size m) is $S_m(X) = K/m$, where K is the number of data points greater than or equal to

X, the observed cumulative distribution for the other sample is $S_n(X) = K/n$ (Siegel and Castellan, 1988).

- Bootstrapping the difference in mean rank

Bootstrapping is a computationally intensive method which involves resampling data from a given data set. The main advantage of bootstrapping over more traditional methods is that it does not make any assumptions about the distribution of the data. Here the difference in mean rank is bootstrapped; the procedure for doing this is outlined below and is taken from the re-analysis of the ECASS II data (Efron and Tibshirani, 1993, Stingele et al., 2001):

1. Take a data set, which contains *N* observations with sample size p in the control group and q in the treated group
2. Draw a sample with replacement of size *N* (using replacement means that some of the original observations may appear in the new sample more than once and some not at all)
3. The first p values are assigned to the control group and the next q values to the treatment group
4. Estimate the parameter of interest (here the difference in mean rank) and store the result
5. Repeat 2 and 3 many times (here three sets of 3,000 iterations were used)
6. Compare the distribution of the stored results to the actual point estimate from the original data set

3.2.3 Excluded statistical tests

Three non-parametric tests were excluded as they are inappropriate for assessing differences between groups of ordinal data or a close alternative is being used:

- Wald-Wolfowitz runs test

Assesses if the number of 'runs' in an ordering is random or not, where a run is repetition in a sequence. If the two groups are from different distributions the number of runs would be mutually independent (Conover, 1971).

- Siegel-Tukey test

Tests for differences in scale between two groups (Siegel and Castellan, 1988).

- Cramer-von Mises two-sample test

This was excluded as it is very similar to the Kolmogorov-Smirnov test (Conover, 1971).

3.2.4 Comparison of statistical tests

Where possible, each data set was analysed using each statistical test. The absolute z scores were then ordered within each trial and given a rank, with the lowest rank given to the test which produced the most significant result, i.e. the largest absolute z score, within that trial. A two-way analysis of variance test (Friedman's) was then used to assess which statistical test had produced the lowest ranks. The statistical tests were then ordered in terms of their efficiency in identifying treatment effects using Duncan's multiple range test (Duncan, 1955).

The number of statistically significant (at 5%) results were also assessed for each test compared.

To assess the validity and reliability of the results, a number of supplementary analyses were carried out. Firstly, the comparison of statistical tests was repeated within sub group of trials sharing similar characteristics. Secondly, the statistical assumptions of the tests were assessed. Lastly, the sensitivity of the tests was explored to make sure treatment effects were only detected when they truly existed (the type one error rate). The availability of the tests in popular statistical packages (SAS, Stata, SPSS) was also assessed. These analyses are discussed in more detail below.

3.2.5 Sub group analysis

Sub group analyses were performed by assessing the efficiency of the different tests for differing trial characteristics:

- Type of intervention tested (thrombolysis, anticoagulation, antihypertensive, antiplatelet, feeding, neuroprotection, occupational therapy, procoagulant, and stroke unit)
- Trial setting (acute drug treatment, rehabilitation, stroke unit)
- Trial size (<500, ≥500 participants), 500 falls between the mean and median trial size included and is used to define smaller and larger trials
- Time between randomisation and stroke onset (≤6, >6 hours), sub acute trials (≤6 hours) tend to include more severe patients and using more aggressive interventions
- Patient age (≤70, >70 years), this cut was chosen as the median age of patients in the included trials was 71
- Baseline severity (median control group death rate adjusted for length of follow up, ≤0.05, >0.05), this was used because baseline severity was not available for many trials or had been measured using a variety of scales
- Outcome measure (BI, mRS, 3Q)
- Length of follow up (≤3 months, >3 months)
- Intervention result, as published (beneficial, harmful)

3.2.6 Statistical assumptions

The principal statistical assumptions underlying the tests which performed well were assessed to ensure that their use was appropriate for stroke trial data.

- Ordinal logistic regression - proportionality of odds across response categories

Ordinal logistic regression makes the assumption that the odds ratio for the difference between the treatments groups is constant across categories of the outcome. Because of this assumption, this model is sometimes referred to as the proportional odds model. This was tested using a likelihood ratio test, comparing the multinomial logistic model to the ordinal logistic regression model.

- t-test – normal distribution of outcome scores

The t-test assumes the data is normally distributed. Normality was assessed both visually by plotting histograms and using the Shapiro-Wilk test (Shapiro and Wilk, 1965). The equality of variances assumption was not required as the unpooled version of the t-test was used. But to confirm the usage of this version, the F-test was used to see if it might be possible to also use the pooled version of the test.

- Robust rank test – independence of treatment groups

The robust rank test is a non-parametric test and only assumes independence of groups.

3.2.7 Type 1 error rate

A type 1 error occurs when a statistical difference between two groups is observed where no difference truly exists. The type 1 error rate is usually set at 5%, i.e. 5% chance that the observed variation in the data is not true. It is conceivable that an overly sensitive statistical test might have an inflated type 1 error and therefore find statistically significant differences, where none truly exist, greater than 5% of the time. The type I error rate was assessed for the three most efficient statistical tests, using data from three representative trials including one each of the three most used measures of functional outcome (BI: RANTTAS, mRS: NINDS, 3Q: IST). From these data, 1,000 data sets were generated, using random sampling with replacement, in which any treatment difference could have occurred only by chance. Tests maintaining adherence to the nominal type I error rate would expect to see a significant result in around 50 of the 1,000 data sets.

3.2.8 Availability of tests

Currently many use the chi square test for analysing dichotomised functional outcome data. A contributing factor to this may be the ease of use and interpretation. The chi square test is available in every statistical package and can also be calculated online. The availability of each test being compared was assessed for three commonly used statistical packages: SAS, Stata and SPSS.

3.3 RESULTS

3.3.1 Trial characteristics

As previously discussed all 55 data sets were included in this analysis. The characteristics of these trials are presented in Chapter 2.

3.3.2 Comparison of statistical tests

The statistical tests assessed differed significantly in the results they gave for each trial (two way ANOVA $p<0.0001$). The ordering of the tests showed that those which maintain and analyse the original ordinal data generally perform better than those which collapse the data into two or more groups. The most efficient tests included ordinal logistic regression, t-test, robust rank test, bootstrapping the difference in mean rank, and the Wilcoxon test (Table 3.2). All of the tests which do not take into account the ordering of the data ranked the lower.

Where tests had been repeated under different conditions (dichotomous, three levels, four levels, raw data) and the ordering of the groups was assessed, a greater the number of levels resulted in greater statistical power. This is reflected in the results for ordinal logistic regression and the Cochran-Armitage test for trend. The same pattern was not seen for the unordered chi square tests.

The median test, which dichotomises the data at the median, which some have suggested increases power, performed poorly. The lowest performing test was the Kolmogorov-Smirnov test.

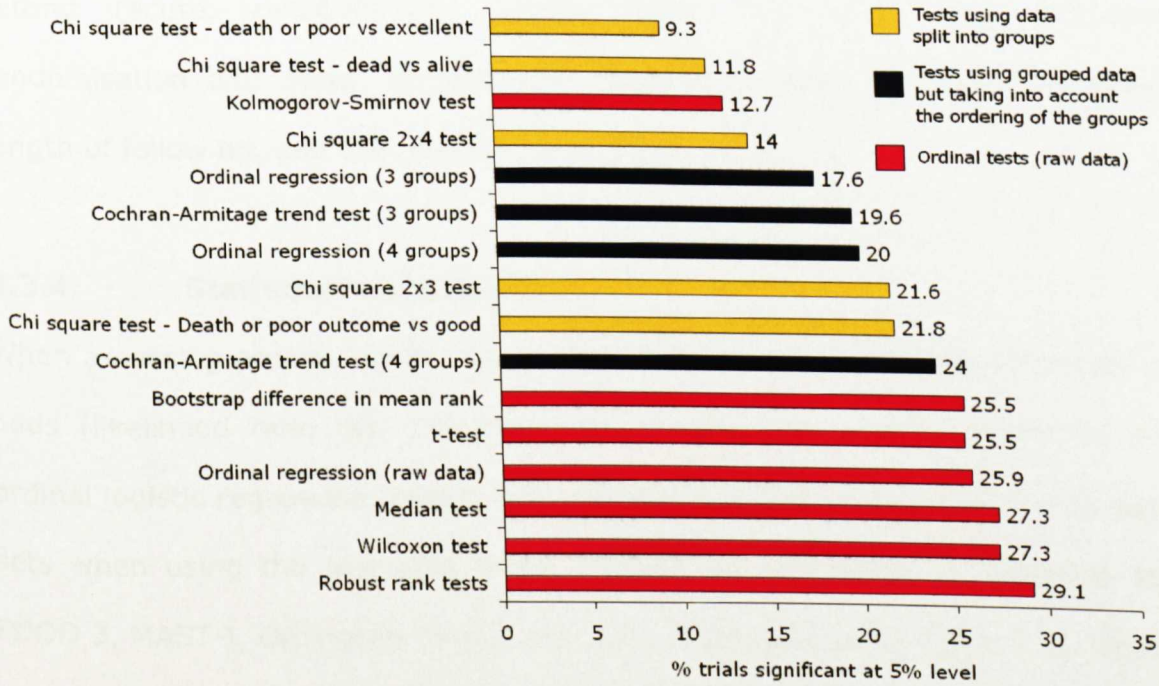
When assessed by how many trials were statistically significant, those tests which did not collapse the data into groups again out-performed the other

approaches; for example, ordinal logistic regression (using raw data) gave a statistically significant result in 25.9% of trials whereas the 2x2 chi-square test comparing death or poor outcome to an excellent outcome only gave a significant result in 9.3% of the trials (Figure 3.3).

Interestingly, the median test performed well here compared to the two way analysis of variance results. This may be because the median test collapses the data at the median and therefore compares groups of roughly even size.

FIGURE 3.3

The number of statistically significant results found for each test, where $p<0.05$ signifies statistical significance.



3.3.3 Sub group analysis

Table 3.3 shows the two way ANOVA results and test rankings by intervention. Statistically significant differences in the ranking of the tests were seen for thrombolysis, anticoagulation, antiplatelet, neuroprotection and occupational therapy. Ordinal regression performed well in trials of antiplatelets, feeding, neuroprotection, occupational therapy and stroke unit trials. Ordinal methods seem to perform poorly in trials of thrombolysis, in contrast, the four level and three level chi square tests performed well for these trials. Ordinal logistic regression using raw data ranked 11th out of the 16 tests for trials of thrombolysis.

The sub group analysis showed similar ordering of tests irrespective of the trial setting (acute, rehabilitation, stroke unit), trial size, time between randomisation and onset, patient age, baseline severity, outcome measure, length of follow up, and trial result (Table 3.4).

3.3.4 Statistical assumptions

When assessing ordinal logistic regression, the assumption of proportionality of odds (likelihood ratio test comparing the multinomial logistic model to the ordinal logistic regression model) was not met ($p<0.05$) in eight of the 55 data sets when using the raw data (ASK, ASSIST 07, ATLANTIS A, citicoline 10, FOOD 3, MAST-I, Orpington domiciliary care, Orpington team, Table 3.5). Three of these eight trials were testing thrombolytics, and this may be part of the reason why ordinal logistic regression may not perform well in thrombolysis trials (Table 3.5). Similar results were seen for the ordinal logistic regression when based on three and four levels of data. Figures 3.4 a-c show the distribution of the MAST-I data for aspirin and streptokinase versus control data. These plots show that aspirin and streptokinase increase the proportion

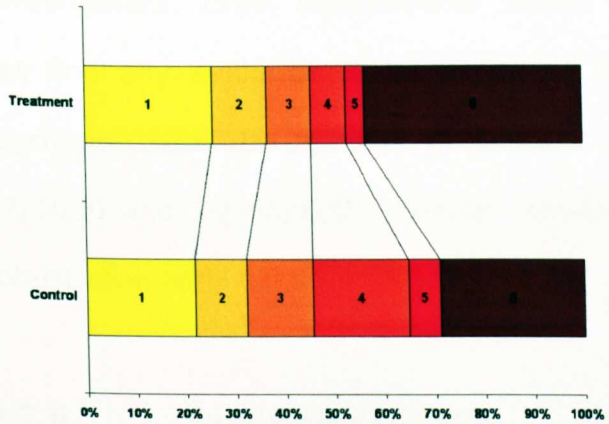
of patients with a good outcome (26% of patients score an mRS of one in the treated group compared to 22% in the control group) but there is also an increase in the proportion of patients who die (44% died in the treated group compared to 29% in the control group), hence the proportional odds assumption is not met.

The assumption of normality required for the t-test did not hold for all but one of the data sets (Table 3.6). The equality of variance F test was statistically significant in 12/55 data sets, implying that using the unpooled version of the t-test was a necessary approach. Additionally, when the two way ANOVA analysis was repeated with both the pooled and unpooled t-test included, no difference was found between the two (Table 3.7).

In contrast, the assumption of independence of groups for the robust rank test was met in all cases.

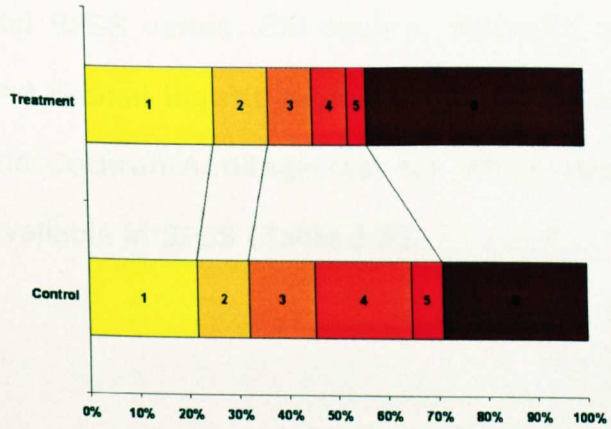
FIGURES 3.4 A-C

Distribution of the mRS in the factorial MAST-I trial of aspirin and streptokinase versus control, demonstrating non proportional odds (Multicentre Acute Stroke Trial-Italy (MAST-I) Group, 1995).



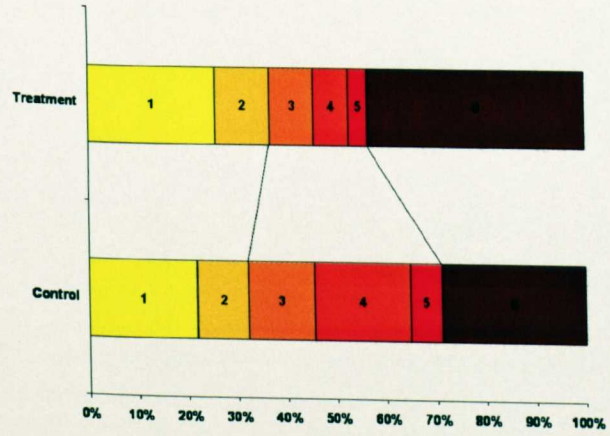
3.4.a

Raw data



3.4.b

Four levels



3.4.c

Three levels

3.3.5 Type 1 error rate

Table 3.8 shows the type 1 error rate results. Analysis of 1,000 re-sampled random data sets from the three trials (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995, The RANTTAS Investigators, 1996, International Stroke Trial Collaborative Group, 1997) did not find any evidence of an increased type I error rate for ordinal logistic regression with the number of 'positive' data sets being: BI 39/1000; mRS 57/1000 and 3Q 56/1000. Similar results were found for both the t-test and robust rank test.

3.3.6 Availability of tests

The availability of the compared tests within the three packages - SAS, Stata and SPSS varied. Chi square, Wilcoxon, median, Kolmogorov-Smirnov, t-test and ordinal logistic regression were available in all three packages. However, the Cochran-Armitage test for trend, robust rank and bootstrapping are not available in SPSS (Table 3.9).

3.4 DISCUSSION

These results show that statistical approaches which analyse the original ordinal data for functional outcome perform better than those which work on pre-processed data, which has been collapsed into two or more groups. In particular, ordinal logistic regression, the t-test, the robust rank test, bootstrapping the difference in mean rank, and the Wilcoxon test performed well and appear to be useful irrespective of the type of stroke trial, patient or setting. However, ordinal methods may not be appropriate for trials of thrombolytic agents.

Although individual tests based on dichotomised data using chi-square analysis (e.g. 'dead/dependent' versus 'independent') were effective for some data sets, they performed poorly in many and therefore cannot be recommended as a general solution for analysing stroke trials. From a historical perspective, it is quite possible that trials which collapsed mRS or BI into two groups may have used a sub-optimal analysis, and this may have contributed to false neutral findings in some cases in the past. For example, The International Stroke Trial comparison of aspirin against control was neutral on its primary outcome but shows a statistically significant treatment effect when re-analysed using ordinal logistic regression on the raw data (International Stroke Trial Collaborative Group, 1997).

Ordinal logistic regression assumes the intervention will exert effects of similar magnitude and direction at each transition of the outcome scale, i.e. 'proportionality of odds'. This is unlikely to be the case for treatments where symmetrical benefits occur (i.e. the intervention is effective across a spectrum of severity) but hazard is asymmetrical tending to effect mainly those with severe stroke. Thrombolysis is an example and its overall effect is to reduce

dependency and, to a lesser extent, increase death (largely through promoting fatal intracerebral haemorrhage). Specifically, thrombolysis probably reduces dependency across all levels of the mRS, but increases haemorrhage in patients with severe stroke who are likely to have a poor outcome. Hence, thrombolysis may be considered, in the context of stroke severity, to have symmetrical effects on efficacy but asymmetrical effects on hazard, and therefore ordinal methods are probably not appropriate.

Several comments can be made about this part of the OAST project. Firstly, the search for all possible statistical tests relevant to the problem of analysing ordered categorical data was not exhaustive. Instead, the focus was concentrated on those approaches which are available in standard statistical textbooks and computer packages (all tests assessed were available in SAS and Stata). Additionally, some tests used in recent trials could not be included, e.g. patient specific outcomes and Cochran Mantel-Haenszel test, since these require access to individual data for both baseline and outcome variables, and these data were not available uniformly. These will be assessed in a subsequent chapter using a sub-set of the OAST data.

Secondly, some of the statistical assumptions underlying the more efficient tests were not met in all trials. For example, the t-test assumes data are normally distributed while ordinal logistic regression assumes that any treatment effect is similar across outcome levels. Nevertheless, the robustness of these tests to deviations from their underlying assumptions means that they remain relevant for analysing functional outcome data from stroke trials. Indeed some have recommended the use of the t-test for measures with seven or more ordered categories (Walters et al., 2001).

If alternative approaches to analysing functional outcome data are to be used in the future, it is pertinent to ask how sample size should be calculated at the trial design stage. Historically, most calculations assumed that functional outcome would be dichotomised and analysed using a Chi-square test approach (Weaver et al., 2004). Although future trials could continue to calculate sample size in the same way (and then gain extra power by analysing their data using an ordinal approach), specific sample size calculations are available when data are to be analysed using ordinal logistic regression, or the Wilcoxon or t-tests. Ideally, it might be considered that the extra power gained by using an ordinal statistical approach should not be used to reduce sample size; stroke trials have been too small in the past, as shown in a recent meta analysis (Weaver et al., 2004), and this may also have contributed to the failure of some studies. Assessment of sample size in the OAST data set will be addressed in Chapter 4.

A further issue with using a statistical test which analyses ordered categorical data is how to report the results to patients, carers, clinicians, and health policy makers. The results of dichotomous tests may be summarised easily as the proportion of patients who benefit (or suffer) with a treatment, i.e. alteplase reduced absolute death or dependency (mRS>1) by 13% in the NINDS part two trial (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995). In contrast, ordinal tests will need to be presented as the average absolute improvement in outcome, e.g. alteplase improved the mRS by 1 (of 7) point and BI by 22.5 (of 100) points (unpublished). Alternatively, the combined odds ratio and its confidence intervals would have to be reported if ordinal logistic regression was used. In this respect, health consumers will need to decide what differences in mRS and BI are worthwhile, both clinically and in terms of health economics. In reality, it is reasonable to present the effect on functional outcome using both absolute percentage change (as a secondary

outcome) and mean or median change in functional outcome score (as the primary outcome).

Interestingly, a recent study which sent questionnaires about the design and analysis of stroke trials to 300 neurologists found that of the 152 who replied, 54% would chose a method of analysis for the mRS which looked for changes across the whole scale whereas only 39% would choose a dichotomous endpoint; although, 20% still felt they did not fully understand the results from shift analyses (Savitz et al., 2008).

3.5 SUMMARY

These results suggest that ongoing and future trials should consider using statistical approaches which utilise the original ordered categorical data in the primary analysis of functional outcome measures. Such ordinal tests include ordinal logistic regression, the robust rank test, bootstrapping the difference in mean rank, and the Wilcoxon test; the t-test may also be used although its assumptions were not met in many trials.

TABLE 3.1

Definitions of outcomes on the Barthel Index, modified Rankin Scale, Three Questions and Nottingham Activities of Daily Living Scale.

	Barthel Index	Modified Rankin Scale	3 Questions Scale	Nottingham ADL
Good vs. poor	≥60 vs. <60	≤2 vs. >2	≥3 vs. <3	≥6 vs. <6
Excellent vs. poor	≥95 vs. <95	≤1 vs. >1	4 vs. <4	≥9 vs. <9
Alive vs. dead	≥0 vs. -5	≤5 vs. 6	>1 vs. 1	≥0 vs. -1
Good vs. poor vs. dead	≥60 vs. <60-≥0 vs. -5	≤2 vs. >2-≤5 vs. 6	≥3 vs. <3->1 vs. 1	≥6 vs. <6-≥0 vs. -1
Excellent vs. good vs. poor vs. dead	≥95 vs. <95-≥60 vs. <60-≥0 vs. -5	≤1 vs. <1-≤2 vs. >2-≤5 vs. 6	4 vs. 3 vs. <3->1 vs. 1	≥9 vs. >9-≥6 vs. <6-≥0 vs. -1

TABLE 3.2

Comparison of rank scores for 16 statistical tests; lower ranks imply the test is more efficient. Analysis by non parametric two-way ANOVA and Duncan's multiple range test; tests joined by the same band are not significantly different from each other at $p<0.05$.

Test	No. of data sets	Mean rank	Banding
Ordinal logistic regression (raw data)	54	6.11	
t-test	55	6.51	
Robust rank test	55	6.53	
Bootstrap difference in mean rank	55	6.85	
Wilcoxon test	55	7.31	
Cochran-Armitage trend test (4 groups)	50	7.36	
Ordinal logistic regression (4 groups)	50	7.50	
Ordinal logistic regression (3 groups)	51	7.92	
Cochran-Armitage trend test (3 groups)	51	8.27	
Chi Sq – dead or poor outcome vs. good	55	8.87	
Chi Sq – dead or poor outcome vs. excellent	54	9.24	
Median test	55	9.47	
Chi Sq – 2x3 test	51	9.96	
Chi Sq – dead vs. alive	51	9.98	
Chi Sq – 2x4 test	50	10.02	
Kolmogorov-Smirnov test	55	11.29	

TABLE 3.3

Comparison of statistical tests by type of intervention. The numbers in the table reflect the rank of that test for each intervention. Statistically significant results ($p<0.05$) are shown in bold, the green shading highlights the top three tests.

Intervention	Data sets	Ranking															
		OR	TT	RRT	BS	WIL	Tr 4	OR 4	OR 3	Tr 3	X^2 c	X^2 e	MED	$X^2$3	X^2 d	X^2 4	KS
Thrombolysis	**10**	**11**	**15**	**13**	**10**	**12**	**7**	**6**	**3**	**4**	**5**	**14**	**16**	**2**	**8**	**1**	**9**
Anticoagulation	**3**	**5**	**7**	**10**	**6**	**11**	**4**	**8**	**1**	**13**	**9**	**2**	**3**	**14**	**16**	**12**	**15**
Antihypertensive	4	10	13	11	5	12	4	3	6	2	14	8	16	9	1	7	15
Antiplatelet	**4**	**3**	**1**	**2**	**7**	**6**	**4**	**5**	**11**	**8**	**14**	**12**	**10**	**13**	**9**	**15**	**16**
Feeding	1	2	1	4	5	3	8	12	13	11	9	7	10	16	15	14	6
Neuroprotection	**17**	**1**	**3**	**2**	**4**	**5**	**8**	**7**	**10**	**12**	**11**	**9**	**6**	**14**	**13**	**15**	**16**
OT	**7**	**1**	**2**	**5**	**7**	**6**	**4**	**3**	**11**	**12**	**10**	**8**	**9**	**13**	**14**	**15**	**16**
Procoagulant	1	5	3	1	2	4	6	7	8	9	11	12	10	13	15	14	16
Stroke unit	8	2	1	7	6	8	12	14	5	3	4	16	13	11	10	15	9
Total	**55**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**

OR: ordinal logistic regression (raw data); TT: t-test; RRT: robust rank test; BS: bootstrap; WIL: Wilcoxon; Tr 4: trend test (4 levels); OR 4: ordinal logistic regression (4 levels); OR 3: ordinal logistic regression (3 levels); Tr 3: trend test (3 levels); X^2 c: Chi Sq – death or poor outcome vs. good; X^2 e: Chi Sq – death or poor outcome vs. excellent; MED: median test; X^2 3: Chi Sq – death vs. poor outcome vs. good; X^2 d: Chi Sq – death vs. alive; X^2 4: Chi Sq – death vs. poor outcome vs. good vs. excellent; KS: Kolmogorov-Smirnov test; bold = $p<0.05$.

TABLE 3.4

Comparison of the rankings of statistical tests by trial and patient characteristics. The top 3 ranked tests are given for each characteristic.

Rank	Scale			Size		Setting			Follow-up		Severity †		Age		Recruitment time		Outcome	
	BI	mRS	3Q	<500	>500	Acute	Rehab	SU	3 M	>3 M	Mild	Severe	<70	>70	<6 h	>6 h	-	+
1	OR	OR	TT	OR	BS	OR	OR	TT	TT	OR	OR	OR	RR	OR	TT	OR	OR	TT
2	RR	RR	Tr 4	TT	OR	RR	TT	Tr 3	OR	OR 4	TT	BS	OR	TT	OR	RR	RR	OR
3	BS	TT	BS	RR	RR	BS	RR	X^2 c	RR	BS	OR 4	RR	TT	BS	Tr 3	Tr 4	Tr 3	RR
p	**	**	**	***	*	***	**		***	*	**	**	***	**	*	***	*	***

BI: Barthel Index; BS: bootstrap; mRS: modified Rankin Scale; 3Q: three questions; OR: ordinal logistic regression (raw data); OR 3: ordinal logistic regression (3 levels); OR 4: ordinal logistic regression (4 levels); RRT: robust rank test; Tr 3: trend test (3 levels); Tr 4: trend test (4 levels); TT: t-test; X^2 c: Chi Sq – death or poor outcome vs. good. *p<0.05; **p<0.01; ***p<0.0001 reflect statistically significant differences between the tests within the sub group; †Severity assessed as death rate per month of follow up in control group (baseline severity was not used since it was not available for many trials)

TABLE 3.5

Testing the proportionality of odds assumption for ordinal logistic regression. Data given are the p values from the likelihood ratio test. Statistically significant values are shown in bold and signify that the assumption is not met.

Data set	Raw data	3 levels	4 levels
Acute:			
AbESTT	0.4229		
ASK	**0.0286**	**0.0003**	**0.0011**
ASSIST 07	**0.0024**	0.2606	0.3685
ASSIST 10	0.8285	0.2873	0.5455
ATLANTIS A	**0.0099**	**0.015**	**0.0355**
ATLANTIS B	0.2208	0.1585	0.1678
BEST pilot aten	0.0791	0.485	**0.0538**
BEST pilot prop	0.4244	0.5407	0.6823
BEST aten	0.2634	0.0797	0.1835
BEST prop	0.2752	0.2332	0.493
CAST	0.3424	0.1368	0.3424
Citicoline 01	0.5702	0.7586	0.7539
Citicoline 07	0.6205	0.6509	0.7963
Citicoline 10	**0.0042**	0.2107	0.1265
Citicoline 18	0.121	0.6952	0.1228
DCLHb	0.3371	0.413	0.2291
EAST	0.5075	0.962	0.7138
Ebselen		0.9356	0.9379
ECASS II	0.3011	0.1906	0.0942
Edaravone	0.4007	0.8521	0.9695
Factor VIIa	0.9105		
FISS high	0.6601	0.8576	0.6601
FISS low	0.2543	0.1599	0.2543
FISS-TRIS	0.7100		
FOOD 3	**0.0396**	0.8416	0.6793
INWEST high	0.2482	0.0741	**0.058**
INWEST low	0.5118	0.6202	0.849
IST	0.8975	0.639	0.8975
MAST-E	0.1043	0.0768	0.1398
MAST-I A	0.6082	0.4198	0.6551

MAST-I S	0.0746	0.3616	0.2077
MAST-I AS	**0.0031**	**0.0002**	**0.0007**
NINDS	0.1148	0.1685	**0.0212**
PROACT II	0.2495	0.1342	0.305
RANTTAS	0.2832	0.5802	0.5189
RANTTAS II	0.0554	0.8603	0.0947
STIPAS	0.5632	0.7217	0.9417
Streptokinase pilot	0.1347	0.6326	0.7726
TESS	0.9911	0.4618	0.7372
TESS II	0.7744	0.7763	0.7415
Rehabilitation:			
Corr	0.1894	0.7648	0.8463
Gilbertson	0.0708	0.2866	0.4689
Logan	0.2583	0.5619	0.504
Parker ADL	0.5315	0.4086	0.6649
Parker leisure	0.8493	0.9985	0.9689
Walker I	0.511		
Walker II	0.0631		
Young	0.5777		
Stroke unit:			
Dover	0.3844	0.2142	0.4318
Helsinki	0.2541	**0.047**	0.1387
Kuopio	0.1452	0.5874	0.6746
Nottingham	0.2086	0.6037	0.7157
Orpington team	**0.0182**	**0.0174**	**0.0228**
Orpington dom	**0.0181**	0.1937	0.2773
Newcastle	0.4631		

TABLE 3.6

Testing the assumptions of the t-test. The Shapiro-Wilk test assess the normality assumption, statistically significant values indicate the assumption is not met. The F test assesses the equality of variance between the treatment groups. Statistically significant values signify non equal variance across the groups. Statistically significant values are shown in bold.

Data set	Shapiro-Wilk test		F test	
	W	P value	F	P value
Acute:				
AbESTT	**0.90**	**<0.0001**	2.73	0.10
ASK	**0.80**	**<0.0001**	2.22	0.14
ASSIST 07	**0.83**	**<0.0001**	0.88	0.35
ASSIST 10	**0.83**	**<0.0001**	0.50	0.48
ATLANTIS A	**0.73**	**<0.0001**	0.87	0.35
ATLANTIS B	**0.90**	**<0.0001**	0.81	0.37
BEST pilot aten	**0.79**	**<0.0001**	0.00	0.95
BEST pilot prop	**0.76**	**<0.0001**	0.91	0.35
BEST aten	**0.80**	**<0.0001**	0.52	0.47
BEST prop	**0.81**	**<0.0001**	1.02	0.31
CAST*	**0.23**	**<0.01**	**5.34**	**0.02**
Citicoline 01	**0.79**	**<0.0001**	0.73	0.39
Citicoline 07	**0.79**	**<0.0001**	0.01	0.92
Citicoline 10	**0.93**	**0.0001**	0.24	0.63
Citicoline 18	**0.80**	**<0.0001**	0.07	0.80
DCLHb	**0.93**	**0.0003**	**11.35**	**0.001**
EAST	**0.91**	**<0.0001**	**8.24**	**0.004**
Ebselen	**0.78**	**<0.0001**	**5.16**	**0.02**
ECASS II	**0.91**	**<0.0001**	1.63	0.20
Edaravone	**0.88**	**<0.0001**	2.81	0.09
Factor VIIa	**0.90**	**<0.0001**	0.13	0.72
FISS high	**0.86**	**<0.0001**	3.71	0.06
FISS low	**0.84**	**<0.0001**	0.74	0.39
FISS-TRIS	**0.86**	**<0.0001**	3.35	0.07
FOOD 3	**0.74**	**<0.0001**	2.04	0.15
INWEST high	**0.83**	**<0.0001**	**13.41**	**0.0003**

INWEST low	**0.82**	**<0.0001**	2.99	0.09
IST*	**0.26**	**<0.01**	**4.90**	**0.03**
MAST-E	**0.83**	**<0.0001**	0.00	0.99
MAST-I A	**0.87**	**<0.0001**	3.34	0.07
MAST-I S	**0.85**	**<0.0001**	0.76	0.38
MAST-I AS	**0.82**	**<0.0001**	0.65	0.42
NINDS	**0.90**	**<0.0001**	**10.22**	**0.002**
PROACT II	**0.90**	**<0.0001**	0.98	0.32
RANTTAS	**0.74**	**<0.0001**	**4.14**	**0.04**
RANTTAS II	**0.79**	**<0.0001**	1.87	0.17
STIPAS	**0.68**	**<0.0001**	2.85	0.09
Streptokinase pilot	**0.82**	**0.002**	0.00	0.95
TESS	**0.81**	**<0.0001**	0.62	0.43
TESS II	**0.81**	**<0.001**	1.05	0.30
Rehabilitation:				
Corr	**0.91**	**<0.0001**	2.44	0.12
Gilbertson	**0.79**	**<0.0001**	0.13	0.72
Logan	**0.82**	**<0.0001**	1.14	0.29
Parker ADL	**0.93**	**<0.0001**	0.28	0.60
Parker leisure	**0.93**	**<0.0001**	0.96	0.33
Walker I	0.96	0.31	1.38	0.25
Walker II	**0.74**	**<0.0001**	**5.77**	**0.02**
Young	**0.90**	**<0.0001**	**7.32**	**0.008**
Stroke unit:				
Dover	**0.82**	**<0.0001**	**4.79**	**0.03**
Helsinki	**0.82**	**<0.0001**	1.18	0.28
Kuopio	**0.82**	**<0.0001**	0.08	0.78
Nottingham	**0.89**	**<0.0001**	**6.47**	**0.01**
Orpington team	**0.89**	**<0.0001**	8.56	0.004
Orpington dom	**0.90**	**<0.0001**	3.16	0.08
Newcastle	**0.90**	**<0.0001**	0.13	0.72

*used the Kolmogorov-Smirnov test as too many observations for the Shapiro-Wilk test.

TABLE 3.7

Comparison of the pooled and unpooled t-test. Analysis by non parametric two-way ANOVA and Duncan's multiple range test; tests joined by the same band are not significantly different from each other at $p<0.05$.

Test	Mean rank	Banding
Ordinal logistic regression (raw data)	6.65	
t-test (pooled)	**7.13**	
t-test (unpooled)	**7.20**	
Robust rank test	7.31	
Bootstrap difference in mean rank	7.39	
Wilcoxon test	7.81	
Cochran-Armitage trend test (4 groups)	7.86	
Ordinal logistic regression (4 groups)	8.04	
Cochran-Armitage trend test (3 groups)	8.66	
Ordinal logistic regression (3 groups)	8.67	
Chi Sq – dead or poor outcome vs. good	9.53	
Chi Sq – dead or poor outcome vs. excellent	10.14	
Median test	10.28	
Chi Sq – 2x3 test	10.28	
Chi Sq – 2x4 test	10.37	
Chi Sq – dead vs. alive	10.51	
Kolmogorov-Smirnov test	11.90	

TABLE 3.8

Assessment of the type 1 error rate for the top three statistical tests using data from three trials, each using a different functional outcome scale. The data given are the number and percentage of statistically significant results found from 1,000 simulations.

Scale	Trial	Ordinal logistic regression		t-test		Robust rank test	
		n significant	% significant	n significant	% significant	n significant	% significant
Barthel Index	RANTTAS	39	3.9	44	4.4	41	4.1
Modified Rankin Scale	NINDS	57	5.7	41	4.1	44	4.4
Three Questions	IST	56	5.6	46	4.6	52	5.2

TABLE 3.9

Availability of statistical tests compared in three statistical packages.

Test	SAS version 9	Stata version 8	SPSS version 15
Chi Square 2*2	PROC FREQ; TABLE TRT*RESP2CAT/CHISQ; RUN;	tabulate trt resp2cat, chi2	Analyze -> Descriptive statistics -> crosstabs Press statistics button for chi-square test
Chi Square r*n	PROC FREQ; TABLE TRT*RESPNCAT/CHISQ; RUN;	tabulate trt respncat, chi2	Analyze -> Descriptive statistics -> crosstabs Press statistics button for chi-square test
Cochrane-Armitage test	PROC FREQ; TABLE TRT*RESPNCAT/TREND; RUN;	ptrend n1 n2 respncat	N/A
Wilcoxon test	PROC NPAR1WAY WILCOXON; VAR RESP; CLASS TRT; RUN;	ranksum resp, by(trt)	Analyze -> Non-parametric -> Two-independent-samples tests Select Mann Whitney U test

Median test	PROC NPAR1WAY MEDIAN; VAR RESP; CLASS TRT; RUN;	median resp, by(trt)	Analyze -> Non-parametric -> Tests for several independent-samples tests Select median test
t-test	PROC TTEST; VAR RESP; CLASS TRT; RUN;	ttest resp, by(trt)	Analyze -> Compare means -> Independent-samples t test
Kolmogorov-Smirnov test	PROC NPAR1WAY KS; VAR RESP; CLASS TRT; RUN;	ksmirnov resp, by(trt)	Analyze -> Non-parametric -> Two-independent-samples tests Select Kolmogorov-Smirnov test
Robust rank test	User written macro available at http://www.sociology.ohio-state.edu/people/ptv/macros/fligner policello.htm	fprank command available in version 9.2	N/A
Ordinal Regression	PROC LOGISTIC; MODEL RESP=TRT; RUN;	xi: ologit resp i.trt	Analyze -> Regression -> Ordinal

Bootstrapping	Macro can be downloaded from http://support.sas.com/ctx/samples/index.jsp?sid=479&tab=downloads to perform bootstrap analyses	Use commands such as bsample and simulate within a user written program	N/A

Variables used in examples above: RESP: Raw scores from outcome scale; RESP2CAT: Dichotomised version of Resp; RESPNCAT: *Categorised version of Resp; TRT: 1=Active treatment, 0=Control;* N1: Number of patients with that outcome in active treatment group; N2: Number of patients with that outcome in control group; N/A: not available.

CHAPTER 4

RESULTS

SAMPLE SIZE FOR BINARY AND ORDERED DATA

PUBLICATIONS/PRESENTATIONS CONTRIBUTING TO THIS CHAPTER

The Optimising the Analysis of Stroke Trials (OAST) Collaboration, **Gray L.J**, Bath P.M.W, Collier T. (2008) Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches. *International Journal of Stroke. 3:78-84.*

Gray L.J, Bath P.M.W, Collier T, on behalf of the OAST Collaboration. (2008) Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches. *Poster presentation at European Stroke Conference, France, May 2008. Cerebrovascular Diseases 25(suppl 2):1-192.*

4.1 INTRODUCTION

The previous chapter showed that statistical tests that use the original ordered categories describing death or dependency are statistically more efficient than those which dichotomise the data (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2007); suitable approaches include ordinal logistic regression, the t-test, the robust rank test, bootstrapping the difference in mean rank, and the Wilcoxon test.

If the analysis of stroke trials should be changed from using dichotomous to polytomous functional outcome data, then it is critical to consider how sample size should be calculated. Sample size estimation is an important part of trial design and is now a compulsory element when applying for funding and publishing completed trials (The CONSORT Statement, 1996, Gardner and Altman, 1989). Key components in any sample size calculation include the intended power ($1-\beta$) and significance (α), and expected treatment effect (Weaver et al., 2004).

This part of the project compares sample size estimations obtained using different methods based on dichotomous, ordinal and continuous outcomes.

4.2 METHODS

4.2.1 Trial data

As with the previous chapter, all 55 data sets in the OAST project were included in this analysis as only data on functional outcome and treatment assignment were required. See Chapter 2 for a description of the OAST data set.

4.2.2 Sample size estimation

Four methods of sample size estimation were chosen for comparison; one is based on the proportion of events and is currently used in many acute stroke trials (Weaver et al., 2004). The other three estimate sample size for ordinal or continuous outcomes (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2007). As with the previous chapter, the sample size methods compared are those which are available in standard statistical packages. All the methods of sample size estimation assume that the treatment groups are of equal size. In all cases z_α and z_β are the appropriate values from the standard normal distribution based on the significance level (α) and power ($1-\beta$) chosen by the investigator (see Table 4.1). None of the methods take into account drop out or non compliance and it is customary to inflate any given sample size by around 10% to take into account these factors. The methods of sample size estimation used are described in the next section.

Comparison of proportions

The formula for estimating the sample size when the outcome is binary is:

$$n = \frac{(z_\alpha + z_\beta)^2 (p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2}$$

where n is the number of patients required in each group, p_1 and p_2 are the proportions of interest in the two treatment groups (Weaver et al., 2004). This method was carried out using Stata.

Parametric comparison

If a trial has an outcome which is continuous then the investigator may choose a comparison of means as the method of analysis for the primary outcome, e.g. using the t-test. The appropriate sample size calculation is based on:

$$n = \frac{2\sigma^2 (z_\alpha + z_\beta)^2}{(\mu_2 - \mu_1)^2}$$

where μ_1 and μ_2 are the expected means in the two treatment groups and σ is the overall expected standard deviation (Bland, 2000). This method was carried out using Stata.

Non parametric comparison

This method of sample size estimation for comparing ordinal data was proposed by Payne (Payne, 1993) as part of the Genstat (GenStat, 2005) statistical program and is relevant when the Wilcoxon test or the robust rank test (Fligner and Policello, 1981) will be used to analyse the primary outcome once the trial is completed. The method calculates an approximate sample size needed based on the probability of response (i.e. the probability that an observation in one sample will be greater than the equivalent observation in the other sample) that should be detectable by initially assuming a normal approximation. This is

then refined by calculating powers for a range of replications centred around that approximation (Payne, 1993).

This was the only method available in the statistical packages assessed that carried out a non parametric sample size calculation. The method is specific to the GenStat program and no algorithm has been published. There are other published non parametric methods in the statistical literature, such as that according to Noether (Noether, 1987), but these are not available in standard statistical software.

Comparison of ordinal data

Sample size estimation for comparing two groups of ordinal data using the technique of ordinal regression was proposed by Whitehead (Whitehead, 1993). An estimate of the expected odds ratio and proportion of patients expected to fall into each category on the scale for the control group is required.

The sample size per group is given by:

$$n = \frac{6\left[(z_\alpha + z_\beta)^2 / (LogOR)^2\right]}{\left[1 - \sum_{i=1}^{k} \bar{\pi}^3\right]}$$

where OR is the odds of being in category i or less for one treatment group compared to the other, k is the number of categories on the scale of interest, and $\bar{\pi}$ is the mean proportion of patients expected in category i. This method was carried out using GenStat.

4.2.3 Comparison of methods

Each method of sample size estimation was carried out on each data set. The parameters needed within the calculation of each sample size were derived from each data set; these were then used to calculate sample size as if these treatment effects were desired. The comparison of proportions method was carried out twice using two different definitions of a functional outcome:

(i) 'Good': death or poor outcome (BI <60, mRS 3-6, 3Q 1/2) versus good outcome (BI 60-100, mRS 0-2, 3Q 3/4)

(ii) 'Excellent': death or poor outcome (BI <95, mRS 2-6, 3Q 1-3) versus excellent outcome (BI 95/100, mRS 0/1, 3Q 4)

See Chapter 3 for definitions of outcomes for the other scales used. The use of two definitions reflects that most trials used, historically, the poor/good outcome, whilst there has been a tendency recently, to rely on the poor/excellent outcome, largely based on the results of the NINDS tPA trial (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995).

In all cases significance was set at 5% with a power of 90%. The use of a fixed power of 90% will have ensured that the risk of a false negative was held constant. These sample sizes were then ordered within each trial and given a rank, with the lowest rank given to the method which produced the smallest sample size. A two-way analysis of variance test was then used to see on average which method had produced the lowest ranks and therefore the lowest sample sizes. The methods were than ordered in terms of the average sample sizes given using Duncan's multiple range test (Duncan, 1955).

Each method of sample size calculation was then compared to the proportion method for a 'good' outcome (as this is the most common method used in stroke trials). The median multiplier by type of intervention was then calculated, i.e. a value <1 shows that the method produces a smaller sample size than the proportion method and >1 shows that a larger sample size will result.

Analyses were carried out in SAS (version 8.2), Stata (version 7) and GenStat (version 8.1, for the methods of Payne and Whitehead) and significance was taken at $p<0.05$.

4.3 RESULTS

4.3.1 Trial characteristics

The characteristics of the 55 data sets used are presented in Chapter 2.

4.3.2 Comparison of sample size methods

The sample size methods differed significantly in estimating sample sizes for each trial ($p<0.0001$). The ordering of the methods showed that the ordinal method of Whitehead and comparison of means method produced significantly lower sample sizes than the other approaches, with the comparison of medians method of Payne giving the largest sample sizes (Table 4.2).

Table 4.3 shows the change in sample size in relation to the current standard method based on comparison of proportions for a good outcome (mRS ≤ 2 or BI ≥ 60). The ordinal method of Whitehead and comparison of means appear to reduce sample size by 28% and 30% respectively, relative to comparison of proportions (Table 4.3). In contrast, the method of Payne produces 12% larger sample sizes. Whilst this finding appears to be true for most interventions, it

may not be correct for trials of thrombolytics where ordinal (Whitehead, Payne) and continuous (comparison of means) approaches produce larger sample sizes. Interestingly, comparison of proportions based on an 'excellent' outcome also led to an increase in sample size as compared with comparisons based on a 'good' outcome.

The following figures give examples of the sample size required with varying levels of statistical power for each method, for three trials. Overall these plots show that the 'best' method of sample size calculation may vary slightly by trial but on average the Whitehead and comparison of means method produce the smallest sample size.

FIGURE 4.1

Sample size comparisons at varying levels of power (β) for the IST mega trial of aspirin versus control (International Stroke Trial Collaborative Group, 1997).

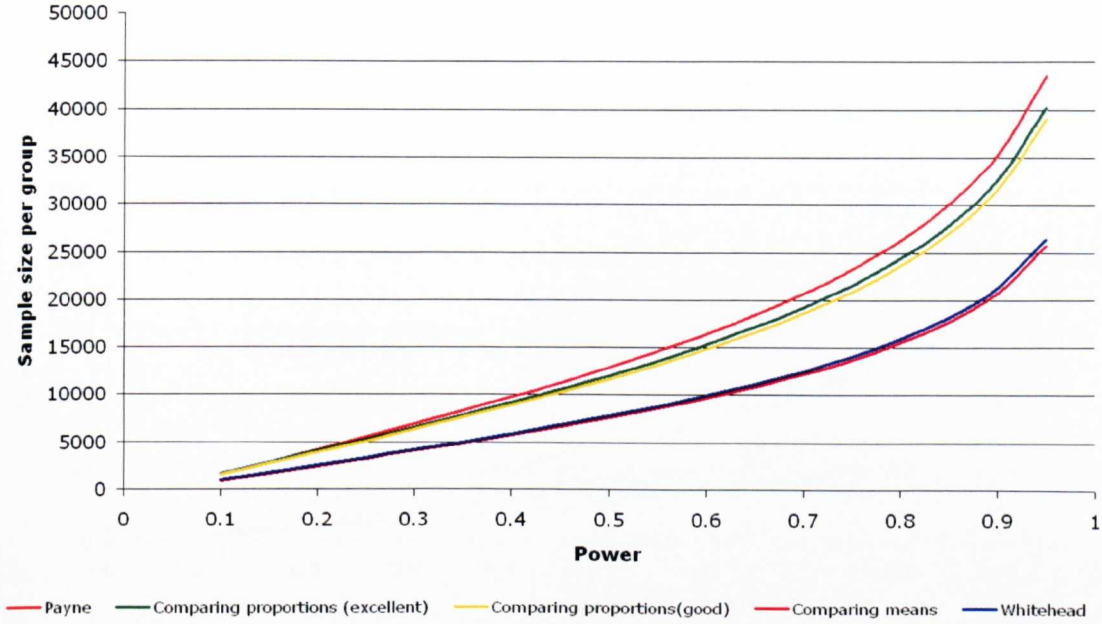
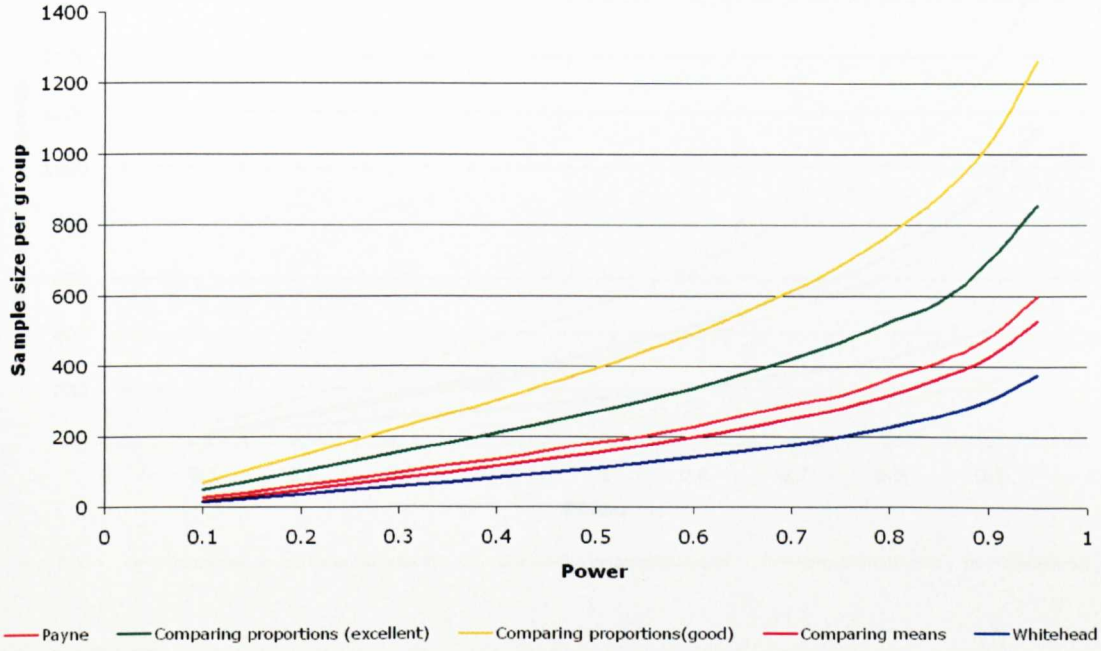


Figure 4.1 shows that across all levels of statistical power the comparison of means and Whitehead method out perform the other methods, consistently producing lower sample sizes. The method of Payne produces the highest sample sizes, with little difference between the two comparisons of proportions.

FIGURE 4.2

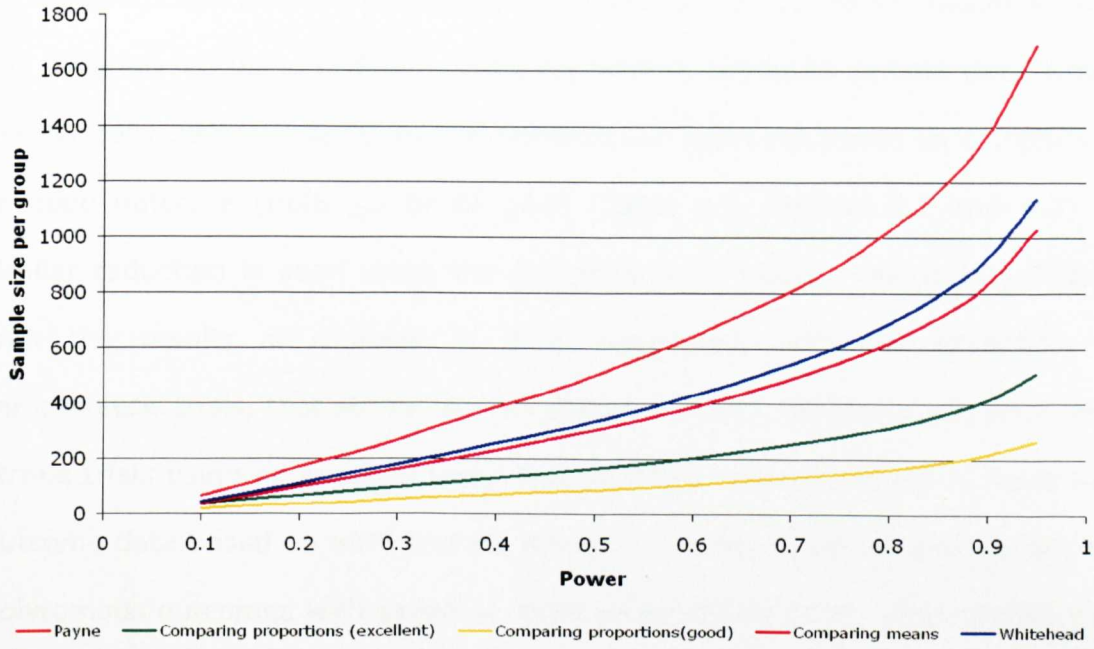
Sample size comparisons at varying levels of power (β) for a trial of edaravone (The Edaravone Acute Brain Infarction Study Group (Chair: Eiichi Otomo MD), 2003).



In the edaravone trial (Figure 4.2) the Whitehead method gives the smallest sample sizes across all levels of power. Here the dichotomous outcomes gave the largest sample sizes.

FIGURE 4.3

Sample size comparisons at varying levels of power (β) for the PROACT II trial of intra-arterial prourokinase (Furlan et al., 1999).



The PROACT II trial tested a thrombolytic agent (prourokinase) versus control. In contrast to the previous two examples, here the two dichotomous comparisons of proportion sample sizes are much smaller than both the comparison of means and Whitehead method. As discussed previously (Chapter 3), this is likely to reflect that the assumption of proportionality of odds is not being met in trials of thrombolytic therapies. The likelihood ratio test, which here tests the proportional odds assumption, for the PROACT II trial was not statistically significant (p=0.24), but this test is known to have low power and therefore may not indicate all cases where this assumption fails.

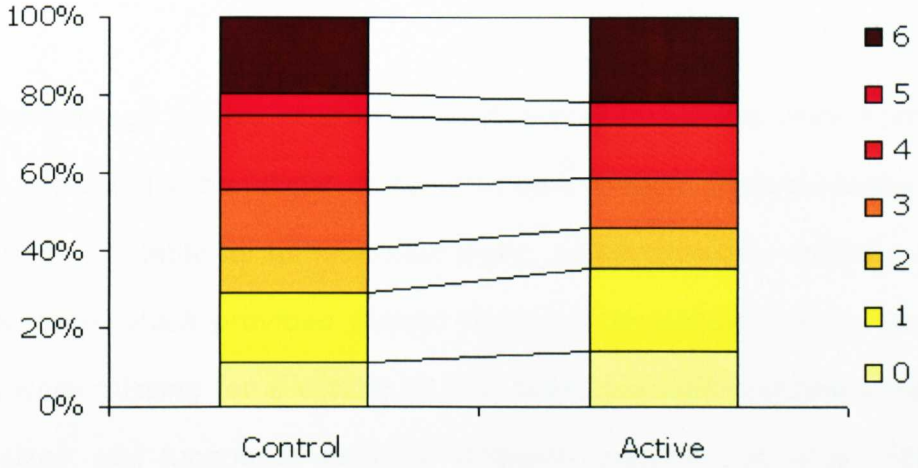
4.4 DISCUSSION

The results support the contention that trials designed to use an ordinal analysis of functional outcome will, on average, be smaller than those using a dichotomous outcome. In particular, Whitehead's method, which assumes trials will be analysed using ordinal logistic regression, produces sample sizes which are typically 28% smaller than the dichotomous approach based on comparison of good outcome (mRS $\leq$2 or BI $\geq$60) (Table 4.3, Figures 4.1 and 4.2). A similar reduction is seen using the comparison of means. Taking this finding with the results of Chapter 3, it is suggested, with the exception of thrombolysis trials, that stroke trialists should consider designing and analysing stroke trials using approaches which maintain the ordered nature of functional outcome data based on mRS and BI. Analysis of means may be appropriate for polytomous outcomes with seven or more levels (Song et al., 2006, Walters et al., 2001), as occurs with the BI.

As discussed previously, ordinal logistic regression assumes the intervention will exert effects of similar magnitude and direction at each transition of the outcome scale; this is unlikely to be the case for treatments such as thrombolytic agents, which both reduce dependency and, to a smaller extent, increase death (Figure 4.4). This is evident in Table 4.3 and Figure 4.3 where the ordinal (Whitehead, Payne) and continuous methods did not deliver smaller thrombolysis trials, e.g. PROACT II (Furlan et al., 1999). In contrast, most other interventions are likely to move patients up (efficacy) or down (hazard) by a part (or whole) of a mRS level (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2007), therefore fulfilling the key assumption underlying proportionality of odds. Table 4.3 shows that the ordinal method of Whitehead leads to smaller sample sizes for a wide range of interventions including antiplatelets, neuroprotectants, occupational therapy, and stroke units.

FIGURE 4.4

Distribution of the modified Rankin Scale for the six combined data sets of thrombolytic therapy (ECASS II, MAST E, MAST I (streptokinase vs. control and streptokinase and aspirin vs. control), NINDS, PROACT II and ATLANTIS B).



Applying methods of analysis which enable investigators to reduce the sample size needed for a trial may increase the feasibility of completing stroke clinical trials. A meta analysis of recruitment into stroke trials showed that over the last 15 years the number of recruiting centres within each completed trial has increased significantly over time with a non significant decrease in recruitment efficacy (subjects enrolled per study centre per month of recruitment) (Elkins et al., 2006). Another review of sample size in stroke trials found that the sample sizes of trials is increasing over time (Weaver et al., 2004). These two studies show that recruitment into stroke trials is becoming more complex and expensive, with more trial centres being needed to meet the sample size requirement. Therefore any solution which reduces the number of patients needed will lower the cost and complexity of trials and increase the potential to recruit the sample size needed (Elkins et al., 2006, Weaver et al., 2004).

The advantage of our study is that the different methods for estimating sample size have been tested on data from a large number of real stroke trials. As a

result, the findings are likely to exhibit external validity. It is evident that stroke trials are inherently heterogeneous in their design and results in that interventions, patients and results differ. Modelling approaches which synthesise data or use data from a single study cannot adequately take account of this heterogeneity.

A disadvantage of this study is that it aimed to include data from all stroke trials assessing a beneficial or harmful intervention. Unfortunately, data were not made available for all identified trials; where possible, individual data from publications which provided patient numbers by outcome score were created. Data were missing for a variety of trial types (acute/rehabilitation/stroke unit) and sizes, and functional outcome measure (mRS/BI), so it is unlikely that a systematic bias was introduced into the findings; however, the precision of the results may have been attenuated by the missing trials.

Another possible criticism of these results is the use of the actual trial parameters in the estimation of the sample sizes. Most of the trials (30/47, 64%) included in this project individually showed no treatment effect and were therefore included as part of a meta analysis showing a statistically significant effect. Therefore the parameters used in the calculations were, in the most part, determined for very small treatment differences. When repeating the analysis on only those data sets where a beneficial treatment difference was seen (16 data sets from 14/47 trials) and hence more 'realistic' parameters were used in the calculations, ordinal methods still ranked highest (see Table 4.4). Using very small treatment effects may add to the validity of these results as many have argued that sample sizes for stroke trials have been based on unrealistically large clinically meaningful differences between treatments (Furlan, 2002, Samsa and Matchar, 2001, Weaver et al., 2004), and small

effects may still be worthwhile if the treatment can be used across a wide range of patients.

4.5 SUMMARY

In summary, it is suggested that trialists designing future stroke studies of treatments which are likely to act uniformly across populations should consider analysing functional outcome using an ordinal method that retains the natural ordering of the outcome data. In doing so, they will be able to maintain study power for a smaller sample size which will reduce the complexity (less centres), length and cost of trials (Elkins et al., 2006). However, trials of thrombolysis (or other interventions where a likely asymmetrical hazard will be present alongside a symmetrical efficacy) should use current approaches which combine outcomes. In this respect, the decision to use excellent (mRS 0, 1/2-6), good (mRS 0-2/3-6) or moderate (mRS 0-3/4-6) splits in functional outcome will depend on the expected severity of patients.

In contrast, many argue that stroke trials have been underpowered (Weaver et al., 2004, Furlan, 2002). Therefore, investigators may choose to determine sample size using a binary cut but increase the statistical power to find a treatment difference by using an ordinal method of analysis. Using this approach would also give investigators increased power to assess treatment effects within certain groups of patients sub group analysis. By carrying out sub group analyses, investigators are able to assess for whom the treatment works best, which may be useful if assessing very expensive novel treatments (Warlow, 2002). Nevertheless, it is apparent that there is no perfect method for calculating sample size for stroke trials and other factors related to trial design and patient type should be considered. Software is available to calculate sample size using the approaches tested here (Whitehead, 1993, GenStat, 2005).

TABLE 4.1

Lookup table for values of $(z_\alpha + z_\beta)$ for various level of α and β (Bland, 2000).

β	Significance level, α	
	0.05	0.01
0.70	6.2	9.6
0.80	7.9	11.7
0.90	10.5	14.9
0.95	13.0	17.8
0.99	18.4	24.0

TABLE 4.2

Comparison of sample sizes produced by five methods. Lower ranks imply the method produces lower sample sizes. Analysis by two-way ANOVA and Duncan's multiple range test; tests joined by the same band are not significantly different from each other at $p<0.05$.

Method	Mean rank	n	Banding
Comparing ordinal data (Whitehead, 1993)	2.15	53	
Comparing means	2.28	55	
Comparing proportions (good outcome)	3.18	55	
Comparing proportions (excellent outcome)	3.37	54	
Comparing medians (Payne, 1993)	3.92	54	

TABLE 4.3

Comparison of sample sizes using four methods of calculation relative to the proportion method for a good outcome (modified Rankin Scale ≤ 2 or Barthel Index ≥ 60) with results subcategorised by type of intervention. Data are median (inter-quartile range) multiplier.

Intervention	Trials n	Ordinal	Means	Proportion (excellent)	Medians
Thrombolysis	10	1.22 (0.73, 2.15)	1.36 (0.52, 38.84)	1.92 (0.43, 6.70)	2.06 (1.22, 3.35)
Anticoagulation	3	0.97 (0.59, 1.08)	1.03 (0.55, 1.03)	0.57 (0.16, 1.47)	1.64 (1.01, 1.78)
Antihypertensive	4	0.42 (0.34, 1.29)	0.88 (0.35, 2.43)	0.83 (0.01, 8.72)	0.27 (0.53, 1.98)
Antiplatelet	4	0.51 (0.28, 0.67)	0.48 (0.38, 0.66)	0.83 (0.35, 1.03)	0.82 (0.44, 1.11)
Feeding	1	0.07 (- , -)	0.04 (- , -)	0.11 (- , -)	0.14 (- , -)
Neuroprotection	17	0.71 (0.22, 1.09)	0.70 (0.42, 0.92)	0.92 (0.23, 2.34)	1.08 (0.41, 1.43)
Occupational therapy	7	0.44 (0.04, 2.20)	0.37 (0.03, 3.46)	0.30 (0.07, 20.77)	0.73 (0.06, 3.38)
Procoagulant	1	0.79 (- , -)	0.68 (- , -)	1.06 (- , -)	1.17 (- , -)
Stroke unit	8	0.88 (0.35, 24.32)	0.96 (0.22, 4.21)	4.36 (1.75, 31.82)	1.36 (0.56, 5.53)
Total	55	0.72 (0.47, 0.86)	0.70 (0.55, 0.94)	0.99 (0.71, 1.79)	1.12 (0.80, 1.40)

TABLE 4.4

Percentage reduction (-)/ increase (+) in sample size in comparison to the Whitehead ordinal data method for a sub-group of the OAST trials where a beneficial treatment effect was shown in the original trial publication. Highlighted cells indicate a greater sample size required in comparison to the ordinal method of Whitehead.

	Sample size method			
	Means	Proportion (good)	Proportion (excellent)	Medians
CAST	-11	+39	+34	+40
Citicoline 1	+4	+73	-21	+35
Edaravone	+29	+71	+57	+37
Factor VII	-14	+21	+25	+32
FISS High	-5	-5	+26	+39
FISS Low	-43	+3	-83	+41
NINDS	+7	+14	-34	+36
PROACT II	-10	-77	-55	+33
Walker I	-14	+56	+70	+40
Walker II	+44	+98	+60	+44
Bradford	+4	+61	+56	+37
Helsinki	+48	-45	+17	+39
Kuopio	-97	-99	-95	-85
Nottingham	-3	+25	+83	+35
Orpington Team	-42	+53	+95	+36
Orpington Domiciliary	-53	+73	+99	+37

CHAPTER 5

RESULTS

ADJUSTMENT FOR PROGNOSTIC FACTORS

PUBLICATIONS/PRESENTATIONS CONTRIBUTING TO THIS CHAPTER

The Optimising the Analysis of Stroke Trials (OAST) Collaboration, **Gray L.J**, Bath P.M.W, Collier, T (2008) Should stroke trials adjust functional outcome for baseline prognostic factors? *In Press Stroke*

Gray L.J, Collier T, Bath P.M.W (2008) Should stroke trials adjust their primary outcome for prognostic factors? *Poster presentation at the International Stroke Conference, New Orleans, February 2008. Stroke. 2008;39:527-729.*

Gray L.J, Bath P.M.W, Collier T (2006) Optimising the statistical analysis of functional outcome in stroke clinical trials – should trials adjust their primary outcome for age, sex and severity? *Poster presentation at Joint World Congress on Stroke, Cape Town, October 2006, International Journal of Stroke. 1 (suppl 1): 111-174.*

5.1 INTRODUCTION

Results from the 'Optimising Analysis of Stroke Trials' (OAST) Collaboration have shown that the univariate analysis of stroke trials can be improved by using the inherent ordering of functional outcome rather than collapsing data into two or more groups (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2007). Specifically, use of ordinal logistic regression, the robust rank test, the t-test, bootstrapping the difference in mean rank and the Wilcoxon test, were more powerful methods than those based on collapsed data. This efficiency can be translated into increased statistical power for a given sample size, or a reduced sample size for a given power (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2008). The next stage of the OAST project will look at the effect of adjusting for prognostic factors.

When considering an adjusted analysis, the choice of covariates is of prime importance. Three main methods have been proposed for selecting covariates (Raab et al., 2000):

1. Variables which are known to be imbalanced across the treatment groups, although this requires a *post hoc* decision
2. Prognostic factors which are related to the primary outcome
3. A combination of adjusting for those variables which are both related to outcome and imbalanced across treatment groups

Senn suggested that the latter approach may be the most sensible as the reliability of unadjusted tests is affected by both the correlation between the outcome and covariate, and the level of imbalance (Senn, 1989). However, accounting for imbalances requires a post hoc decision and therefore is not practical in clinical trials where models have to be specified in the statistical analysis plan prior to database closure, lock and analysis.

The process of randomisation, whilst reducing bias, does not guarantee the matching of baseline variables between treatment groups. Imbalances at baseline between prognostic factors have complicated the interpretation of several acute stroke trials (International Stroke Trial Collaborative Group, 1997, De Deyn et al., 1997, Mayer et al., 2007). Further, imbalances reduce statistical power and it is likely that analysis methods which take account of pre-randomisation factors will be more efficient than those which do not make such adjustment. Finally, adjustment reduces the variability in the data so that more precise comparisons of treatment can be made (Pocock et al., 2002).

In 2000 a review of randomised clinical trials published in high quality journals (British Medical Journal, Journal of the American Medical Association, The Lancet and the New England Journal of Medicine) was carried out, looking specifically at the use of baseline data (Assmann et al., 2000). The review found that in terms of adjustment for covariates, most trials did carry out an adjusted analysis of the primary outcome (72%), but that the majority of studies placed emphasis on the unadjusted analysis (76%). Most trials took into account between five and nine covariates, with the choice being based on prognostic significance or imbalance in the bulk of cases.

Several studies have examined adjustment for prognostic variables when using functional outcome scales. Re-analysis of data from the NINDS trial of alteplase (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995) using a logistic regression model adjusted for an estimate of prior risk found a 13% reduction in sample size (Johnston et al., 2004). A study using data from brain injury trials measuring outcome on the Glasgow Outcome Scale found that covariate adjustment lead to a 25% reduction in sample size when using logistic regression (Hernandez et al., 2006). Other studies have

found similar reductions in sample size with time to event analyses (Hernandez et al., 2006, Hauck et al., 1998). However, no studies to date have looked at the effect of adjustment on ordinal logistic regression.

Furthermore, none of these studies discussed the inherent differences between adjusted and unadjusted models. Adjusted models are conditional on the covariates included in the model and therefore interpretation of the results is at the patient level whereas unadjusted models (which do not account for covariates) have a population level interpretation.

The aim of the analysis presented in this chapter was to assess whether stroke trials using ordinal logistic regression should routinely adjust for important prognostic factors in their primary analyses. The reduction in the sample size needed for a specific power will be used to assess the effect of covariate adjustment.

5.2 METHODS

5.2.1 Trial data

Trials were included from the OAST individual patient database where covariate (age, sex and severity) data had been provided. Three extra trials have been added to the OAST database since the initiation of the project. Tables 5.1 and 5.2 show the baseline characteristics and primary outcome data for these trials. Trials of thrombolytic agents were excluded, since the previous two chapters showed that their analysis does not benefit from ordinal methods (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2008).

5.2.2 Outcome and covariate data

Data on demographics (age, sex), stroke severity (National Institutes of Health Stroke Scale [NIHSS], Orgogozo Stroke Scale, Unified Neurological Stroke Scale, or other similar measures), treatment group and functional outcome variables were collected for each trial.

5.2.3 Statistical methods

All analyses were carried out in Stata (version 8). Statistical significance relates to $p<0.05$.

Relationship of covariates with functional outcome

Ordinal logistic regression was used to assess the relationship between each covariate and outcome within each trial.

Baseline imbalances in covariates

Although statistical testing for baseline imbalances should be discouraged, this was carried out in this study so that the effect of imbalance on adjustment could be assessed. Baseline imbalances between each covariate and treatment

were assessed using t-tests for age and severity, and the chi-square test for sex.

Models

Two models were compared:

(i) unadjusted model, which contained treatment assignment only (as a binary variable)

(ii) adjusted model, which contained treatment and sex (as binary variables), and age and baseline severity (continuous variables)

The adjusted model was restricted to these data as age, sex and severity were the only prognostic variables available for all trials. Additionally, these three consist of the key demographic and clinical variables.

Simulations

Although some included trials were individually significant on their assessment of functional outcome, others were neutral but had been included because they tested effective or hazardous treatments (as determined in published meta analyses). Therefore significant treatment benefits with three levels of effect (coefficients of -0.05, -0.30 or -0.56 equivalent to unadjusted odds ratios of 0.95, 0.74, or 0.57 (Hernandez et al., 2006), respectively) were simulated. By reference, trials of hemicraniectomy (Juttler et al., 2007), thrombolysis (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995), stroke units (Stevens and Ambler, 1982), and aspirin (International Stroke Trial Collaborative Group, 1997) achieved odds ratios of 0.24, 0.63, 0.60, and 0.94 respectively. For consistency across studies, BI and 3Q scales were reversed so that higher scores related to a worse state of outcome, as with the mRS; hence, an OR less than one reflects a beneficial treatment effect across all trials and scales. Simulations were based on the

method proposed by Hernandez *et al* for logistic regression (Hernandez et al., 2006), but extended for outcomes of an ordinal nature by using ordinal logistic regression.

The probability of having an unfavourable outcome was estimated using ordinal logistic regression (containing age, sex and baseline severity). Patients were randomly assigned to each treatment group (with active and control groups of the same size as the original trial); an artificial treatment effect was then added to the active group. A new outcome variable was generated by comparing the probability of an unfavourable outcome (based on the probability from the prognostic model and the added treatment effect) to a random variable with values between zero and one, this comparison adds noise into the new outcome variable produced. Unadjusted and adjusted ordinal logistic regression models were then applied to the new outcome and the Z-score for the estimate of treatment effect for each model was saved. This procedure was then carried out 10,000 times for each of the 23 trials and repeated for each level of treatment effect.

Reduction in sample size

The reduction in sample size was used to assess the increase in power gained from adjustment. The Z scores from the unadjusted and adjusted models were compared and the reduction in sample size calculated using (Hernandez et al., 2004):

$$\text{Reduction} = 100 - 100 \times \left[\frac{\text{Mean Z score unadjusted}}{\text{Mean Z score adjusted}} \right]^2$$

5.3 RESULTS

5.3.1 Trial data

The present data set compared individual patient data from 23 trials (20 from the original OAST data set and three new trials (Blanco et al., 2007, Juttler et al., 2007, Lampl et al., 2007)) including 25,674 patients. The characteristics of the trials included are given in Table 5.3. Thirteen trials measured outcome using the BI, nine used the mRS, and one used the 3Q scale. Fourteen trials measured baseline severity using the NIHSS with others using another measure such as the Orgogozo Stroke Scale. Trial sizes ranged from 32 to 19,435 patients (median 259) (Table 5.3).

5.3.2 Relationship of covariates with functional outcome

Table 5.4 shows the relationship between age, sex and severity with functional outcome. A highly statistically significant ($p<0.0001$) relationship between severity and functional outcome was found for the majority of trials (22/23), with greater baseline severity leading to worse functional outcome. Twenty two trials showed a significant relationship between age and outcome, and six showed a significant relationship with sex. Figures 5.1-5.3 show these relationships graphically in those trials which measured outcome using the mRS.

FIGURE 5.1

Relationship between age (n=9), and outcome (modified Rankin Scale), the data shown are means and standard deviations.

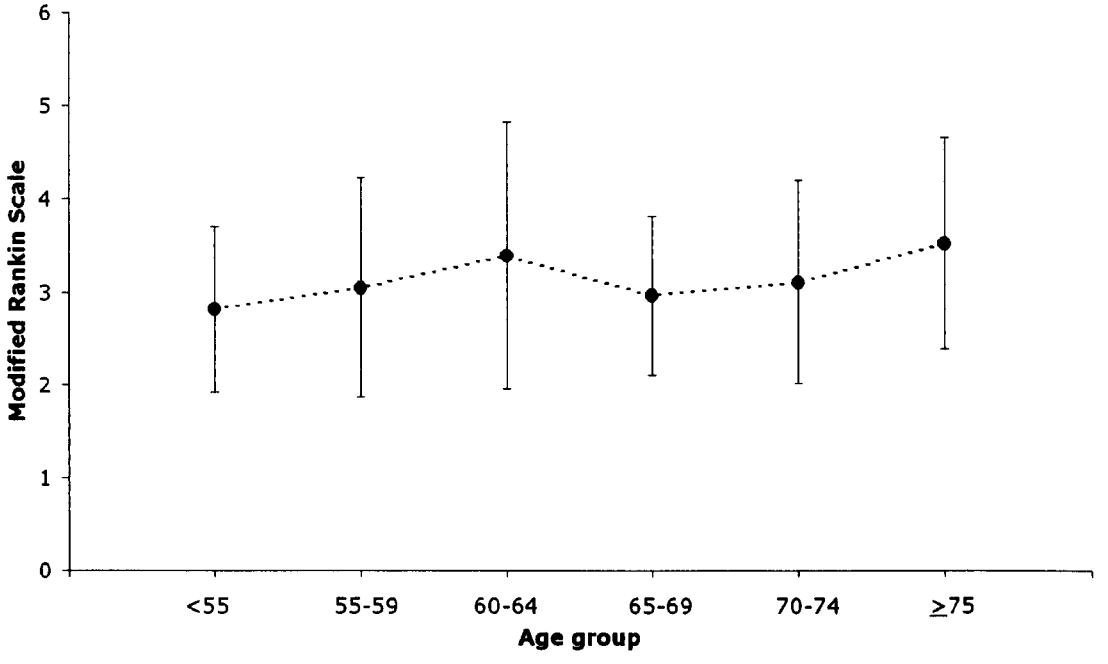


FIGURE 5.2

Relationship between severity (n=6) and outcome (modified Rankin Scale), the data shown are means and standard deviations.

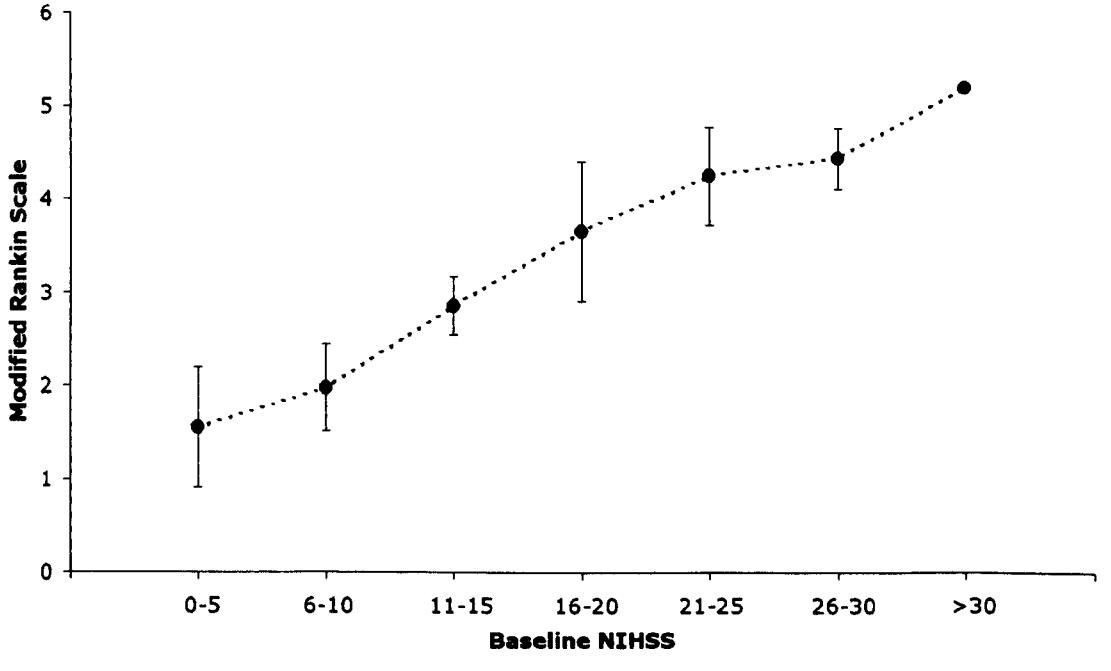
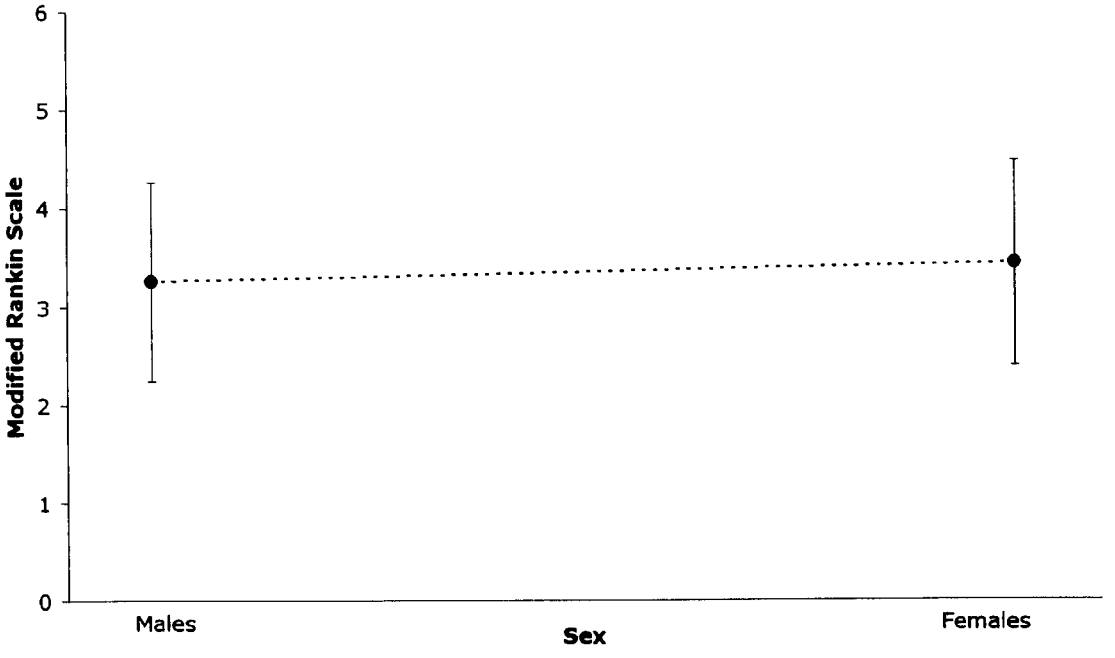


FIGURE 5.3

Relationship between sex (n=9) and outcome (modified Rankin Scale), the data shown are means and standard deviations.



5.3.3 Baseline imbalances in covariates

Statistically significant differences in baseline covariates were only seen in three of the included trial data sets, one for age (in the ASSIST 07 trial the treatment groups differed by 3.6 years, a difference which has borderline biological significance) and two for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance in the Dover trial, but a difference of 2.82 on the NIHSS in the DESTINY trial is clinically relevant) (Table 5.5).

5.3.4 Reduction in sample size

Table 5.6 shows the median reduction in sample size for the three levels of treatment effect. Trial sample size was reduced by 35-38% when covariates were introduced and was independent of the magnitude of treatment effect. A conservative figure for this reduction could be set at the lower end of the interquartile range, i.e. 20-30%. The adjusted coefficients and odds ratios are closer to and more tightly packed around the actual simulated treatment effect than for the unadjusted models (Table 5.6 and Figure 5.4). Table 5.7 shows that as the treatment effect increased, the proportion of simulations where odds ratios and treatment coefficients were larger in the adjusted models compared to the unadjusted also increased.

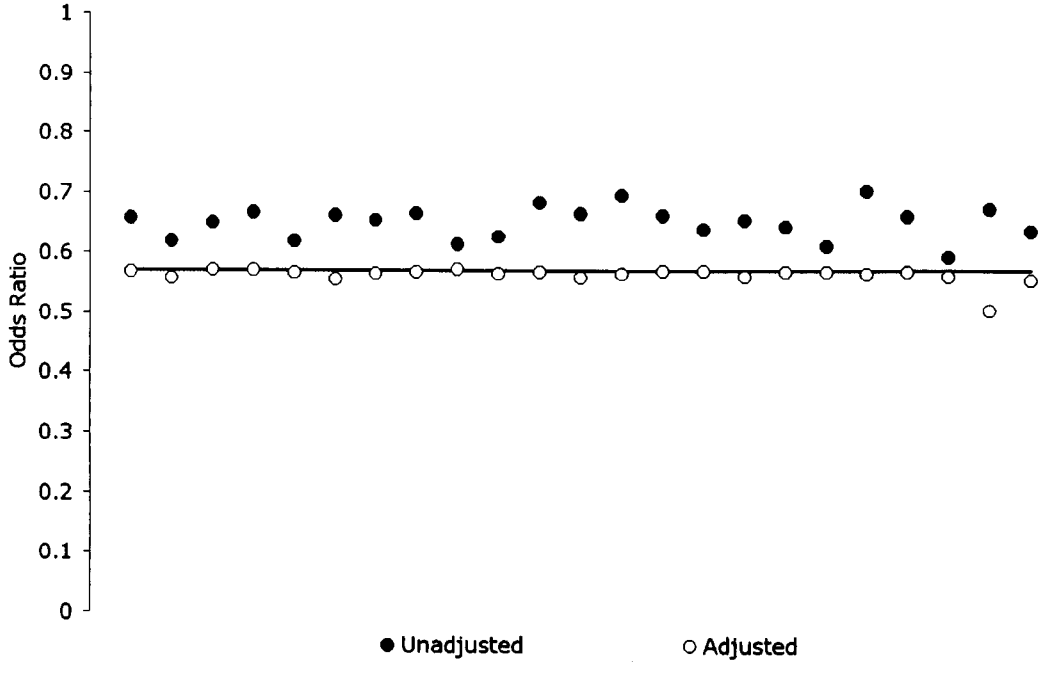
5.3.5 Sub group analysis

The results from the sub group analyses are shown in Table 5.8. The biggest reduction in sample size was seen for trials using the BI (40%) as compared to 21-29% for mRS and 20% for 3Q. Trials using the NIHSS as a measure of stroke severity also had a greater reduction in sample size (37-39%) than those using other severity scales (29-30%). However, different studies used

different measures of severity and outcome and it was not possible to compare directly the relative benefits of using any particular scale.

FIGURE 5.4

Odds ratios for the unadjusted models and the adjusted models for a simulated treatment effect of 0.57; the points are the mean effect from the 10,000 simulations. Each point on the x axis is an individual trial.



5.4 DISCUSSION

The increasing number and size of stroke trials, and failure to identify effective acute treatments, are threatening the viability of future studies. Any method which reduces sample size (and hence, the cost and duration of trials) or increases statistical power and thereby improving the likelihood of finding effective interventions, will be welcome. These results show that the efficiency of analyses of functional outcome in stroke trials is improved when outcome is adjusted for three prognostic factors: age, sex and stroke severity. Such inclusion of covariates allows a substantial reduction in sample size to be achieved, in this case by approximately one-quarter (lower end of the interquartile range), for a given power; conversely, statistical power can be increased for a given sample size. Maintaining sample size has the added benefit of improving the robustness of sub group analyses. Importantly, covariate adjustment appeared to be effective irrespective of the scales used to measure baseline severity and functional outcome.

Other studies have shown that adjustment for baseline covariates improves statistical power. The IMPACT study assessed ways of improving the design and analysis of brain injury trials and found that adjustment for seven predictors of outcome reduced sample size by around 16-23% when analysed using logistic regression on a dichotomised Glasgow Outcome Scale (Hernandez et al., 2006). In contrast to the results presented here, Hernandez looked at two types of covariate adjustment, an adjustment for seven prognostic factors and then a model adjusted for the three strongest predictors of outcome from the seven prognostic variables. They found that adjusting for more variables gave a greater reduction in the sample size required, ~25% compared to ~20% (Hernandez et al., 2006). I have only looked at adjusting for one set of covariates, but as baseline severity is such a strong predictor of outcome, the

addition of others would probably not greatly alter the results found. Another previous paper by Hernandez also looked at the effect of adjustment on logistic regression; this project was more comprehensive and compared different levels of treatment effect, covariate effect, outcome incidences and covariate prevalences (Hernandez et al., 2004). They found, akin with this current analysis, that the reduction in sample size gained was independent of level of treatment effect. Interestingly, they found that adjustment for covariates which were imbalanced across treatment groups did not increase power and therefore they advised against this. They found that the greatest reductions in sample size were associated with adjustment for moderate to strong predictors of outcome. They conclude that randomised controlled trials should consider adjusted analyses and that the covariates included should be either prognostically important and therefore pre-specified in the trial protocol, or are shown to have a statistically significant relationship to outcome. Similar results have been reported for time to event analyses using the Cox proportional hazards model (Hernandez et al., 2006). However, this OAST analysis is the first to look at the effect of adjustment on ordinal logistic regression, and assessment of potential benefits on sample size.

Adjustment addresses imbalances in baseline prognostic factors which occur by chance with simple randomisation. Historically, the interpretation of several stroke trials has been confounded by imbalances at baseline. For example, the large 20,000 patient 'International Stroke Trial' was neutral in its primary univariate analysis but statistically significant following adjustment with a model predictive of outcome (International Stroke Trial Collaborative Group, 1997). Similarly, the SAINT-I trial had a statistically significant result when adjusted for prognostic factors but showed no effect when analysed without covariate adjustment (Lees et al., 2006). Such imbalances in baseline factors

may be reduced using adaptive randomisation (minimisation), a technique which also moderately improves statistical power (Weir and Lees, 2003).

Adjustment for covariates increases the precision of the estimated treatment effect and changes the interpretation of the results, as these are now conditional on the chosen covariates. It is therefore crucial that adjustment is considered at the protocol development stage of setting up a clinical trial and that the covariates are chosen and stated *a priori*; the decision to include covariates, and which ones, at the time of analysis would be incorrect and result in misleading data-driven analyses.

There are several limitations to the present analysis. Firstly, only 20 of the original 55 OAST data sets could be used since many studies did not share baseline data. Although this is unlikely to have changed the present findings qualitatively, it will have reduced the power of the analyses. In this respect, it is vitally important that trialists, both academic and commercial, share data following publication of the main trial paper for use in other projects (such as OAST and VISTA (Ali et al., 2007)) so that its value is maximised. Secondly, only three covariates (age, sex severity) were used so as to maximise the number of included data sets. However, this limitation is not important since, although there are many baseline characteristics which have prognostic significance (e.g. atrial fibrillation, temperature, blood pressure, and serum glucose), severity has been consistently identified as the most powerful predictive factor of outcome (Sprigg et al., 2007) and explains most of the variation in covariate-adjusted analyses (as shown here). Age and sex are added since they are key biological variables. Thirdly, beneficial effects on study power/sample size may not translate to other clinical areas; stroke is unusual in having such a strong predictor of outcome in the form of baseline

severity and, as such, the reduction in sample size gained by adjusting for covariates will be greatly influenced by the strength of the relationship between severity and outcome. Lastly, methods of analysis which assess shifts in outcome over the entire distribution, although popular with physicians, may not be thoroughly understood and therefore greater input may be needed from statisticians (Savitz et al., 2008). Additionally, further work needs to address what magnitude of shift in outcome is meaningful to patients, healthcare professionals and health funders.

5.5 SUMMARY

In summary, trialists should consider using key prognostic variables in the analysis of functional outcome in stroke trials when using ordinal analyses. This will allow trials to be smaller for a given statistical power, or to achieve greater statistical power for a given sample size. Nevertheless, existing knowledge that covariate adjusted logistic regression is more powerful than unadjusted analyses has not led to all trials moving to this approach, perhaps because of uncertainty about the interpretation and presentation of trial results based on adjusted analyses. Hence, in practical terms trialists may, at least in the short term, want to power their study for an unadjusted analysis and then analyse the completed trial with adjustment for covariates, thereby increasing the statistical power but maintaining a large enough sample size to carry out an unadjusted analysis as a secondary endpoint. Nevertheless, the results need to be reported in the context of the included covariates.

TABLE 5.1

Baseline data for the extra trials added to the OAST database.

Trial	Trial characteristics					Baseline		
	Sample size	Intervention	Time (hr)	Active groups	Follow up (mo)	Age (median [IQR])	Male (%)	Baseline severity (NIHSS) (median [IQR])
Acute:								
DESTINY (Juttler et al., 2007)	32	Decompressive surgery	12-36	1	6	45 [38-52]	47	22 [20-24]
Minocycline (Lampl et al., 2007)	151	Minocycline	6-24	1	3	69 [59-75]	65	6 [5-9]
Statin withdrawal (Blanco et al., 2007)	89	Statin withdrawal	24	1	3	69 [63-76]	51	14 [10-18]

IQR: Inter quartile range; NIHSS: National Institute of Health Stroke Scale; hr: hours; mo: months

TABLE 5.2

Primary outcome for the extra trials added to the OAST database.

	Barthel Index (median [IQR])	Rankin Scale (median [IQR])	Death rate (%) per month (control group)	Outcome scale	Type of analysis	Analysis approach used in the primary publication	Trial result (+/0/-)
Acute							
DESTINY (Juttler et al., 2007)		4 [3-6]	8.9	mRS	O	Wilcoxon test mRS	+
Minocycline (Lampl et al., 2007)	100 [70-100]	1 [1-3]	0	mRS	C	t-test mRS	+
Statin withdrawal (Blanco et al., 2007)		2 [1-4]	0.2	mRS	D	Chi square test (mRS >2)	-

D: Dichotomised or data collapsed into multiple groups; O: Ordinal method; C: Continuous method +: Beneficial intervention effect; -: Harmful intervention effect; 0: No intervention effect but part of a meta analysis showing a treatment effect; IQR: Inter quartile range.

TABLE 5.3

Included trials.

Trial	Intervention	Outcome scale	Baseline severity scale	Sample size
AbESTT	Abciximab	mRS	NIHSS	400
ASSIST 07	Selfotel	BI	NIHSS	138
ASSIST 10	Selfotel	BI	NIHSS	432
Citicoline 1	Citicoline	BI	NIHSS	259
Citicoline 7	Citicoline	BI	NIHSS	394
Citicoline 10	Citicoline	mRS	NIHSS	100
Citicoline 18	Citicoline	BI	NIHSS	899
DCLHb	DCLHb	mRS	NIHSS	85
DESTINY	Decompressive surgery	mRS	NIHSS	32
Dover	Stroke unit	mRS	Own	235
Ebselen	Ebselen	BI	Own	298
FOOD 3	NG tube	mRS	Own	321
INWEST HIGH	Nimodipine	BI	ORGO	194
INWEST LOW	Nimodipine	BI	ORGO	201
IST	Aspirin	3Q	Own	19435
MAST-I	Aspirin	mRS	Own	309
Minocycline	Minocycline	mRS	NIHSS	151
RANTTAS I	Tirilazad	BI	NIHSS	660
RANTTAS II	Tirilazad	BI	NIHSS	126
Statin withdrawal	Statin withdrawal	mRS	NIHSS	89
STIPAS	Tirilazad	BI	NIHSS	111
TESS I	Tirilazad	BI	UNSS	450
TESS II	Tirilazad	BI	UNSS	355

BI: Barthel Index; mRS: modified Rankin Scale; NIHSS: National Institute of Health Stroke Scale; ORGO: Orgogozo Scale; UNSS: Unified Neurologic Stroke Scale.

TABLE 5.4

Relationship between age, sex and severity and outcome using ordinal logistic regression. Statistically significant results ($p<0.05$) are given in bold.

Trial	Relationship to outcome		
	Age	Sex	Severity
AbESTT	**<0.001**	0.126	**<0.001**
ASSIST 07	**0.001**	0.327	**<0.001**
ASSIST 10	**<0.001**	0.280	**<0.001**
Citicoline 1	**<0.001**	0.072	**<0.001**
Citicoline 7	**<0.001**	0.234	**<0.001**
Citicoline 10	**0.03**	**0.030**	**<0.001**
Citicoline 18	**<0.001**	0.056	**<0.001**
DCLHb	**0.036**	0.724	**<0.001**
DESTINY	**0.001**	0.687	0.301
Dover	**0.004**	0.257	**<0.001**
Ebselen	**<0.001**	**0.003**	**<0.001**
FOOD 3	**<0.001**	0.135	**<0.001**
INWEST HIGH	**<0.001**	**0.044**	**<0.001**
INWEST LOW	**<0.001**	0.078	**<0.001**
IST	**<0.001**	**<0.001**	**<0.001**
MAST-I	**<0.001**	**0.012**	**<0.001**
Minocycline	0.49	0.50	**<0.001**
RANTTAS I	**<0.001**	**0.002**	**<0.001**
RANTTAS II	**<0.001**	0.458	**<0.001**
Statin withdrawal	**0.007**	0.12	**<0.001**
STIPAS	**0.005**	0.651	**<0.001**
TESS I	**<0.001**	0.912	**<0.001**
TESS II	**<0.001**	0.442	**<0.001**

TABLE 5.5

Baseline imbalances for age, sex and severity using t-test for age and severity and chi square test for sex. Statistically significant results ($p<0.05$) are given in bold.

Trial	Baseline imbalance		
	Diff in mean age (yrs)	Diff in % male	Diff in mean severity
AbESTT	1.32	7.50	0.50
ASSIST 07	**3.60**	7.61	0.69
ASSIST 10	1.93	3.10	0.29
Citicoline 1	2.54	2.80	0.19
Citicoline 7	0.56	3.13	0.58
Citicoline 10	4.20	2.08	0.31
Citicoline 18	0.58	4.37	0.55
DCLHb	2.56	11.00	0.60
DESTINY	2.83	0.39	**2.82**
Dover	0.81	0.86	**0.14**
Ebselen	0.15	5.02	4.20
FOOD 3	0.23	0.41	<0.0001
INWEST HIGH	1.08	4.32	3.79
INWEST LOW	0.91	4.45	1.54
IST	0.03	1.07	0.01
MAST-I	0.88	2.97	0.19
Minocycline	1.03	2.72	0.04
RANTTAS I	0.48	4.88	0.67
RANTTAS II	2.18	1.79	1.35
Statin withdrawal	1.47	5.66	0.86
STIPAS	3.17	10.18	1.18
TESS I	1.47	1.00	0.09
TESS II	0.63	2.59	1.07

TABLE 5.6

The median (interquartile range) odds ratios obtained from unadjusted and adjusted models with the reduction in sample size gained from using an adjusted analysis.

Treatment effect (odds ratio (OR))	Unadjusted treatment OR	Adjusted treatment OR	Reduction in sample size (%)
0.95	0.96 (0.96 – 0.96)	0.95 (0.95 – 0.95)	35.3 (21.0 – 42.1)
0.74	0.79 (0.78 – 0.80)	0.73 (0.73 – 0.74)	38.4 (29.4 – 42.7)
0.57	0.65 (0.63 – 0.66)	0.57 (0.56 – 0.57)	38.4 (27.4 – 42.2)

TABLE 5.7

Comparison of z scores and treatment coefficients from the adjusted models with those from the unadjusted models; data given as median percentage and interquartile range.

	Treatment effect		
	0.95	0.74	0.57
Z score, adjusted > unadjusted (%)	52.5 (51.5 - 53.6)	65.0 (59.8 - 69.6)	75.8 (67.6 - 82.7)
Treatment coefficient, adjusted > unadjusted (%)	52.8 (52.0 - 54.3)	67.3 (62.6 - 73.5)	79.2 (71.8 - 87.6)

TABLE 5.8

Reduction in sample size (%) for trials sub grouped by type of severity scale and functional outcome scale; median and inter quartile range.

	N trials	Treatment effect (Odds ratio)		
		0.95	0.74	0.57
Overall	23	35.3 (21.0 – 42.1)	38.4 (29.4 – 42.7)	38.4 (27.4 – 42.2)
Outcome scale:				
Modified Rankin Scale	9	21.0 (18.0 – 39.9)	29.4 (20.7 – 42.6)	27.4 (20.6 – 42.0)
Barthel Index	13	39.6 (33.1 – 45.1)	39.5 (36.1 – 44.4)	39.9 (36.3 – 43.8)
Three Questions	1	20.1	20.7	20.3
Severity scale:				
NIHSS	14	37.5 (31.3 – 42.9)	38.9 (34.4 – 43.2)	39.2 (34.1 – 42.7)
Other scale	9	30.2 (20.4 – 43.2)	29.6 (20.9 – 42.0)	29.4 (21.2 – 41.4)

CHAPTER 6

AN ASSESSMENT OF OTHER METHODS OF ANALYSES USED IN STROKE TRIALS

PUBLICATIONS/PRESENTATIONS CONTRIBUTING TO THIS CHAPTER

Gray L.J, Bath P.M.W, Collier T (2006) Analysis of the effect of the Cochran Mantel-Haenszel test and patient specific outcome (sliding dichotomy) in randomized stroke trials. *Poster presentation at Joint World Congress on Stroke, Cape Town, October 2006, International Journal of Stroke. 1 (suppl 1): 111-174.*

Gray L.J, Bath P.M.W, OAST Collaborators (2004) Do global outcomes increase efficiency in stroke clinical trials? *Poster presentation at the Research Students Conference, Sheffield, April 2004.*

6.1 INTRODUCTION

The OAST project so far has assessed using various univariate methods of analysis and the effect of taking into account covariates on the results produced from functional outcome data. As discussed previously in the introduction chapter, other types of analysis have also been used; namely the global outcome analysis, patient-specific outcome, and the Cochran Mantel-Haenszel test. This chapter will consider these approaches.

The global outcome analysis, where data from more than one outcome scale is combined, has been used in a number of stroke trials. The NINDS trial tested the thrombolytic agent alteplase against placebo; during the development of this trial, it was decided that choosing one primary outcome scale was too limiting. Instead the trialists chose four scales (mRS, BI, NIHSS and GOS) to cover a number of aspects of stroke recovery rather than focussing on one disability scale. In 1992 the NINDS trial group held a workshop to discuss methods of statistical analysis for trials with multiple pre-specified outcomes (Tilley et al., 1996). The consensus of the participants was that a global test, utilising generalised estimating equations (GEE) modelling, should be used. Here, two or more dichotomised outcomes can be tested simultaneously using a Wald test statistic; the NINDS trial combined the following dichotomised outcomes:

- BI ≥ 95
- mRS ≤ 1
- NIHSS ≤ 1
- GOS $= 1$

The NINDS trial showed a beneficial treatment effect, both on the global outcome and for each individual scale. The "Intravenous Magnesium Efficacy in Stroke" (IMAGES) trial changed their analysis plan during the trial to include a global measure (BI≥95 and mRS≤1) as the primary outcome, after a study using simulated data showed that global outcomes were more powerful than using BI dichotomised at ≥60, which was the trial's original primary outcome (Intravenous Magnesium Efficacy in Stroke (IMAGES) Study Investigators, 2004, Young et al., 2003). Applying a post hoc global analysis to the ECASS trial (BI≥95, mRS≤1, NIHSS ≤1) gave a statistically significant result, compared to the neutral finding of the original analysis (median BI and median mRS) (Hacke et al., 1998).

The second type of analysis which has been suggested takes into account the patient's initial level of stroke severity. Here, the definition of a good outcome varies depending on the baseline severity instead of being constant for all patients (Adams et al., 2004, Berge and Barer, 2002). In the literature this type of analysis has been termed:

- "patient-specific" (Young et al., 2003)
- "responder" (Adams et al., 2004)
- "prognosis-adjusted" (Young et al., 2005)
- "sliding dichotomy" (Murray et al., 2005)

This approach has been taken by a few completed trials. The "Stroke Treatment with Ancrod Trial" (STAT) used a variation of this and defined a favourable outcome as either ≥95 on the BI or at least equal to their pre stroke value at the day 90 assessment (Sherman et al., 2000). The "Abciximab in Emergent Stroke Treatment Trial" (AbESTT) was the first trial to use a full responder

analysis approach as a secondary outcome. This trial used three dichotomisations of the mRS to define a favourable outcome based on the patients baseline NIHSS score (mRS=0 for NIHSS ≤7, mRS≤1 for NIHSS 8-14 and mRS≤2 for NIHSS>14). The trialists found, in line with their primary outcome, that the patient specific outcome showed increased response in the abciximab group (Adams et al., 2004, Abciximab Emergent Stroke Treatment Trial (AbESTT) Investigators, 2005). Unfortunately, the follow-on phase three trial failed to confirm this finding (Adams et al., 2008). A comparison of outcomes in thrombolytic trials found that a patient specific outcome and a normal dichotomisation, which does not take into account baseline severity, gave similar proportions of patients with an excellent outcome, but that the types of patients within this category were quite different. The patient specific outcome categorised fewer mild stroke patients as having an excellent outcome and more patients with a severe stroke. This study found the patient specific outcome to be a better and more clinically relevant outcome (Thomassen et al., 2005). There is also a statistical argument for using this type of analysis, since it increases statistical power, as compared to approaches which do not take baseline severity into account (Young et al., 2005).

The final type of analysis also takes into account covariates, such as severity, but stratifies the analysis by using a Cochran Mantel-Haenszel test, rather than by setting varying definitions of a favourable outcome. This method was used in the SAINT trials, where the primary end point of the mRS was adjusted for the stratification variables: NIHSS, side of infarct and use of alteplase (Lees et al., 2006, Shuaib et al., 2007).

The aim of this part of the OAST project was to test whether these three approaches improve the efficiency of stroke clinical trials.

6.2 METHODS

6.2.1 Trial data

For this part of the project, trials from the OAST database which had measured both mRS and BI were used for the global outcome analysis, and those which had collected data on baseline severity using the NIHSS were used for assessing the patient specific outcome and Cochran Mantel-Haenszel test.

6.2.2 Global outcome

Global outcome analysis (GO) was calculated using the GEE method, as used in the NINDS study (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995, Tilley et al., 1996). In this analysis, a multivariate model was used to combine two dichotomous outcomes BI≥95 and mRS≤1. The model used has the following form:

As two binary variables are being combined $K = 2$, this can be extended to any number of binary variables. Y_{ijk} is the Kth response: $K = 1, 2$ in the ith group: $i = 0\,(\text{control}), 1\,(\text{treatment})$ for the jth subject: $j = 1,2,.....n_i$. The observation vectors for each subject are independent, with mean μ_i and variance $Y_{ijk} = \Phi\mu_{i\,jk}\left(1 - \mu_{i\,jk}\right)$, where Φ allows for over dispersion.

The multivariate model uses a logistic model which models the probability of a good outcome on each scale. The model for the mean $E\left(Y_{ijk}\right) = \mu_{ik}$ is therefore $logit\ \mu_{ik} = \alpha + \beta x_i$ (Tilley et al., 1996).

The GEE method of Lefkopoulou and Ryan is then used to obtain a Wald statistic which simultaneously tests the null hypothesis that the two outcome

measures are equal in the two treatment groups (Lefkopoulou and Ryan, 1993).

6.2.3 Patient specific outcome

The definitions of a favourable outcome suggested by Adams et al were used with equivalent cuts being used for the BI (Adams et al., 2004). A chi square test, without continuity correction, was then applied to the patient specific outcome. See Table 6.1 for definitions.

6.2.4 Cochran Mantel-Haenszel test

The same strata for NIHSS (<8, 8-14, >14) were used for the Cochran Mantel-Haenszel test as used in the patient specific outcome. A favourable outcome was defined as BI≥95 and mRS≤1. The Cochran Mantel-Haenszel test statistic is:

$$M^2 = \frac{\left[\sum_k (n_{11k} - \mu_{11k})\right]^2}{\sum_k Var(n_{11k})}$$

For a set of $2 \times 2 \times K$ tables. Where $\mu_{11k} = E(n_{11}) = \frac{n_{1+k} n_{+1k}}{n_{++k}}$, which is the expected frequency of the first cell in the Kth table, and the variance of cell $(1,1)$ is:

$$Var(n_{11k}) = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{n_{++k}^{\ 2}(n_{++k} - 1)}$$

6.2.5 Comparison of statistical tests

The z scores from the three novel approaches were compared to the z scores from ordinal logistic regression (for trials not testing thrombolytic agents) and the t-test (all trials); ordinal logistic regression and t-test were carried out on the primary outcome scale for the trial. The difference between the z scores was then assessed using a Wilcoxon test, to see if the z scores produced by one test were significantly different to those given by the other. Analyses were carried out in SAS (version 8.2) and Stata (version 7) and significance was taken at $p<0.05$.

6.3 RESULTS

6.3.1 Included trials

Table 6.2 shows the data sets included for each type of analysis. Twelve trials from a mixture of acute and rehabilitation trials had provided data on both the mRS and BI and therefore the global outcome could be calculated. Seventeen and sixteen data sets from acute trials were included in the patient specific outcome and Cochran Mantel-Haenszel test respectively.

6.3.2 Global outcome

Table 6.3 and Figure 6.1 show the comparison of the global outcome with the t-test. There was no significant difference between the z scores produced by the global outcome and those produced by the t-test ($p=0.69$). The comparison with ordinal logistic regression for those trials not testing a thrombolytic agent showed similar results ($p=0.89$, Table 6.4), with Figure 6.2 showing that the global outcome and ordinal logistic regression generally give comparable results.

6.3.3 Cochran Mantel-Haenszel test

Tables 6.3 and 6.4 show no statistical difference between the Cochran Mantel-Haenszel test and both the t-test and ordinal logistic regression (p=0.60 and p=0.77 respectively), although this may be due, in part, to lack of power owing to the limited number of data sets included. Figures 6.3 and 6.4 show that although the z scores are similar for the Cochran Mantel-Haenszel test and the t-test, and the Cochran Mantel-Haenszel test and ordinal logistic regression, the Cochran Mantel-Haenszel test produced consistently smaller z scores (smaller treatment effects) than both other tests (seen when green line falls below zero).

6.3.4 Patient specific outcome

Similar results to the Cochran Mantel-Haenszel test are seen for the patient specific outcome (Tables 6.3 and 6.4, Figures 6.5 and 6.6), with analogous but lower z scores compared to the t-test and ordinal logistic regression (p=0.69 and p=0.70 respectively).

FIGURE 6.1

Z scores from the global outcome and the t-test, with difference between the two. Where the difference falls below the line, the global outcome produces a smaller z score than the t-test. Each point on the x axis is an individual trial.

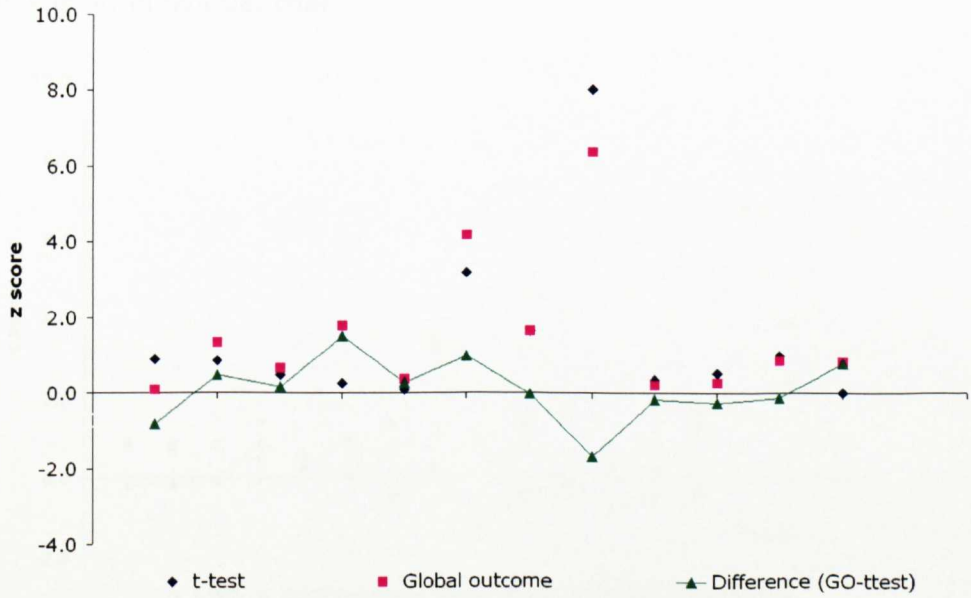


FIGURE 6.2

Z scores from the global outcome and ordinal logistic regression, with difference between the two. Where the difference falls below the line, the global outcome produces a smaller z score than ordinal logistic regression. Each point on the x axis is an individual trial.

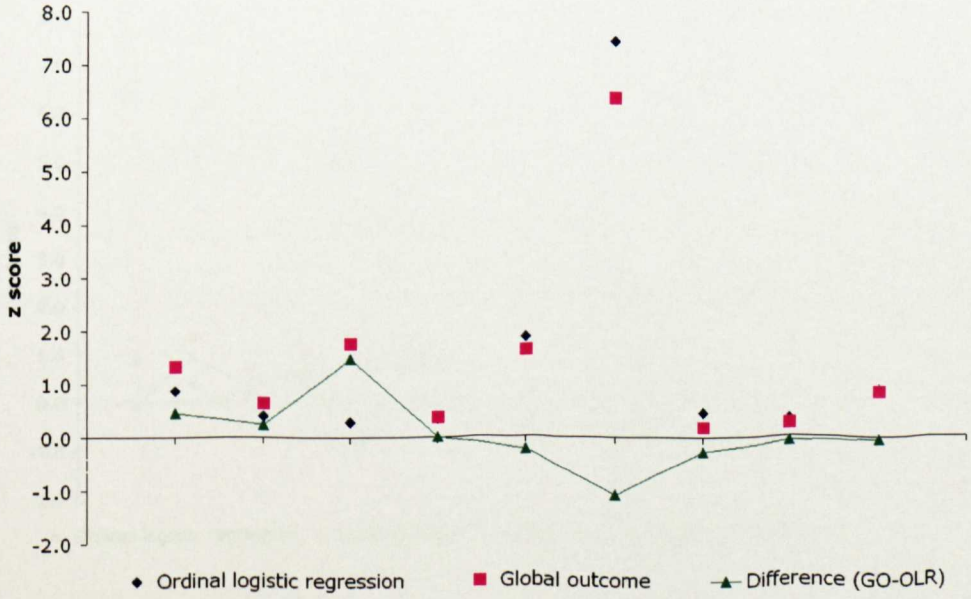


FIGURE 6.3

Z scores from the Cochran Mantel-Haenszel test and the t-test, with difference between the two. Where the difference falls below the line, the Cochran Mantel-Haenszel test produces a smaller z score than the t-test. Each point on the x axis is an individual trial.

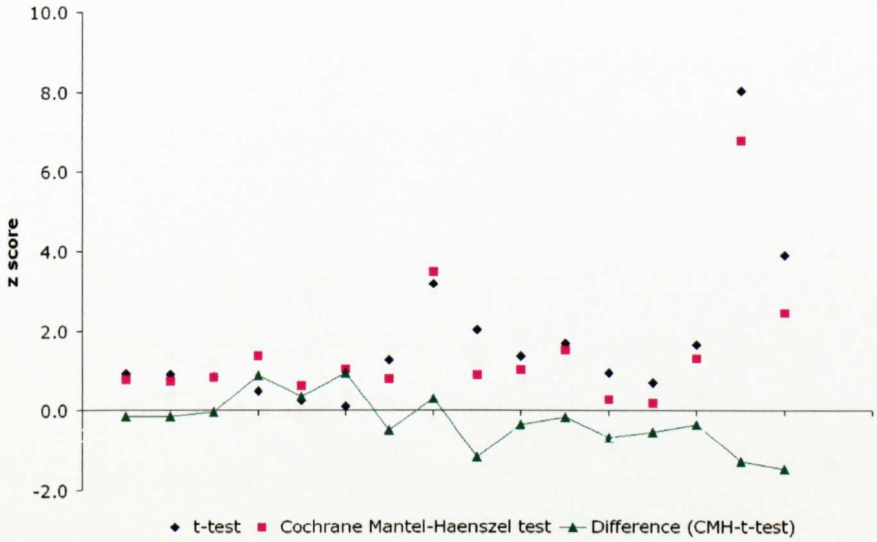


FIGURE 6.4

Z scores from the Cochran Mantel-Haenszel test and ordinal logistic regression, with difference between the two. Where the difference falls below the line, the Cochran Mantel-Haenszel test produces a smaller z score than ordinal logistic regression. Each point on the x axis is an individual trial.

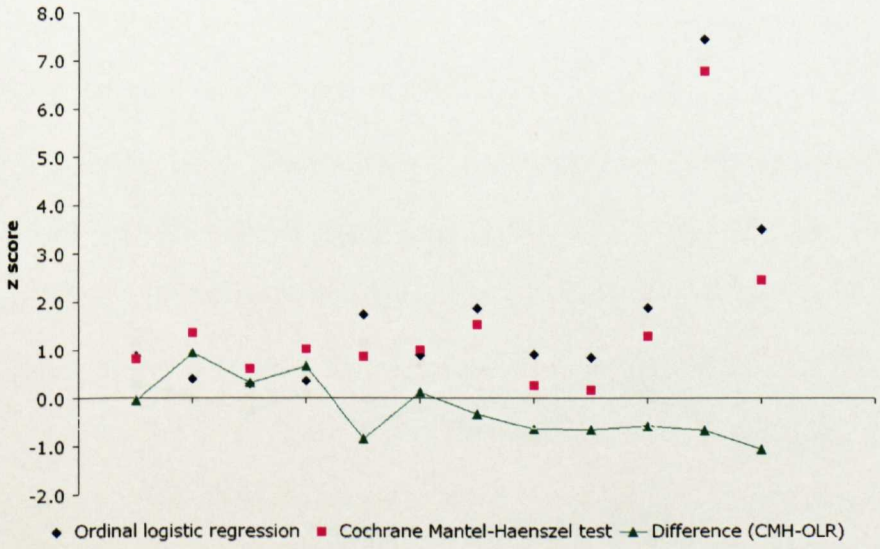


FIGURE 6.5

Z scores from the patient specific outcome and the t-test, with difference between the two. Where the difference falls below the line, the patient specific outcome produces a smaller z score than the t-test. Each point on the x axis is an individual trial.

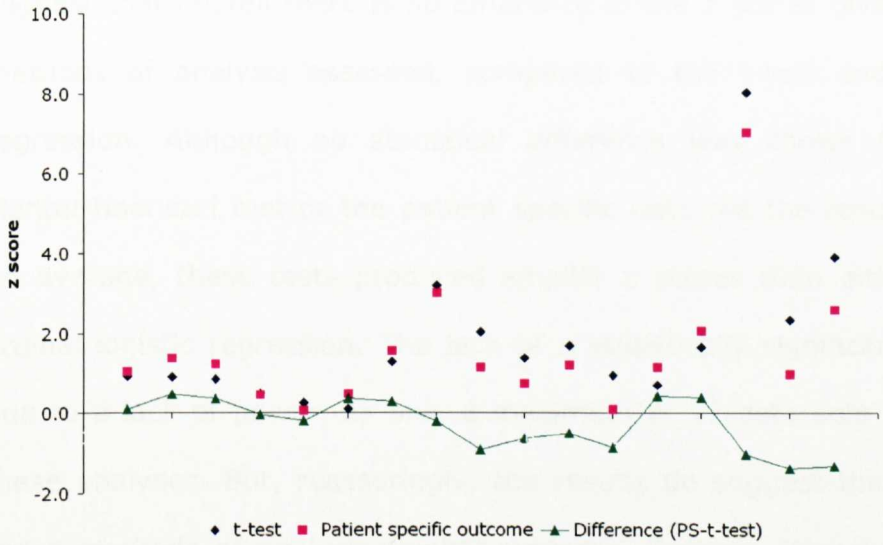
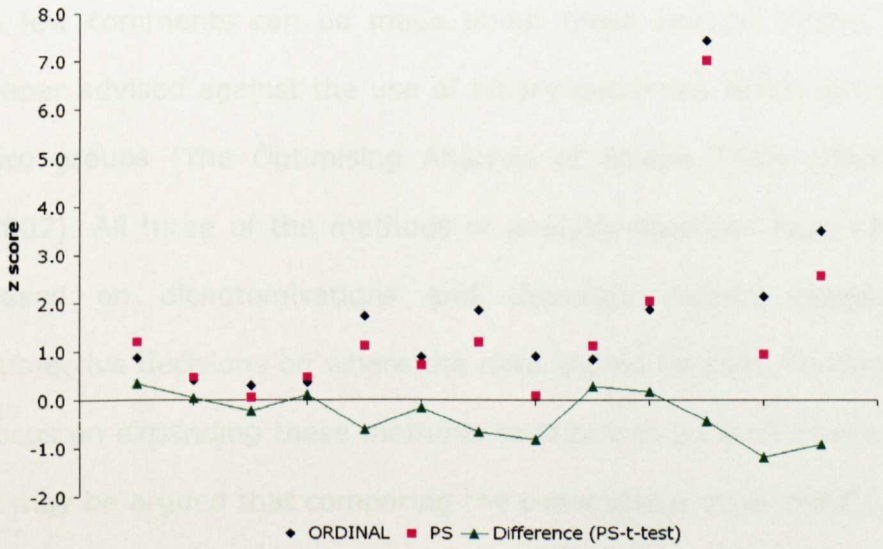


FIGURE 6.6

Z scores from the patient specific outcome and ordinal logistic regression, with difference between the two. Where the difference falls below the line, the patient specific outcome produces a smaller z score than ordinal logistic regression. Each point on the x axis is an individual trial.



6.4 DISCUSSION

This final part of the OAST project has focussed on methods of analysis which have been used in stroke trials but which have not been considered so far: global outcome, Cochran Mantel-Haenszel test and patient specific outcome. These were compared with the t-test and ordinal logistic regression. The results suggest that overall there is no difference in the z scores given with the three methods of analysis assessed, compared to the t-test and ordinal logistic regression. Although no statistical difference was shown for the Cochran Mantel-Haenszel test or the patient specific outcome the results suggest that, on average, these tests produced smaller z scores than either the t-test or ordinal logistic regression. The lack of a statistically significant result may be due to a lack of power, as only a maximum of 17 data sets were included in these analyses. But, reassuringly, the results do suggest that using either of these methods of analysis (global outcome, Cochran Mantel-Haenszel test or patient specific outcome) or the comparators (t-test or ordinal logistic regression) produce very similar results, and therefore one would expect for a beneficial treatment a statistically significant result would be seen with any of these tests.

A few comments can be made about these results. Firstly, the initial OAST paper advised against the use of binary outcomes which dichotomise data into two groups (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2007). All three of the methods of analysis assessed here used here are also based on dichotomisations and therefore require researchers to make subjective decisions on where the data should be split. Further research should focus on expanding these methods to take into account ordinal data. Secondly, it may be argued that comparing the t-test and ordinal logistic regression which analyse data from one scale, with methods which combine data from two more

scales is not valid. However, it is important to assess whether the methods of analysis which combine scales are better than those that do not.

No difference was seen between the methods of analysis assessed and the t-test or ordinal logistic regression. If a difference was seen, then these methods of analysis could be recommended for use in stroke trials. As no difference was seen, and they also have the intrinsic problems of dichotomisation, it might be advantageous to still consider ordinal logistic regression or another univariate approach when deciding how the primary outcome of a trial will be analysed. As shown in the previous chapter, ordinal logistic regression can easily be adjusted for prognostic factors if needed. As the methods of analysis assessed here and ordinal logistic regression performed similarly, an adjusted ordinal logistic regression is likely, therefore, to out perform the global outcome statistic, patient specific and the Cochran Mantel-Haenszel test.

It may be argued that patient-specific outcomes may be useful in trials of agents which both increase the odds of a good outcome, but also have an associated increase in risk, i.e. bleeding in trials of thrombolytic agents. Here, ordinal logistic regression analysis is not suitable and the t-test can not be adjusted for covariates.

The global outcome, patient specific and the Cochran Mantel-Haenszel test may answer interesting clinical questions which are uniquely different to the question posed by the ordinal logistic regression analysis. For example, the responder outcome which sets differing definitions of a "good outcome", depending on the patient's initial level of severity, is assessing a severity related treatment effect. This would therefore presumably classify more patients with a good outcome as compared to an analysis based on a set

definition for all patients. It could also be argued that the global outcome is assessing overall outcome across a number of domains, rather than placing emphasis on one scale.

6.5 SUMMARY

In conclusion this work has shown no additional statistical benefit in using either the global outcome, patient specific outcome, or the Cochran Mantel-Haenszel test over the t-test or ordinal logistic regression.

TABLE 6.1

Definitions of a good outcome for various levels of baseline severity (Adams et al., 2004).

Baseline severity (NIHSS)	Good outcome Barthel Index	Good outcome Rankin Scale
<8	95, 100	0
8-14	75-90	≤1
>14	60-70	≤2

TABLE 6.2

Data sets used for each type of analysis.

	Global outcome	Outcome calculated Cochran Mantel-Haenszel test	Patient specific
Acute			
AbESTT	X	X	X
ASSIST 07		X	X
ASSIST 10		X	X
ATLANTIS A		X	X
ATLANTIS B	X	X	X
Citicoline 01	X	X	X
Citicoline 07	X	X	X
Citicoline 10	X	X	X
Citicoline 18	X	X	X
DESTINY			X
ECASS II		X	X
MAST-E	X		
Minocycline	X	X	X
NINDS	X	X	X
RANTTAS		X	X
RANTTAS II		X	X
Statin withdrawal		X	X
STIPAS		X	X
Rehabilitation			
Gilbertson	X		
Parker ADL	X		
Parker leisure	X		
Total Trials	12	16	17

TABLE 6.3

Comparison with the t-test.

	Number of trials	Mean z score [SD]	Mean diff compared to t-test [SD]	Wilcoxon p value
Global outcome	12	1.55 [1.88]	0.10 [0.84]	0.69
Cochran Mantel-Haenszel test	16	1.50 [1.63]	-0.27 [0.69]	0.60
Patient specific outcome	17	1.53 [1.62]	-0.27 [0.64]	0.69

TABLE 6.4

Comparison with ordinal logistic regression.

	Number of trials	Mean z score [SD]	Mean diff compared to ORL (95% CI)	Wilcoxon p value
Global outcome	9	1.50 [1.92]	0.06 [0.68]	0.89
Cochran Mantel-Haenszel test	12	1.52 [1.76]	-0.23 [0.64]	0.77
Patient specific outcome	13	1.47 [1.81]	-0.31 [0.50]	0.70

CHAPTER 7

EXTENDING THE OAST PROJECT TO STROKE PREVENTION TRIALS

PUBLICATIONS/PRESENTATIONS CONTRIBUTING TO THIS CHAPTER

Bath P.M.W, Geeganage C, **Gray L.J,** Collier T, Pocock S. (2008) Use of ordinal outcomes in vascular prevention trials: comparison with binary outcomes in published stroke trials. *Stroke DOI: 10.1161/STROKEAHA.107.509893.*

Sare G.M, **Gray L.J**, Bath P.M.W (2008) Association between hormone replacement therapy and subsequent cerebrovascular, cardiovascular and thromboembolic disease: a meta analysis. *European Heart Journal DOI:10.1093/eurheartj/ehn299.*

Bath P.M.W, Geegangage C, **Gray L.J**, Collier T, Pocock S (2007) Can we improve the statistical analysis of vascular prevention trials? Assessment of ordinal outcomes. *Poster presentation at International Stroke Conference, San Francisco, USA, February 2007. Stroke, 38 (2): 523.*

Gray L.J, Sare G.M, Bath P.M.W (2007) Association between hormone replacement therapy and subsequent cerebrovascular, cardiovascular and thromboembolic disease: a meta analysis. *Platform presentation at the UK Stroke Forum, Harrogate, December 2007.*

Geeganage C.M, Bath P.M.W, **Gray L.J**, Collier T, Pocock S (2006) Optimising the analysis of stroke prevention trials (OAST-P): assessment using ordered rather than dichotomous outcomes? *Oral presentation at Annual Scientific Meeting of the British Hypertension Society. September 2006. Journal of Human Hypertension 20: S3.*

Bath P.M.W, **Gray L.J**, Geeganage C, Collier T, Pocock S (2006) Optimising the analysis of stroke prevention trials (OAST-P): pilot assessment using ordered rather than dichotomous outcomes. *Poster presentation at the European Stroke Conference, Belgium. May 2006. Cerebrovascular Diseases 21(suppl 4): 121.*

Bath P.M.W, **Gray L.J** (2005) Association between hormone replacement therapy and subsequent stroke: a meta analysis. *British Medical Journal 330 (7487):342.*

7.1 INTRODUCTION

The OAST project has shown that the design and analysis of acute stroke trials can be improved through the use of ordinal methods of analysis. The application of ordinal methods to stroke trials could increase the statistical power to find treatment differences or reduce sample size, which in turn will improve the quality of stroke trials and reduce their complexity and cost.

Trials looking at the prevention of first (primary prevention) or recurrent (secondary prevention) strokes have been more successful in finding new treatments than acute stroke trials, with effective strategies being based on antithrombotic agents, carotid endarterectomy, blood pressure and cholesterol lowering. However, this success has made subsequent trials more difficult as the absolute risk of recurrence, and therefore event rates, have fallen dramatically over time. Figure 7.1 demonstrates this trend by plotting the stroke rate in the control group for each trial included in the OAST prevention project. The regression line shows that the stroke rate has decreased in recent years ($p=0.01$). Figure 7.2 shows the increase in the sample size of stroke prevention trials in recent years. This trend is likely to continue as new and effective interventions are added. Since absolute event rates are a key component in sample size calculations for binary (stroke/no stroke) outcomes, low rates equate to larger trials. Another pressure on performing prevention trials is that their number has increased as new prophylactic strategies are tested (Figure 7.3). The combination of more and larger trials means it is becoming increasingly difficult to find sufficient patients to enrol into new studies.

FIGURE 7.1

Control group stroke rate (%) by date of trial publication for all trials included in the OAST prevention project. The red line gives the regression slope, for every year increase the stroke rate decreases by -0.2 (p=0.01).

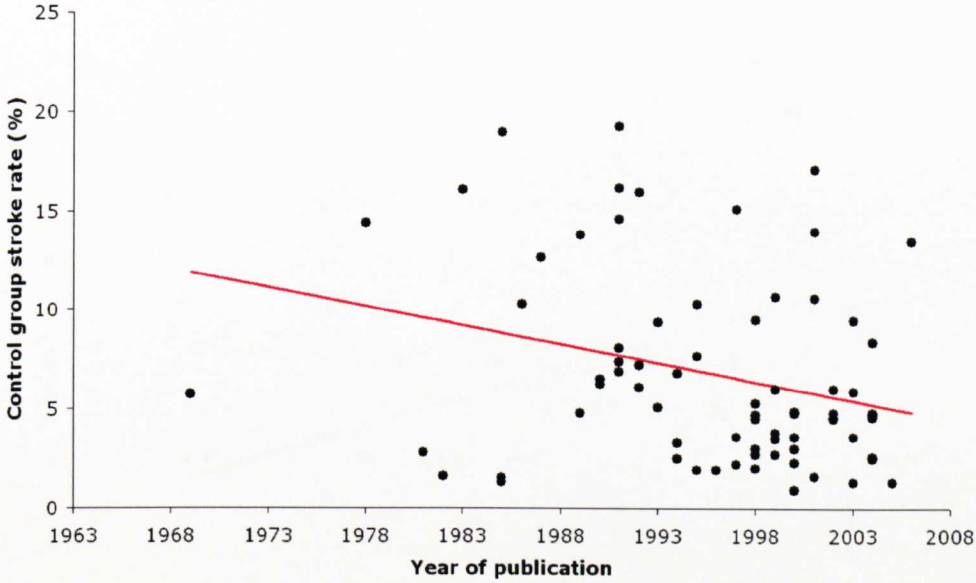


FIGURE 7.2

Sample size by date of trial publication for all trials included in the OAST prevention project. The blue line gives the regression slope, for every year increase sample size increases by 144 patients (p=0.03).

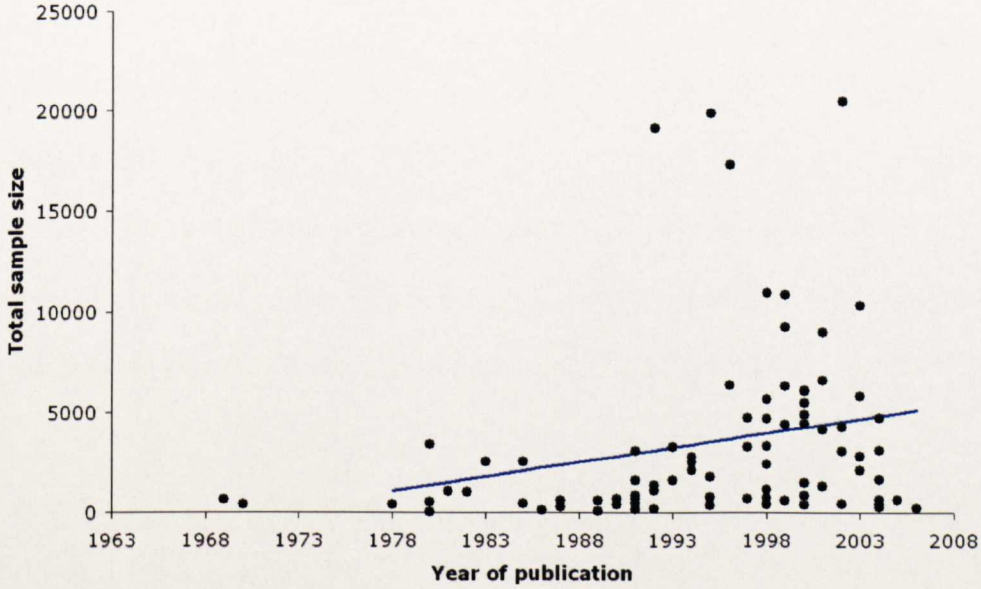
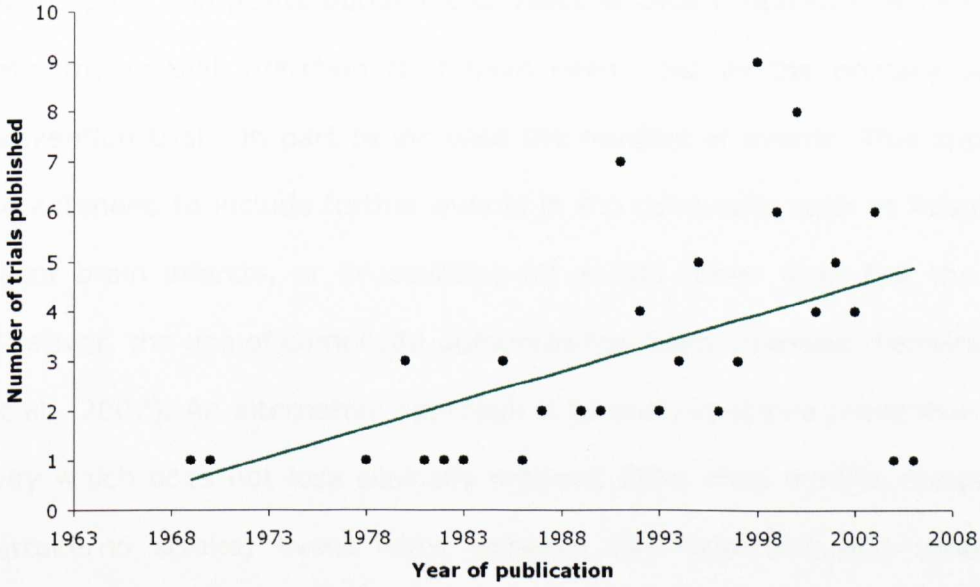


FIGURE 7.3

Number of trials published by year for all trials included in the OAST prevention project. The green line gives the regression slope, for every year increase the number of trials published increases by 0.1 (p=0.01).



It may be possible to use the results of the acute OAST project to influence the design and analysis of stroke prevention trials, in the hope of bringing sample sizes down while maximising the potential to demonstrate benefit.

In the past, composite outcomes of vascular death, non-fatal stroke, and non-fatal myocardial infarction (MI) have been used as the primary outcome in prevention trials, in part to increase the number of events. This approach can be extended to include further events in the composite such as hospitalisation, silent brain infarcts, or by counting all events rather than just the first one. However, the use of composite outcomes has been criticised (Ferreira-González et al., 2007). An alternative approach is to analyse stroke prevention trials in a way which does not lose clinically relevant data. Most studies compare binary (stroke/no stroke) event rates between the treatment and control group. However, stroke events may be fatal or non-fatal, so trichotomous ordinal outcomes (fatal event/non-fatal event/no event) can be analysed. This approach can be extended to four (fatal stroke/severe non-fatal stroke/mild stroke/no stroke) or five (fatal stroke/severe non-fatal stroke/mild stroke/TIA/no event) levels. Similar ordered categorical outcomes can be developed for MI and composite vascular outcomes, as well as other vascular events, such as heart failure and bleeding. Such polytomisation of events assumes that the ordering of events is meaningful, i.e. that fatal stroke events are considered more severe than non fatal ones. If so, ordinal outcomes may be more informative to patients, carers, healthcare professionals and government than binary outcomes.

This part of the project aims to compare the relative efficiencies of using and analysing binary and polytomous ordinal outcomes from vascular prevention trials. This part of the OAST project will be referred to as 'OAST prevention'.

Vascular trials involving non stroke patients and those measuring non stroke outcomes are included since, stroke patients suffer subsequent non stroke vascular events, and those with other vascular conditions can go on to have a stroke. Here the term 'vascular event' refers to stroke, or MI. Taking this approach means the findings are generalisable across the field of vascular medicine.

7.2 METHODS

7.2.1 OAST prevention data set

In contrast to the acute OAST project, the OAST prevention data set is entirely extracted from the trial publications and individual trial data was not sought. All data was extracted and collated by Dr Chamila Geeganage, for full details see (Bath et al., 2008). In brief, data were collated from randomised controlled trials assessing primary or secondary vascular prevention, i.e. preventing first or recurrent events respectively, which were either beneficial or harmful according to the trial publication, or were included in a meta analysis showing benefit or harm; trials in a meta analysis showing no statistically significant treatment effect were excluded. This approach follows the acute OAST project (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2007).

Published studies fulfilling these criteria were identified from electronic searches of the Cochrane Library and included studies of antithrombotic, blood pressure or lipid lowering therapy, carotid endarterectomy, and hormone replacement therapy. Trials were excluded if they did not include adequate ordered categorical information for at least one vascular outcome.

The numbers of subjects at the end of follow-up having a vascular event were obtained, where available, for each treatment group (active, control) from the

primary trial publication. In factorial trials, or those having more than two treatment groups, data were analysed for each active comparison versus control. Data were assessed by intention-to-treat where possible.

7.2.2 Statistical tests

Ten different statistical tests for assessing treatment effect were compared:

(i) Chi-square 2x2 test – stroke versus no stroke

(ii) Chi-square 2x2 test - death versus alive

(iii) Chi-square test across all categories (unordered data) – e.g. fatal stroke/ non fatal stroke/ no stroke

(iv) Cochran-Armitage trend test (ordered data) – e.g. fatal stroke/ non fatal stroke/ no stroke

(v) Ordinal logistic regression

(vi) Median test

(vii) Wilcoxon test (adjusted for ties)

(viii) Robust rank test

(ix) t-test

(x) Bootstrap of difference in mean rank (with 3x3,000 cycles)

The tests compared were used in the same way as in Chapter 3 (see Chapter 3 for detail). Analyses were carried out in SAS (version 8.2) and Stata (version 7); significance was taken at $p<0.05$ for analyses of trials and $p<0.01$ for ANOVA.

7.2.3 Comparison of statistical tests

Each data set was analysed using each statistical test. The results were then ordered within each trial and given a rank, with the lowest rank given to the test which produced the smallest p-value within that trial. A two-way analysis of variance test (Friedman's with adjustment for ties) was then performed to assess which statistical test produced the lowest ranks (i.e. the most statistically significant values). Duncan's multiple range test was used to assess the ordering of tests and determine where significant differences between tests were present. The number of statistically significant (at 5%) results found for each test was also assessed.

The analysis was repeated for six types of vascular outcome:

(i) Three level stroke: fatal stroke/ non fatal stroke/ no stroke

(ii) Four level stroke: fatal stroke/ severe non fatal stroke/ mild non fatal stroke/ no stroke

(iii) Four level stroke/TIA: fatal stroke/ non fatal stroke/ TIA/ no stroke

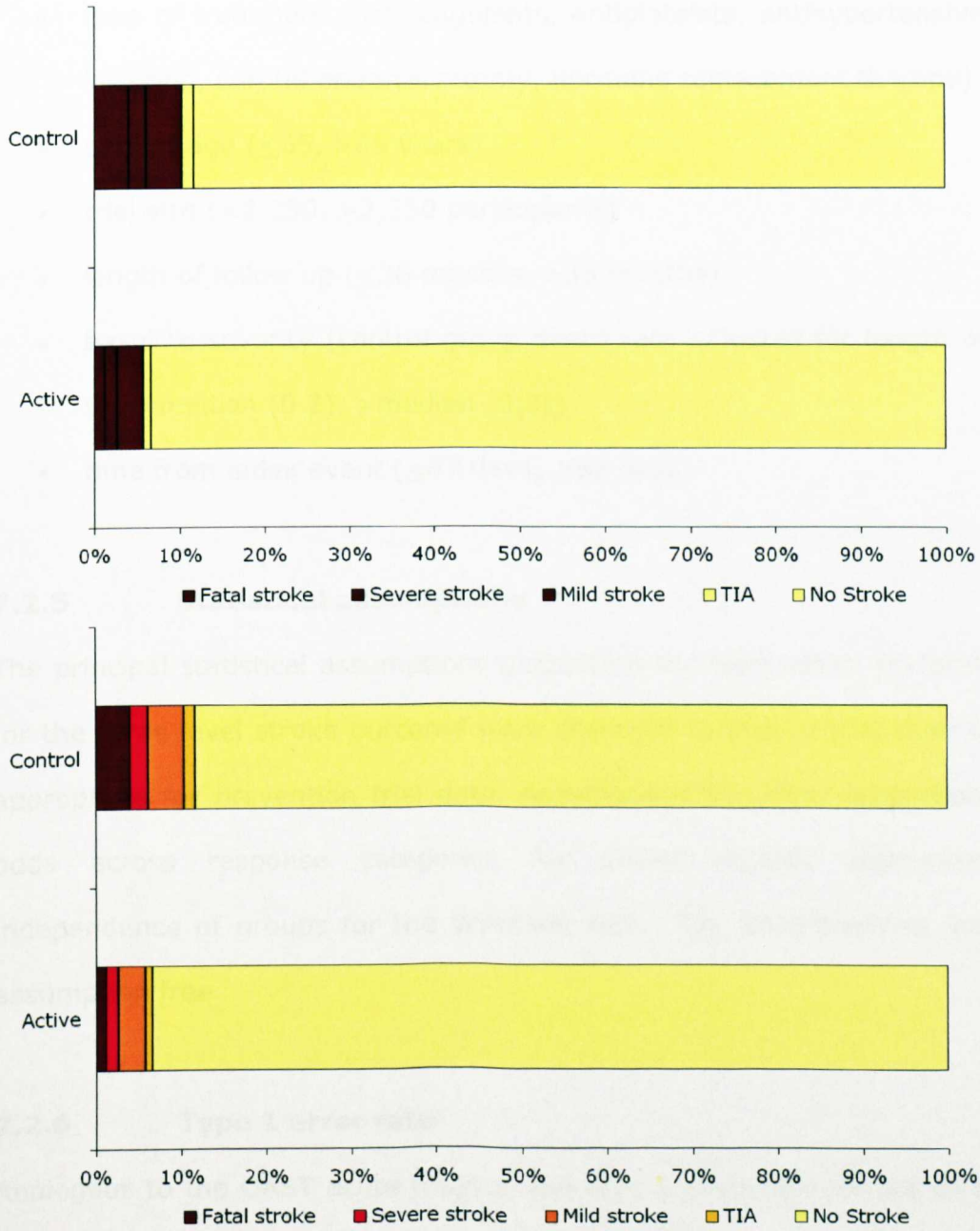
(iv) Five level stroke/TIA: fatal stroke/ severe non fatal stroke/ mild non fatal stroke/ TIA/ no stroke (see Figure 7.4 for an example)

(v) Three level MI: fatal MI/ non fatal MI/ no MI

(vi) Three level vascular (composite of stroke or MI) event: fatal vascular event/ non fatal vascular event/ no vascular event

FIGURE 7.4

Example of the five level stroke/TIA outcome compared to a standard stroke versus no stroke outcome, using data from the HEP trial (Coope and Warrender, 1986).



7.2.4 Sub group analysis

Sub group analyses were performed for the three level stroke outcome by assessing the efficiency of the different tests for differing trial characteristics:

- type of prevention (primary, secondary)
- type of treatment (anticoagulants, antiplatelets, antihypertensives, lipid lowering, carotid endarterectomy, hormone replacement therapy)
- patient age ($\leq$65, >65 years)
- trial size (<2,250, $\geq$2,250 participants)
- length of follow up ($\leq$36 months, >36 months)
- baseline severity (control group death rate adjusted for length of follow up, $\leq$median (0.2), >median (0.2))
- time from index event ($\leq$87 days, >87 days)

7.2.5 Statistical assumptions

The principal statistical assumptions underlying the tests which performed well for the three level stroke outcome were assessed to ensure that their use was appropriate for prevention trial data. Assumptions included: proportionality of odds across response categories for ordinal logistic regression, and independence of groups for the Wilcoxon test. The bootstrapping method is assumption free.

7.2.6 Type 1 error rate

Analogous to the OAST acute project, the type 1 error rate for the three most efficient statistical tests for the three level stroke outcome were tested using data from five representative trials. From these 1,000 data sets were generated, using random sampling with replacement, in which any treatment difference could have occurred only by chance. Tests maintaining adherence to

the nominal type I error rate would expect to see a significant result in around 50 of the 1000 data sets (5%).

7.3 RESULTS

7.3.1 Trials

Of 243 identified trials, 101 (416,020 subjects) were included, these comprising 35 primary and 66 secondary prevention studies. There were 142 trials excluded, mainly because their published data did not distinguish between fatal and non-fatal vascular events so that three level data could not be calculated. For full details see (Bath et al., 2008).

7.3.2 Stroke

The results of the statistical tests differed significantly for the three level stroke outcome (85 trials, 335,305 subjects) (ANOVA $p<0.0001$) (Table 7.1); ordinal analyses ranked above binary approaches with the Wilcoxon test, bootstrapping (difference in mean rank) and ordinal logistic regression performing significantly better than the other methods. Similar results were seen for the other stroke outcome assessments: four level stroke outcome, four level stroke/TIA outcome, and the five level stroke/TIA outcome (each ANOVA $p<0.0001$) (Table 7.2).

Although the absolute ordering of the tests varied across the outcomes, ordinal tests always performed better than binary ones. Six trials gave sufficient data to compare qualitatively three, four and five level data; four level stroke/TIA outcome and five level data stroke/TIA outcome appeared to be the most efficient approaches (Table 7.3). When assessed by how many trials were statistically significant with each of the ten tests (beneficial or harmful but not ineffective), those tests which did not collapse the data into groups again out-

performed other approaches. For example, the Wilcoxon test gave a statistically significant result in 44% of trials in comparison with the chi square 2x3 test at 32% (Figure 7.5).

7.3.3 Myocardial infarction

Fifty-eight trials (232,515 subjects) gave data for the three level MI outcome. The analyses differed significantly for the three level MI outcome ($p<0.0001$) with ordinal approaches performing better than binary (Table 7.2).

7.3.4 Composite vascular event

Forty-three trials (204,108 subjects) gave data for the three level composite vascular outcome. Ordinal tests performed best ($p<0.0001$) with the Wilcoxon test, bootstrapping (the difference in mean rank) and ordinal logistic regression ranking highest (Table 7.2).

7.3.5 Sub group analyses

The ordering of statistical tests, with ordinal more efficient than binary, was maintained for all sub groups of trials irrespective of type of prevention and treatment, average age of patients, trial size and length of follow-up, risk of death or stroke, and time from index event for the three level stroke outcome (Table 7.4). When considering the 19 trials (27 data sets) with a high event rate (>10% overall) ordinal tests remained most efficient. Published hazard ratios (which take into account the time to event, as derived from the Cox proportional hazards model) for stroke were available for 36 trials; a comparison of the 11 statistical tests, including Cox results, revealed bootstrapping, Wilcoxon test and ordinal logistic regression to be as good if not slightly superior to the Cox model (Duncan's multiple range test) (Table 7.5).

7.3.6 Statistical assumptions

The proportionality of odds assumption for ordinal logistic regression was not violated ($p>0.05$) in 79 of 85 trials with three level stroke data (see Figure 7.6).

7.3.7 Type 1 error

The type 1 error analysis showed that the top performing statistical tests (ordinal logistic regression, Wilcoxon test) were not overly sensitive and statistically significant treatment effects were only found where they are likely to be present (see Table 7.6). Figure 7.7 shows that the odds ratios were similar for different strata of severity for three level stroke, four level stroke/TIA, and five level stroke/TIA outcome.

FIGURE 7.5

The number of significant trials (p<0.05) for each statistical test for the three level stroke outcome.

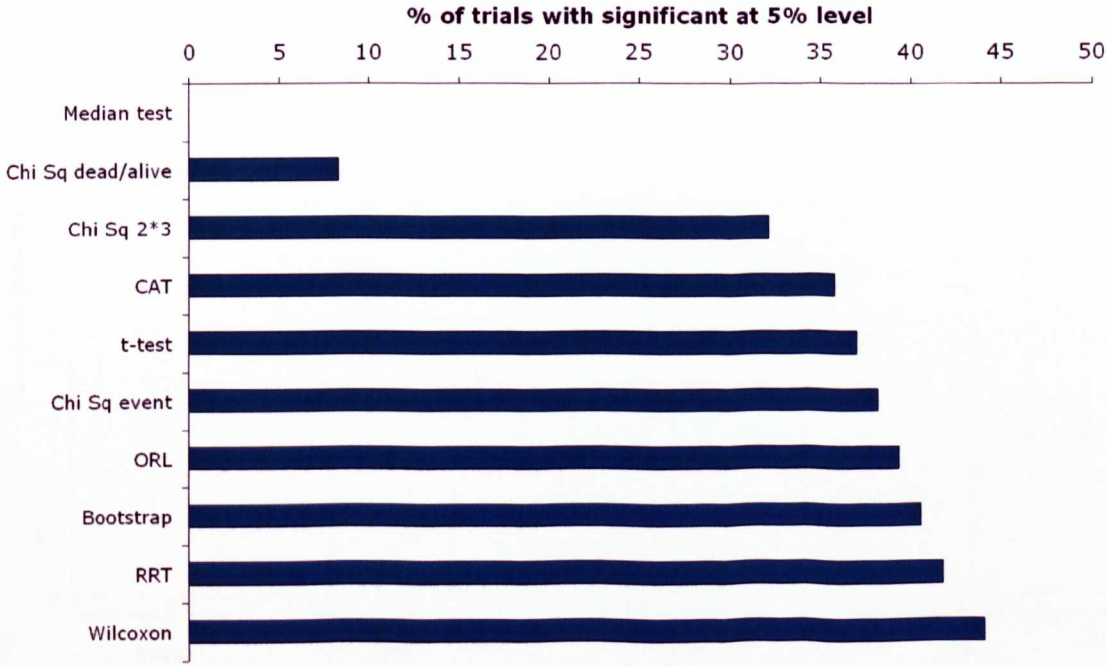


FIGURE 7.6

The p values from the likelihood ratio test for the proportional odds assumption for the three level stroke outcome. P<0.05 indicates non proportional odds. Dotted line is at p=0.05.

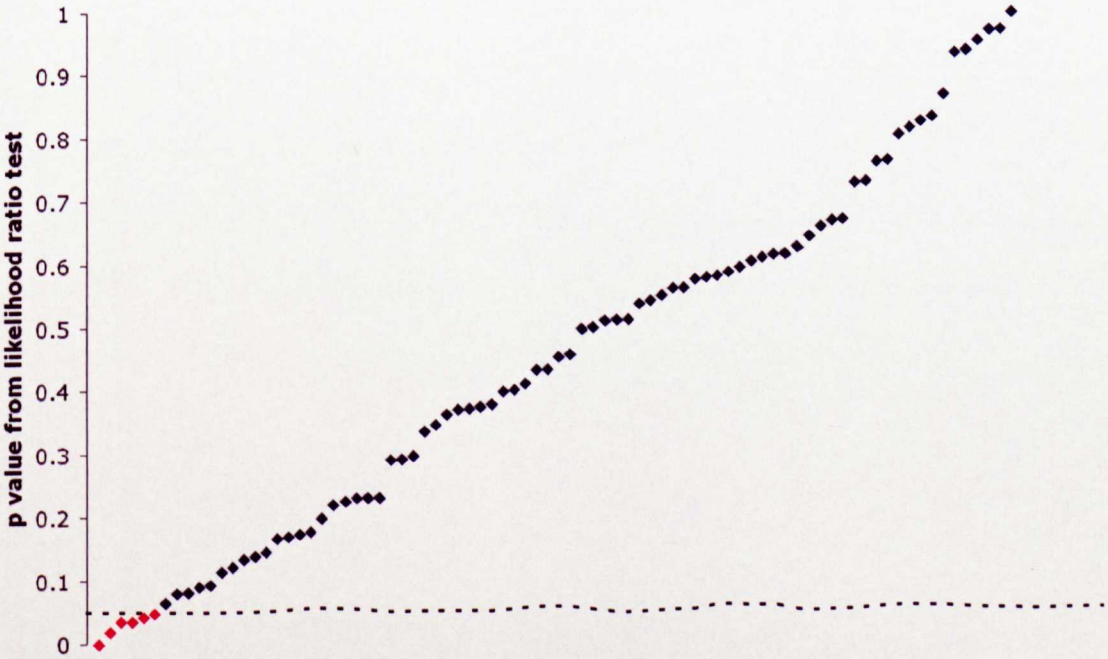
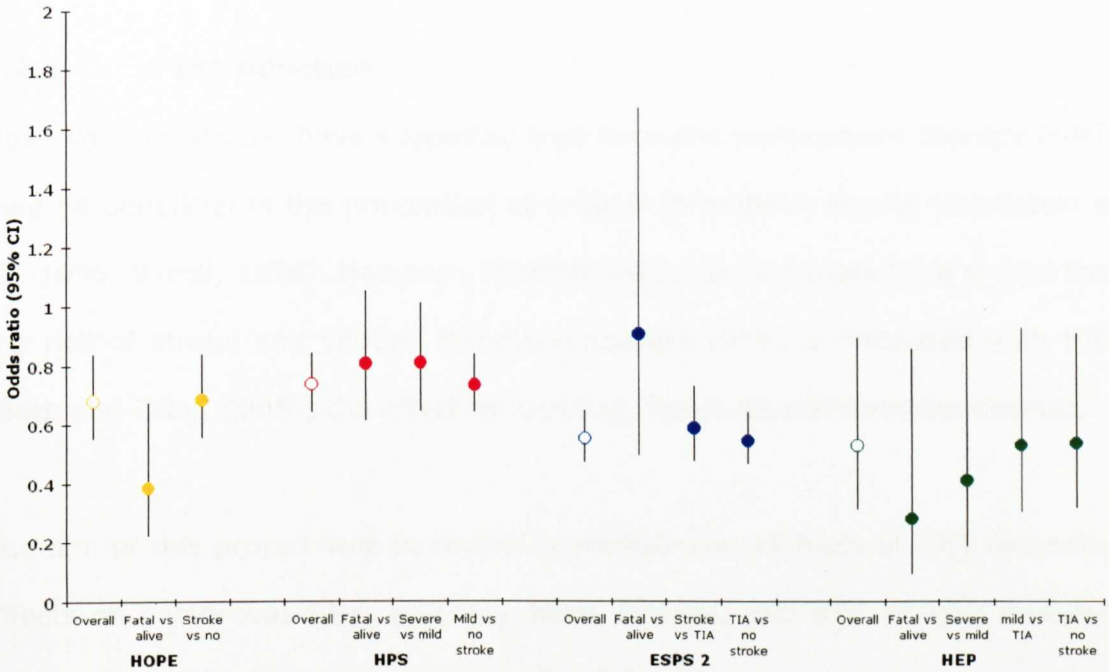


FIGURE 7.7

Odds ratios across four trials (by ordinal logistic regression) and by individual outcome levels to illustrate the assumption of proportionality of odds.



7.4 HORMONE REPLACEMENT THERAPY EXAMPLE

This section describes in more detail an example where the ordinal approach to analysis has been used.

7.4.1 Introduction

Observational studies have suggested that hormone replacement therapy (HRT) may be beneficial in the prevention of arterial thrombotic events (Grodstein et al., 1996, Sarrel, 1996). However, randomised controlled trials have shown that the risk of stroke and venous thromboembolism (VTE) is increased with HRT (Bath and Gray, 2005); the effect on coronary heart disease remains unclear.

The aim of this project was to review systematically all trials of HRT assessing effects on cerebrovascular, coronary heart disease, and VTE events; analyses assessed both the frequency and severity of events.

7.4.2 Identification of trials

Completed and published non-confounded randomised controlled trials of HRT versus no HRT (open or placebo-controlled) were included. Trials had to report event rates for one or more of cerebrovascular (CVD), coronary heart disease (CHD) or venous thromboembolism (VTE). Non-English language publications were excluded. Publications were identified from searches of The Cochrane Library, Embase, Medline (to May 2007), previous reviews (Wren, 1998, Zec and Trivedi, 2002, Collins, 2002, Salpeter et al., 2004, Bath and Gray, 2005, Gabriel et al., 2005), and reference lists from identified articles.

7.4.3 Data extraction

Vascular events (identified as adverse events in some trials) were extracted from the study papers, ideally by intention-to-treat, and included cerebrovascular disease (CVD) (stroke, TIA), coronary heart disease (CHD) (MI, sudden cardiac death, unstable angina (UA)) and VTE disease (deep vein thrombosis, pulmonary embolism, cerebral venous thrombosis). Each outcome (e.g. stroke, TIA, MI etc.) was counted separately and as total outcomes under the pooled headings CVD, CHD and VTE as above. Where sufficient information was given, events were further categorised by severity. If data were taken from lists of adverse events rather than tabulations of outcomes, the trial was only included if it could be determined that adverse events had been reported for each treatment group. Where it was possible to ascertain that more than one event occurred in a single subject, the most severe event was counted, i.e. fatal rather than non-fatal stroke. DVT and PE were counted as separate events but the VTE total represents the most severe event in a single patient.

7.4.4 Statistical analysis

The effect of HRT on dichotomous outcomes was assessed using the odds ratio calculated using a random effects model since the trials were expected to be heterogeneous in their design, patient populations and interventions. Outcomes were recoded in an ordered categorical manner where appropriate data were published:

- Three level stroke (fatal stroke / non-fatal stroke / no stroke)
- Four level stroke/TIA (fatal stroke / non-fatal stroke / TIA / no stroke)
- Three level MI (fatal MI / non-fatal MI / no MI)
- Four level MI/UA (fatal MI / non-fatal MI / unstable angina / no MI)
- Three level PE (fatal PE / non-fatal PE / no PE)

Insufficient data were available to do this for DVT and VTE. These ordinal outcomes were assessed using ordinal logistic regression adjusted for trial. Data were analysed using Stata (version 8).

7.4.5 Results

Table 7.7 shows the results for all outcomes. The control event rate is given to provide information on the background risk of each event; the changes in risk associated with treatment are therefore quantifiable. HRT increased the odds of having any CVD event by 24% (Figure 7.8), and stroke by 32%. Non fatal stroke was increased by 28%; both TIA and fatal stroke showed a trend towards increased odds of having an event with HRT although the statistical power for TIA was limited owing to the small number of events. No relationship was seen between HRT and CHD events, including MI. Those taking HRT had a two-fold increase risk of VTE, this comprising increases in DVT (97%) and PE (74%). Taking all outcomes together in a single analysis, HRT significantly increased a person's odds of having any thrombotic event by 23%. No statistical heterogeneity was found for any outcome apart for overall thrombotic events.

For ordered categorical data, a statistically significant result was seen for stroke severity when assessed as fatal stroke, non-fatal stroke, and no stroke (Table 7.8). The odds ratio of 1.31 (95% confidence interval 1.12 - 1.54) signifies that HRT treatment is associated with a shift to increased stroke severity. Ordinal regression requires the assumption of 'proportionality of odds' to be adhered to and this was present in all of the trials with more than two levels of data. Non-significant trends towards increased severity were seen for stroke/TIA assessed at four levels, and three level PE; both of these assessments suffered from limited published data on event severity thereby restricting the statistical power

of these analyses. No significant difference was seen for three level or four level MI, and no data were available for DVT, and VTE.

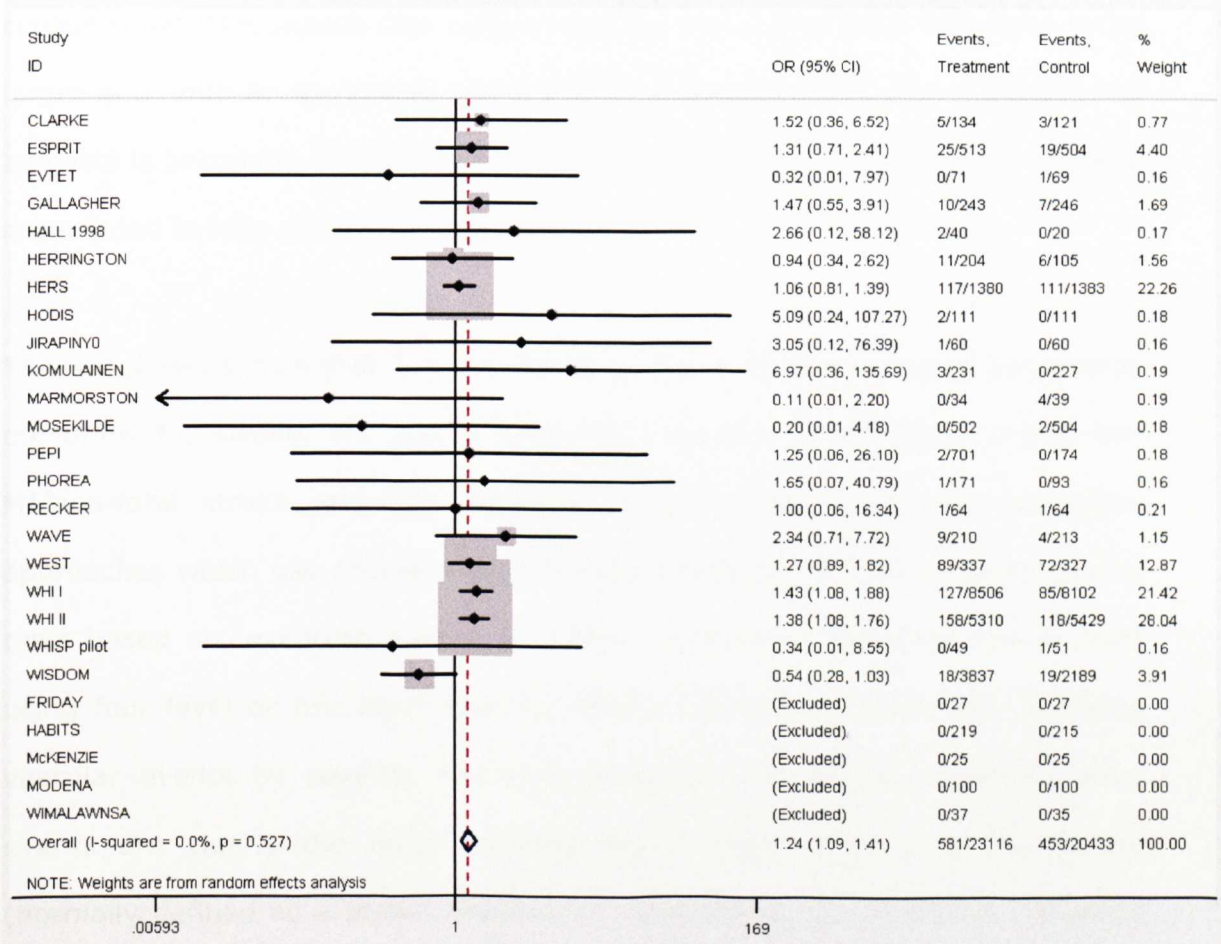
7.4.6 Conclusion

This meta analysis extends the findings of previous trials and meta analyses of HRT with the additional of ordinal regression analysis to assess the effect of HRT on severity. In summary, HRT is associated with increased CVD, stroke and stroke severity, VTE, and its components DVT and PE. In contrast, CHD rates are not increased.

HRT was found to increase the rate of total CVD by 24%. Ordering the severity of stroke by vital status (fatal stroke/non-fatal stroke/no stroke) allowed an ordinal meta analysis to be performed; HRT increased stroke severity by 31%. Since the assumption of proportionality of odds was adhered to in all of the trials reporting more than two levels (and trials which do not adhere to this would tend to attenuate any treatment effect), this finding of increased severity is likely to be genuine. This finding of increased severity is supported by a trend towards more fatal strokes in patients receiving HRT using standard dichotomous analysis (although this analysis is underpowered because of the limited number of events).

FIGURE 7.8

Forest plot of the effect of HRT on cerebrovascular disease.



7.5 DISCUSSION

Improvements in secondary prevention are leading to falling event rates in clinical trials. This means that future vascular prevention trials will need to be larger and, with an increasing number of new interventions, the availability of subjects is becoming limited. Thus, new approaches to trial design and analysis are needed to help reduce sample size.

This study has shown that it is feasible to create three level ordered categorical outcomes for stroke, MI, and a composite vascular event (fatal stroke and MI/non-fatal stroke and MI). Analysis reveals that, in general, statistical approaches which use ordinal data are more efficient than conventional binary tests based on 'event/no event'. A further increase in efficiency comes from using four level or five level data for stroke (with or without TIA). Ordering vascular events by severity has both biological and clinical meaning. Fatal events are clearly the most extreme health state while a severe stroke (normally defined as a stroke resulting in dependency on others) is a disaster for the patient, their carer and society, both for clinical and economic reasons. A mild stroke leaves the patient independent, even if residual impairment remains, and those who are younger can often return to work.

The most efficient statistical tests were those which examine ordinal data, including ordinal logistic regression, the Wilcoxon test, and bootstrapping the mean rank. In addition to improving statistical efficiency, the use of ordered categorical outcomes gives information on the ability of an intervention to reduce or increase the severity of an event, not just the number of events. This was demonstrated in the HRT meta analysis, where HRT not only increases the risk of stroke but also the severity of the event, with those taking HRT being more likely to have a fatal stroke compared rather than a non fatal stroke.

Ordinal logistic regression allows both estimation (with confidence intervals) and inclusion of baseline prognostic covariates in analyses. However, it assumes that any treatment effect is similar across outcome levels, i.e. the odds of moving a treated patient from fatal to severe non-fatal stroke are similar to those for moving from TIA to no event ('proportionality of odds'). This assumption requires justification since it is neither widely recognised nor obvious in most published vascular trial data. Firstly, it is biologically plausible to suggest that prophylactic interventions will reduce severity as well as the total number of events. Since the development of atherosclerosis and increases in thrombosis, coagulation and inflammation are not binary events in nature, and their magnitude is a determinant of the severity of clinical vascular events, it is reasonable to expect that interventions will move patients from fatal to severe, severe to mild, and mild to no events. If this assumption (of proportional odds) is not met, an alternative ordinal model could be considered (Stokes et al., 1995).

Secondly, there is existing published evidence that interventions do alter severity:

- Simvastatin reduces the risk of stroke of different severities by similar risk reductions (Heart Protection Study Collaborative Group, 2002)
- HRT increases both stroke and its severity (Sare et al., 2008)
- Antiplatelet agents reduce both fatal and non-fatal vascular events (Antithrombotic Trialists Collaboration, 2002)

The apparent failure of most vascular prevention trials to show individual effects on death or severe events is largely because they were not powered to assess these specific and, therefore, relatively uncommon events. Thirdly, the odds reduction at each outcome level appeared to be relatively constant when

individual trials were assessed (Figure 7.7); formal statistical assessment using the likelihood ratio test indicated that 'proportionality of odds' was present in most cases (although this test is known to be conservative) (Table 7.6). Lastly, using ordinal statistical tests was more powerful than binary approaches, the central finding of the OAST prevention study. Although this is not a novel idea in the statistical community, ordinal outcomes have not been applied to vascular prevention trials in the past.

Another efficient ordinal test is the Wilcoxon test which is widely available in statistical packages and can produce a point estimate (median difference between groups) with confidence intervals. The major assumption of the test is that the treatment groups should be independent and this is met here. The final efficient statistical approach was bootstrapping the mean rank; this approach is computer intensive and its application and the interpretation of results are not well appreciated by clinicians, although it is free of assumptions (Efron and Tibshirani, 1993).

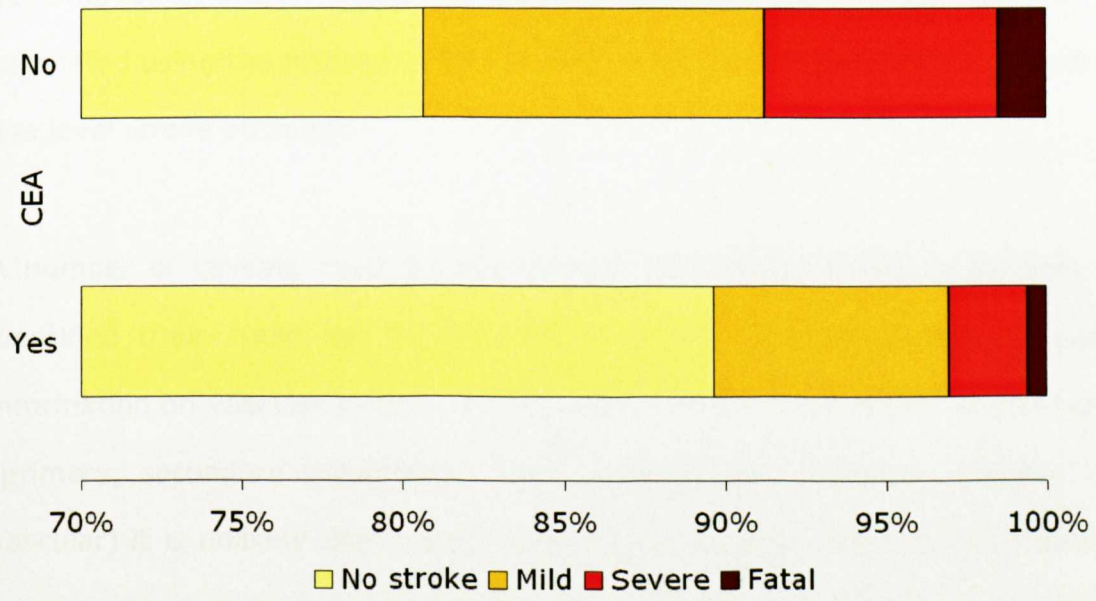
The conventional approach to analysing vascular prevention trials is to perform time to event analyses, as visualised using Kaplan-Meier curves, and analysed with Cox regression. When the frequency of events is high, analyses based on time-to-event are more efficient than those using frequencies (as analysed using logistic regression) (Vittinghoff and McCulloch, 2006). However, the frequency of vascular events in most primary and secondary prevention trials running over three to five years is relatively low; recent vascular prevention trials have tended to report annualised stroke rates of 2-4% (Bhatt et al., 2006, The Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) Investigators, 2006). Logistic and Cox models give similar results when the overall event frequency is less than 10% (Ingram and Kleinman,

1989, Annesi et al., 1989). Where the frequency of events is higher, ordinal data may be analysed using ordinal time to event analyses (Berridge and Whitehead, 1991). In the current data set, the Cox model was slightly less efficient than bootstrapping, the Wilcoxon test and ordinal logistic regression.

Using ordered categorical data means that results will need to be reported differently to those obtained from binary analyses. The results of binary tests are summarised easily as the proportion of patients who benefit (or suffer) with a treatment, i.e. oral anticoagulation reduced absolute stroke recurrence by 1.46% (odds ratio 0.75, p=0.036) in the ASPECT trial (Anticoagulants in the Secondary Prevention of Events in Coronary Thrombosis (ASPECT) Research Group, 1994). In contrast, ordinal tests will need to be presented as the average absolute improvement in outcome, e.g. anticoagulation reduced stroke recurrence and its severity with an odds ratio of 0.60 (or reduced the mean severity by 0.5 points, p=0.013) on a five level scale. In this respect, health consumers will need to decide what odds ratio or difference in events is worthwhile, both clinically and in terms of health economics. In reality, it is reasonable to present the primary result using the odds ratio (or median change in event severity) and to give the absolute percentage change calculated from the binary outcome as a secondary measure. Further, a visual presentation of the data can be displayed as the percentage of patients within each category by treatment group (as shown in Figure 7.9).

FIGURE 7.9

Example four-level ordinal data from the North American Symptomatic Carotid Endarterectomy Trial (NASCET) of carotid endarterectomy (CEA). Note that CEA moves each polytomous level to the right. Statistical comparisons of binary (stroke /no stroke), p=0.002; trichotomous (fatal stroke /non-fatal stroke /no stroke), p=0.001; and quadrotomous (p=0.0009) data. Note, 70% of patients with no events are not shown to emphasise those who had an event (North American Symptomatic Carotid Endarterectomy Trial Collaborators, 1991).



Just as sample size calculations exist for trials using dichotomised analyses, analogous approaches exist for ordinal tests. Since ordinal analyses are statistically more powerful than dichotomous ones, trial size may be reduced for a given power of say 90% e.g. sample size falls by 15-24% as the number of outcome categories increases from three to seven (Whitehead, 1993). This reduction is worthwhile and would reduce competition between trials for patients, and lower trial costs and complexity. Taking the HEP trial (Coope and Warrender, 1986) as an example (and assuming significance=0.05 and power=0.9), the sample size is reduced by 48% from 1,556 for a binary outcome of stroke/no stroke to 810 for a three level stroke outcome as calculated using the method of Whitehead; this is further reduced to 772 with a five level stroke outcome.

A number of caveats must be made about this study. Firstly, a majority of identified trials could not be included since they did not publish adequate information on vascular events. As data were missing for a variety of trial types (primary, secondary prevention), sizes, and outcome measures (stroke, MI, vascular) it is unlikely that a systematic bias was introduced into the findings; however, the precision of the results will have been attenuated by the missing data. Future trial publications should give this information, including vital status for the main vascular outcomes, so that ordered outcome categories can be calculated. Secondly, not use all possible statistical tests relevant to the problem of analysing ordered categorical data were used; instead, the focus was concentrated on those approaches which are readily available in statistical textbooks (Siegel and Castellan, 1988) and computer packages.

The HRT meta analysis shows the first example of an ordinal analysis being applied to vascular prevention data. The ordinal analysis added novel information on the effect of HRT on the severity of stroke suffered.

7.6 SUMMARY

These results show that vascular prevention trials should consider employing statistical approaches which use the inherent ordered categorical data present within vascular outcome events. The resulting trials could be smaller (with savings in patient numbers, numbers of centres, and study cost and complexity) and would allow appreciation of the effect of interventions on severity, as well as absolute number of events, to be highlighted. Appropriate tests include the Wilcoxon test, ordinal logistic regression, and bootstrapping the mean rank.

TABLE 7.1

Assessment of ten statistical approaches for analysing stroke as a three level stroke outcome (fatal/non-fatal/no stroke) in 85 vascular prevention trials. Analysis by two way ANOVA ($p<0.0001$) on the ranked data (1-10 with 1 'best'); comparison of tests by Duncan's multiple range test - those tests joined by the same band are not significantly different from each other at $p<0.05$.

Test	Mean rank	Banding
Wilcoxon test	3.32	
Bootstrap (difference in mean rank)	3.32	
Ordinal logistic regression	4.12	
Robust rank test	4.51	
Cochran-Armitage trend test	4.80	
t-test	5.08	
Chi Sq – 2x3 test	5.94	
Chi Sq – stroke vs. no stroke	6.37	
Chi Sq – death vs. alive	7.58	
Median test	9.97	

TABLE 7.2

Ranking of statistical tests (1-10 with 1 'best') for measures of stroke (three, four, and five levels), myocardial infarction (three level), and composite vascular outcome (three level). The most efficient tests are highlighted and do not differ from each other statistically.

Outcome	Trials	P value	Ranking of tests relative to each other									
			WIL	BS	OLR	RRT	CAT	t-test	X^2 2x3	X^2 Event	X^2 Dead	Median test
Fatal stroke/non-fatal stroke/no stroke	85	<0.0001	1	2	3	4	5	6	7	8	9	10
Fatal stroke/severe non-fatal/mild/no stroke	21	<0.0001	2	1	4	3	5	6	8	7	9	10
Fatal stroke/non-fatal stroke/TIA/no stroke	29	<0.0001	2	1	5	6	3	4	7	8	9	10
Fatal stroke/severe non-fatal stroke/mild stroke/TIA/no stroke	11	<0.0001	3	4	5	6	1	2	8	7	9	10
Fatal MI/non-fatal MI/no MI	58	<0.0001	1	3	5	6	2	4	7	8	9	10
Fatal vascular event/non-fatal vascular event/no vascular event	43	<0.0001	1	2	3	4	5	6	7	8	9	10

BS: bootstrap; CAT: Cochran-Armitage test; WIL: Wilcoxon test; OLR: ordinal logistic regression; RRT: robust rank test

TABLE 7.3

Comparison of effects of treatment on stroke using Chi-square (two level), and Wilcoxon test and ordinal logistic regression (three, four, and five levels) for the six trials where data was available. The data given are p values. The most significant result is highlighted.

Stroke	Test	PPP	SPAF 1	SPAF 2	HEP	ASPECT	BAATAF
Two level (event/no event)	Chi-square test	0.293	0.029	0.170	0.033	0.036	0.008
Three level (fatal/non-fatal/no event)	Wilcoxon test	0.224	0.022	0.129	0.021	0.030	0.003
	Ordinal logistic regression	0.227	0.024	0.131	0.023	0.031	0.011
Four level (fatal/severe/mild/no event)	Wilcoxon test	0.224	0.023	0.129	0.021	0.030	0.003
	Ordinal logistic regression	0.227	0.025	0.130	0.022	0.030	0.011
Four level (fatal/non-fatal/TIA/no event)	Wilcoxon test	0.064	0.008	0.061	0.016	0.013	0.005
	Ordinal logistic regression	0.065	0.008	0.062	0.018	0.014	0.009
Five level (fatal/severe/mild/TIA/no event)	Wilcoxon test	0.064	0.008	0.060	0.016	0.013	0.005
	Ordinal logistic regression	0.065	0.008	0.062	0.017	0.014	0.009

TABLE 7.4

Ranking of statistical tests (1-10 with 1 'best') for three level stroke (fatal, non-fatal, no stroke) in subgroups of vascular prevention trials. The most efficient tests are highlighted and do not differ from each other statistically.

Outcome	Trials	P value	Ranking of tests relative to each other									
			WIL	BS	OLR	RRT	CAT	t-test	X^2 2x3	X^2 Event	X^2 Dead	Median test
Prevention, primary	29	<0.0001	1	2	5	3	6	4	7	8	9	10
Prevention, secondary	56	<0.0001	2	1	3	4	5	6	7	8	9	10
Anticoagulants	12	<0.0001	2	1	7	6	4	3	5	8	9	10
Antiplatelets	33	<0.0001	1	2	3	5	4	6	7	8	9	10
Antihypertensives	23	<0.0001	2	1	3	4	5	6	7	8	9	10
Lipid lowering	10	<0.0001	2	1	3	4	5	6	8	7	9	10
Carotid endarterectomy	4	<0.0001	2	1	5	3	6	4	8	7	9	10
Hormone replacement therapy	2	0.86	-	-	-	-	-	-	-	-	-	-
Age<65 years	34	<0.0001	1	2	3	4	5	6	7	8	9	10
Age >65 years	31	<0.0001	2	1	4	3	6	5	7	8	9	10

Trial, small (n<2,520)	42	<0.0001	1	2	5	6	3	4	7	8	9	10
Trials, large (n>2,520)	42	<0.0001	2	1	3	4	5	6	8	7	9	10
Follow-up, short term (<36 months)	45	<0.0001	1	2	4	5	3	6	7	8	9	10
Follow-up, long term (>36 months)	39	<0.0001	2	1	3	4	5	6	7	8	9	10
Risk of death in control, low (<0.2% per month)	43	<0.0001	1	2	3	4	5	6	7	8	9	10
Risk of death in control, high (>0.2% per month)	41	<0.0001	2	1	3	4	5	6	7	8	9	10
Risk of stroke in control, low (<0.17% per month)	40	<0.0001	1	2	3	4	5	6	7	8	9	10
Risk of stroke in control, high (>0.17% per month)	41	<0.0001	2	1	5	6	3	4	7	8	9	10
Time from index event, short (<87 days)	22	<0.0001	1	2	3	4	5	6	8	7	9	10
Time from index event, long (>87 days)	22	<0.0001	2	1	3	6	4	5	7	8	9	10

BS: bootstrap; CAT: Cochran-Armitage test; WIL: Wilcoxon test; OLR: ordinal logistic regression; RRT: robust rank test

TABLE 7.5

Assessment of ten statistical approaches for analysing stroke as a three level stroke outcome (fatal/non-fatal/no stroke) with the hazard ratio extracted from the trial publication in 36 vascular prevention trials. Analysis by two way ANOVA ($p<0.0001$) on the ranked data (1-10 with 1 'best'); comparison of tests by Duncan's multiple range test - those tests joined by the same band are not significantly different from each other at $p<0.05$.

Test	Mean rank	Banding
Bootstrap (difference in mean rank)	3.42	
Wilcoxon test	3.85	
Ordinal logistic regression	4.46	
Hazard ratio	4.75	
Robust rank test	5.26	
Cochran-Armitage trend test	5.43	
t-test	5.68	
Chi Sq – 2x3 test	6.72	
Chi Sq – stroke vs. no stroke	6.86	
Chi Sq – death vs. alive	8.61	
Median test	10.96	

TABLE 7.6

Assessment of the type 1 error rate for the Wilcoxon test and ordinal logistic regression using data from five trials for the three level stroke outcome. The data given are the number and percentage of statistically significant results found from 1,000 simulations.

Trial	Wilcoxon test		Ordinal logistic regression	
	n significant	% significant	n significant	% significant
SPAF 2	21	2.1	47	4.7
ESPS 2	30	3.0	21	2.1
HOPE	17	1.7	30	3.0
HPS	26	2.6	17	1.7
NASCET	18	1.8	54	5.4

TABLE 7.7

Effect of hormone replacement therapy on arterial and venous events; with odds ratio (95% confidence intervals) using random effects model.

	Trials	Subjects	Events	Control event rate (events per person/year)	Odds ratio (95% CI)	p	Heterogeneity p
Cerebrovascular disease	26	43,549	1,034	0.02	**1.24 (1.09 - 1.41)**	**0.001**	0.53
Stroke	18	36,523	741	0.02	**1.32 (1.14 - 1.53)**	**<0.0001**	0.87
Transient ischaemic attack	7	6,035	153	0.03	1.05 (0.76 - 1.45)	0.78	0.53
Fatal stroke	11	32,935	105	0.003	1.35 (0.89 – 2.03)	0.16	0.39
Non-fatal stroke	10	32,680	581	0.02	**1.28 (1.08 – 1.52)**	**0.004**	0.58
Coronary heart disease	25	43,159	1,636	0.04	1.00 (0.90 – 1.11)	0.97	0.56
Myocardial infarction	21	41,849	1,238	0.03	1.02 (0.91 – 1.15)	0.70	0.78
Fatal MI	15	40,319	396	0.01	1.03 (0.84 – 1.26)	0.77	0.49
Non-fatal MI	15	40,319	846	0.02	1.02 (0.88 – 1.18)	0.77	0.41
Unstable angina	5	9,413	360	0.04	0.97 (0.71-1.40)	0.98	0.23
Venous thromboembolism	22	42,381	547	0.02	**2.05 (1.44 – 2.92)**	**<0.0001**	0.07
Deep vein thrombosis	16	40,417	376	0.01	**1.97 (1.58 – 2.46)**	**<0.0001**	0.58
Pulmonary embolism	12	39,612	230	0.004	**1.74 (1.32 – 2.30)**	**<0.0001**	0.66
All thrombotic events	31	44,113	3,217	0.08	**1.23 (1.07 – 1.41)**	**0.004**	0.06

TABLE 7.8

Effect of hormone replacement therapy on the severity of arterial and venous events; by ordinal logistic regression.

Outcome	Trials	Subjects	Ordinal outcome	Odds ratio	95% confidence interval	P value
Three level stroke (fatal stroke / non-fatal / no stroke)	10	32,679	104 / 581 / 31,997	**1.31**	**1.12 - 1.54**	**0.001**
Four level stroke/TIA (fatal stroke / non-fatal stroke / TIA / no stroke)	4	12,440	57 / 291 / 159 / 11,933	1.10	0.91 - 1.33	0.34
Three level MI (fatal MI / non-fatal MI / no MI)	15	40,252	396 / 846 / 39,010	1.04	0.93 - 1.17	0.49
Four level MI/UA (fatal MI / non-fatal MI / unstable angina / no MI)	5	7765	140 / 248 / 360 / 7,017	1.00	0.85 – 1.17	0.96
Three level PE (fatal PE / non-fatal PE / no PE)	3	7,527	4 / 13 / 7,510	2.57	0.73 – 9.01	0.14

CHAPTER 8

DISCUSSION

PUBLICATIONS/PRESENTATIONS CONTRIBUTING TO THIS CHAPTER

Bath P.M.W, **Gray L.J** (2009) Systematic Reviews as a Tool for Planning and Interpreting Trials *International Journal of Stroke. January 2009.*

8.1 INTRODUCTION

The results from stroke trials have greatly improved the treatment and care, and therefore outcome, of patients who have suffered from a stroke. Stroke units can be used to treat all types of stroke, and combine the skills of a multi-disciplinary team of therapists and clinicians. Aspirin has wide utility, but limited efficacy in ischaemic stroke, while thrombolytic therapy has high efficacy, but with limited usage. Hence, treatment options remain limited for those with stroke, especially for those who have suffered from a haemorrhagic stroke.

Although there have been successes in stroke research, there have also been many failures. For over two decades trials have been assessing neuroprotective agents, treatments which aim to protect brain tissue from cell death, with no success (Kidwell et al., 2001). Many factors have been suggested as reasons for this, including the applicability of animal findings to humans, and trial design and analysis (Rother, 2008). Although the recent SAINT trials were reported to be the "perfect" trial, with animal data fulfilling all of the STAIR criteria, a primary outcome which took into account baseline severity, an early time window, and the allowance of thrombolysis (Lees et al., 2006), NXY-059 was still shown to be ineffective in a second phase three trial (Shuaib et al., 2007). Research is now being carried out to try and find out why such promising initial results in both animal and man lead to the ultimate failure of the phase three trial (Bath et al., 2008). This specific example highlights the need for further research, such as the OAST project, to try and improve aspects of the design and analysis of stroke clinical trials

OAST is the largest data pooling project, to date, in stroke to look at improving the statistical analysis of stroke trials. Previous research had focussed on re-

analysing data from one trial or using simulated artificial data to describe effects. The quirks and complexity of data from stroke trials means that using 'real-life' data from many situations is beneficial. Also, other studies have focussed on only acute trials, whereas this project includes data on not only acute interventions, but also stroke unit trials and those assessing occupational therapy. This section will discuss the main findings of both the acute and prevention projects, reflect on these, and suggest places for further work.

8.2 OAST PROJECT

The acute OAST project gathered individual patient data on over 50,000 patients from 47 completed trials. Re-analysis of these trials with various statistical methods revealed that many stroke trials have been using sub optimal methods for analysing data from functional outcome scales, with the most powerful methods of analysis being: ordinal logistic regression, the t-test, the robust rank test, bootstrapping the difference in mean rank, and the Wilcoxon test. All of these tests take into account the inherent ordering of functional outcome data, whereas traditional methods of analysis, such as the chi square test, lump these categories together to create two or more groups ignoring any ordering. The assessment of sample size showed that by changing to an ordinal method of analysis, trialists could reduce the sample size needed for a given power by 28%. This saving could also be transferred into greater statistical power to find a difference between treatments for a given sample size.

The assessment of sample size showed an interesting finding, with ordinal methods not performing as well in trials of thrombolytic agents, where the lack of proportional odds means that dichotomous outcomes may be more appropriate.

Although finding that ordinal methods are more statistically powerful than those which dichotomise is not surprising or novel in the statistical community, the novelty of this work is in the application to stroke data. Very few stroke trials to date have used an ordinal method of analysis for their primary outcome, and although statistical analysis is receiving more interest in the field of stroke, most studies still choose their method of analysis on hearsay or the results of previous trials. The OAST acute project is a rigorous and thorough examination of the available methods of analysis and the results can therefore be used reliably in future trials.

The next part of the project assessed the impact of taking into account covariates on the sample size required. Adjusting ordinal logistic regression for three prognostic factors (age, sex and severity) can further reduce the sample size needed by around 37%. This part of the project used less data than the preceding analyses, as data was required not only from the primary outcome but baseline variables as well. The initial analysis showed that ordinal methods are not suitable for trials of thrombolytic agents, so trials testing these agents were also excluded. Given the smaller number of trials included, simulation was used to examine the effect of adjustment for covariates on sample size. Using simulations allowed the comparison of three different levels of treatment effect and used the actual covariate structure of those patients in the included trials.

When assessing a global outcome, the Cochran Mantel-Haenszel test, and a patient specific outcome, no difference between these and either the t-test or ordinal logistic regression were found. This may, in part, be due to the low number of trials included in this part of the analysis. It may also be argued that it is not valid to compare outcomes which combine more than one scale with an analysis based on only one scale.

Several comments can be made about the OAST acute project. First, it aimed to include data from all stroke trials assessing a beneficial or harmful intervention. Unfortunately, data were not made available for all identified trials; where possible, individual data from publications that provided patient numbers by outcome score, were created. Data were missing for a variety of trial types (acute/rehabilitation/stroke unit) and sizes, and functional outcome measure (mRS/BI), so it is unlikely that a systematic bias was introduced into the findings. However, the precision of the results may have been attenuated by the missing trials. It is important that data from completed trials are shared with data pooling projects such as OAST or the Virtual International Stroke Trials Archive (VISTA) (Bath and Gray, 2008, Ali et al., 2007). Unlike OAST, VISTA collates data from only the control arms of completed trials. Second, the OAST project only included data from trials of 'beneficial' or 'hazardous' treatment as shown with an individual trial or as part of a meta analysis. The rationale for this is that re-analysing data for interventions known not to have an effect on outcome, looking for more statistically significant findings, could be perceived as data dredging. Theoretically all of the included trials should have shown a beneficial/hazardous outcome if they had been powered correctly and analysed in an appropriate manner.

Overall, this part of the OAST project has shown that improvements can be made to the statistical analysis of functional outcome data in stroke trials. Where distributions meet the proportional odds assumption, i.e. they exert a similar treatment effect across all levels of the scale, it is suggested that trialists use ordinal logistic regression. Using a modelling approach of analysis also allows adjustment for prognostic factors. Where the proportionality of odds assumption is not met, i.e. with interventions such as thrombolytic therapy, trialists can consider other methods which assess treatments across the whole

functional outcome scale, such as the t-test, robust rank test, bootstrapping the difference in mean rank or the Wilcoxon test.

8.2.1 Efficacy of Nitric Oxide in Stroke trial

The results of the OAST project are being used to improve the statistical analysis of the ongoing 'Efficacy of Nitric Oxide in Stroke' (ENOS) trial. The ENOS trial is a factorial randomised phase three trial comparing the efficacy of transdermal glyceryl trinitrate against control, and stopping or continuing pre stroke antihypertensive therapy (The ENOS Trial Investigators, 2006). The initial primary outcome of the trial was a dichotomised death or dependent versus independent on the mRS, cut at two (0-2 vs. 3-6). On the basis of the OAST project, the trial steering committee in April 2008 decided to change this to an analysis of data across the whole mRS scale using ordinal logistic regression and to adjust this for age, sex and baseline severity. The committee decided to retain the planned sample size of 5,000 but to increase the statistical power for finding a treatment difference.

8.2.2 Extensions to the OAST project

There are still many unanswered questions around the analysis of stroke trials and therefore there are many ways this project could be built upon.

The global outcome, patient specific outcome and Cochran Mantel-Haenszel test assessed here were all based on dichotomous data. Even though taking into account baseline severity or merging more than one scale may be beneficial, there are still the inherent problems of defining where scales should be dichotomised and the loss of information associated with collapsing data into groups. Future work could look at developing these outcomes to take into account the ordinal nature of functional outcome data. The global outcome

could be extended to ordinal outcomes by either using an ordinal GEE model (Lumley, 1996) or by using a multivariate t-test, such as Hotelling's t-test, which compares by treatment group correlated data from two or more continuous or ordinal scales (Hotelling, 1931). The patient specific outcome currently uses a chi square test to analyse dichotomous data. Future analysis could look at using a test for trend to take into account the ordering of this data and using more than two categories for collapsing the data. For example, comparing those who are independent versus mildly dependent versus severely dependent versus dead, instead of the binary outcome, independent versus dead or dependent. Research would need to focus on creating well defined and valid categories for various levels of baseline severity. The Cochran Mantel-Haenszel test assessed in Chapter 6 stratifies by collapsed baseline severity. It may be preferable here to use ordinal logistic regression instead, with adjustment for severity.

If trialists decide to use an ordinal approach it is important to consider how the number needed to treat would be calculated. Methods have been developed but these are based on the within patient correlation and therefore require paired data. Cross-over trials are rare in stroke research and therefore it is difficult to calculate an estimate of the within patient correlation (Walter, 2001). Saver has begun to address this problem by using a panel of experts to independently specify a joint distribution, based on the NINDS tPA trial, for samples of 100 patients assigned to placebo and active treatment, and uses these joint distributions to estimate the within patient correlation (Saver, 2004). Development of a method which removed the need for independent experts would save money and time and allow all trialists to present this important data in the trial manuscript to aid interpretation of the results. A possible approach to this could involve creating matched data from a completed trial and using

this to estimate the within patient correlation, i.e. taking a patient from each treatment group who share similar characteristics (e.g. age, sex and baseline severity) and compare their outcomes.

The OAST project is promoting the use of ordinal analyses to stroke trialists. While applying ordinal methods improves the analysis of stroke trials, this will complicate the ability of future researchers to carry out meta analyses. Trials which use binary outcomes will normally present the number and percentage of patients who fall into each category by treatment group. These numbers can be easily extracted and added to a binary meta analysis. Trialists will need to display the number and percentage of patients falling into each category on the scale being used to allow ordinal meta analyses (see example from the FOOD 3 trial).

TABLE 8.1

mRS score, primary outcome, and death taken from the FOOD 3 trial manuscript.

Modified Rankin Scale	Early tube (n=429)	Avoid tube (n=430)	PEG tube (n=162)	Nasogastric tube (n=159)
0	4 (1%)	9 (2%)	2 (1%)	1 (1%)
1	10 (2%)	16 (4%)	0	3 (2%)
2	26 (6%)	19 (4%)	7 (4%)	6 (4%)
3	50 (12%)	41 (10%)	9 (6%)	20 (13%)
4	53 (12%)	42 (10%)	8 (5%)	12 (8%)
5	104 (24%)	95 (22%)	57 (35%)	41 (26%)
Dead	182 (42%)	207 (48%)	79 (49%)	76 (48%)
Unknown	0	1 (<1%)	0	0
MRS 0-3	90 (21%)	85 (20%)	18 (11%)	30 (19%)
MRS 4-5	157 (37%)	137 (32%)	65 (40%)	53 (33%)
Dead or MRS 4-5	339 (79%)	344 (80%)	144 (89%)	129 (81%)

Reprinted from The Lancet, 365. Dennis M et al. Effect of timing and method of enteral tube feeding for dysphagic stroke patients (FOOD): a multicentre randomised controlled trial. 764-72, Copyright (2005), with permission from Elsevier.

Ordinal meta analysis methods are only available for individual patient data (IPD) and not for combining summary ordinal data (Whitehead et al., 2001). If data are presented as in Table 8.1, IPD can be formed for the primary outcome by treatment and then combined using IPD methods. Future work could look at combining the odds ratios from ordinal logistic regression so that summary meta analyses can also be performed.

8.3 OAST PREVENTION PROJECT

The OAST prevention project aimed to improve the analysis of vascular prevention trials. The acute OAST project showed that employing an ordinal approach to analysis could improve statistical power and this idea was used to create ordinal categories for analysis from vascular prevention data. The results showed that creating ordinal categories from binary outcome data and using an approach which looks for changes across these, improved the statistical power to find a treatment effect.

Akin to the acute OAST project not all vascular prevention trials showing benefit or harm either individually or in a meta analysis were included. This was because data was extracted from the trial publication and this was only possible if outcome data by treatment group had been presented for the categories of interest. For example, if the number of fatal and non fatal stroke had not been presented separately, the data could not be included. Of the 151 studies excluded, 128 (85%) did not provide adequate outcome data (see (Bath et al., 2008)). This is in part due to the types of analyses routinely used in prevention trials, for example, if the primary outcome is a composite event, there may not be data on the individual events. Although these missing trials will have reduced the statistical power, the missing trials were from a wide range of trial types (different treatments, primary and secondary trials, smaller and larger

trials etc) so it is assumed that their exclusion will not have induced a systematic bias.

The ordinal event analyses shown here do not take into account the time to the event. It is argued that time to event analyses are more powerful than those based on event counts, although it was shown that the Wilcoxon test and ordinal logistic regression produced similar results to the Cox model. This analysis was only based on 36 trials where hazard ratios and their p value had been presented in the trial manuscript.

8.3.1 Extensions to the OAST prevention project

Further work is still to be done in this area. Firstly, as with the acute OAST project, the effect on sample size could be assessed. The reduction in sample size gained from ordinal analysis is probably more meaningful to trialists carrying out prevention trials and can be converted easily into savings in terms of trial costs, the number of centres required, length of follow up needed etc. In the HEP trial given as an example in the discussion of Chapter 7, a 48% reduction in sample size was seen when changing from a binary to a three level stroke outcome. A similar approach could be used here as carried out in Chapter 4 for the acute project. Secondly, a criticism of the ordinal prevention outcomes is that they do not take into account the timing of the event. Many prevention trials analyse their primary outcome with a time to event analysis, such as the Cox proportional hazards model. There are ordinal survival models described in the literature (Berridge and Whitehead, 1991), and an extension to this work could look at combining the ordinal prevention outcomes with the time of the event. Finally, reflecting the acute OAST project adjustment for prognostic covariates could also be considered. Each of the extensions

discussed above require IPD, and therefore IPD would need to be sought from the Chief Investigators of each trial.

Ordinal methods of analysis are yet to be applied to an actual prevention trial, but have been used in a meta analysis of HRT (Sare et al., 2008).

8.4 OTHER AREAS OF RESEARCH

When considering how the OAST project could be used in other areas of research, there are two main considerations:

1. Other areas in stroke, apart from functional outcome and prevention, which use ordinal scales
2. Other clinical areas, apart from stroke, which use ordinal scales

In stroke research, scales are used to measure many aspects of recovery. The main aspect is functional outcome, which is usually the primary outcome in large phase three clinical trials. However, secondary outcomes also use scales to measure domains of interest, such as quality of life, mood and cognition. Although these are usually secondary outcomes, many agree that these are perhaps the most important outcomes to the patient and therefore novel treatments which show an effect in these more patient centred outcomes could still be clinically beneficial. Akin to functional outcome scales, other scale data is also routinely dichotomised when analysed. It might not be possible to simply apply the results from the functional outcome data to those other domains of recovery. As shown earlier, functional outcome after stroke has a 'U' shaped distribution with around a third of patients dying post stroke, a third returning to full independence and the remaining third being distributed across the scale. Data on domains such as mood and cognition may not follow the same

distribution. Data on quality of life has been shown to be highly related to functional outcome (Gray et al., 2007) and therefore the results presented here may translate and be useful when analysing quality of life data. However, unlike functional outcome scales some quality of life scales, such as the EuroQol (Brooks and with the EuroQol Group, 1996), are linear and therefore ordinal methods, which do not assume linearity, may not be optimal. Analysis of the quality of life data gained from the Short-form 36 has been assessed previously and has been described in the introduction of this thesis (Walters et al., 2001).

To date, no studies have looked at the optimal ways of analysing data from mood or cognition scales. To do this IPD from trials which have assessed quality of life, mood or cognition could be collated and the methods of the OAST acute project repeated to look for the optimal method of analysis.

Other areas of research have also reported problems in the analysis of ordinal data from clinical trials and have taken steps to rectify these. One example is traumatic brain injury, where a group similar to OAST, have looked at some issues with the analysis of the Glasgow Outcome Scale. In line with the OAST findings they reported that statistical power could be increased by using a shift analysis. They have also assessed patient specific outcomes and adjustment of logistic regression analysis.

There may be other areas which could also benefit from the results of the OAST project. For example, problems with the presentation and analysis of ordinal data have also been described in veterinary dermatology (Plant et al., 2007), rheumatology (Lavalley and Felson, 2002), and in nursing literature (Jakobsson, 2004).

8.5 SUMMARY

In summary, the OAST project has shown that many stroke trials have used sub-optimal methods of analysis and this could be a contributing factor to why so many stroke trials have found neutral results. This project has shown that functional outcome scales should always be analysed in a way that retains the ordinal nature of the data. This not only provides greater statistical power and more information on the effect of the intervention, but can also be used to lower sample size and if a modelling approach is chosen to take into account covariates. Ordinal methods can be applied to both acute and prevention trials. Statistical power can be increased in prevention trials by turning binary event accounts into ordinal variables.

Changing the design and analysis of trials to improve statistical power gives new effective interventions the best possible chance of being used in everyday medicine, both reducing the number of people needed to be involved in clinical trials and possibly the actual number of trials needed.

REFERENCES

ABCIXIMAB EMERGENT STROKE TREATMENT TRIAL (ABESTT) INVESTIGATORS (2005) Emergency treatment of Abciximab for treatment of patients with acute ischemic stroke. Results of a randomized phase 2 trial. *Stroke,* 36, 880-890.

ADAMS, H. P., DAVIS, P. H., LEIRA, E. C., CHANG, K. C., BENDIXEN, B. H., CLARK, W. R., WOOLSON, R. F. & HANSEN, M. D. (1999) Baseline NIH stroke scale score strongly predicts outcome after stroke: A report of the trial of Org 10172 in acute stroke treatment (TOAST). *Neurology,* 53, 126-131.

ADAMS, H. P., EFFRON, M. B., TORNER, J., DAVALOS, A., FRAYNE, J., TEAL, P., LECLERC, J. R., OEMAR, B., PADGETT, L., BARNATHAN, E. S., HACKE, W. & FOR THE ABESTT-II INVESTIGATORS (2008) Emergency Administration of Abciximab for Treatment of Patients With Acute Ischemic Stroke: Results of an International Phase III Trial. *Stroke,* 39, 87-99.

ADAMS, H. P., LECLERC, J. R., BLUHMKI, E., CLARK, W., HANSEN, M. & HACKE, W. (2004) Measuring outcomes as a function of baseline severity of ischemic stroke. *Cerebrovascular Diseases,* 18, 124-129.

AGRESTI, A. (1984) *Analysis of ordinal categorical data,* New York, John Whiley & sons.

AGRESTI, A. (1999) Modelling ordered categorical data: Recent advances and future challenges. *Statistics in Medicine,* 18, 2191-2207.

AGRESTI, A. (2002) *Categorical Data Analysis (Second Edition)*, Wiley.

AITKEN, P. D., RODGERS, H., FRENCH, J. M., BATES, D. & JAMES, O. F. W. (1993) General medical or geriatric unit care for acute stroke? A controlled trial (Abstract). *Age Ageing.*

ALBERS, G. W., GOLDSTEIN, L. B., HALL, D., LESKO, L. M. & FOR THE APTIGANEL ACUTE STROKE INVESTIGATORS. (2001) Aptiganel hydrochloride in acute ischemic stroke. *Journal of the American Medical Association,* 286, 2673-2682.

ALI, M., BATH, P., CURRAM, J., DAVIS, S., DIENER, H. C., DONNAN, G., FISHER, M., GREGSON, B. A., GROTTA, J., HACKE, W., HENNERICI, M. G., HOMMEL, M., KASTE, M., MARLER, J. R., SACCO, R. L., TEAL, P., WAHLGREN, N. G., WARACH, S., WEIR, C. J. & LEES, K. R. (2007) The Virtual International Stroke Trials Archive (VISTA). *Stroke,* 38, 1905-1910.

ANNESI, I., MOREAU, T. & LELLOUCH, J. (1989) Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Statistics in Medicine,* 8, 1515-1521.

ANTIOCOAGULANTS IN THE SECONDARY PREVENTION OF EVENTS IN CORONARY THROMBOSIS (ASPECT) RESEARCH GROUP (1994) Effect of long-term oral anticoagulant treatment on mortality and cardiovascular morbidity after myocardial infarction. *Lancet,* 343, 499-503.

ANTITHROMBOTIC TRIALISTS' COLLABORATION (2002) Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. *British Medical Journal,* 324, 71-86.

ASSMANN, S. F., POCOCK, S. J., ENOS, L. E. & KASTEN, L. E. (2000) Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet,* 355, 1064-1069.

BAMFORD, J., SANDERCOCK, P., DENNIS, M., BURN, J. & WARLOW, C. (1990) A prospective study of acute cerebrovascular disease in the community: the Oxfordshire Community Stroke Project 1981-86. 2. Incidence, case fatality and overall outcome at one year of cerebral infarction, primary intracerebral and subarachnoid haemorrhage. *J Neurology Neurosurgery Psychiatry,* 53, 16-22.

BAMFORD, J., SANDERCOCK, P. A. G., DENNIS, M., BURN, J. & WARLOW, C. (1991) Classification and natural history of clinically identifiable subtypes of cerebral infarction. *Lancet,* 337, 1521-1526.

BAMFORD, J. M., SANDERCOCK, P. A. G., WARLOW, C. P. & SLATTERY, J. (1989) Interobserver agreement for the assessment of handicap in stroke patients. *Stroke,* 20, 828.

BARER, D. & NOURI, F. (1989) Measurement of activities of daily living. *Clinical Rehabilitation,* 3, 179-187.

BARER, D. H., CRUICKSHANK, J. M., EBRAHIM, S. B. & MITCHELL, J. R. (1988) Low dose beta blockade in acute stroke ("BEST" trial): an evaluation. *British Medical Journal,* 296, 737-741.

BATH, P. & GRAY, L. J. (2005) Association between hormone replacement therapy and subsequent stroke: a meta analysis. *British Medical Journal,* 330.

BATH, P. M. W., GEEGANAGE, C., GRAY, L. J., COLLIER, T. & POCOCK, S. J. (2008) Use of ordinal outcomes in vascular prevention trials: comparison with binary outcomes in published trials. *Stroke,* DOI: 10.1161/STROKEAHA.107.509893.

BATH, P. M. W. & GRAY, L. J. (2009) Systematic Reviews as a Tool for Planning and Interpreting Trials. *International Journal of Stroke,* January 2009.

BATH, P. M. W., GRAY, L. J., GREEN, A. R. & FOR NEMAS COLLABORATORS (2008) NXY-059 Efficacy Meta-analysis in individual Animals with Stroke (NEMAS). *In preparation*.

BATH, P. M. W., GRAY, L. J. & WAHLGREN, N. G. (2007) Should data monitoring committees assess efficacy when considering safety in trials in acute stroke? *International Journal of Clinical Practice,* 61, 1749-1755.

BERGE, E. & BARER, D. (2002) Could stroke trials be missing important treatment effects? *Cerebrovascular Diseases,* 13, 73-75.

BERRIDGE, D. M. & WHITEHEAD, J. (1991) Analysis of failure time data with ordinal categories of responses. *Statistics in Medicine,* 10, 1703-1710.

BHATT, D. L., FOX, K. A. A., WERNER HACKE, C. B., BERGER, P. B., BLACK, H. R., BODEN, W. E., CACOUB, P., COHEN, E. A., CREAGER, M. A., EASTON, J. D., FLATHER, M. D., HAFFNER, S. M., HAMM, C. W., HANKEY, G. J., CLAIBORNE JOHNSTON, S., MAK, K.-H., MAS, J.-L., MONTALESCOT, G., PEARSON, T. A., STEG, P. G., D, STEINHUBL, S. R., WEBER, M. A., BRENNAN, D. M., FABRY-RIBAUDO, L., BOOTH, J., TOPOL, E. J. & FOR THE CHARISMA INVESTIGATORS (2006) Clopidogrel and aspirin versus aspirin alone for the prevention of atherothrombotic events. *The New England Journal of Medicine,* 354, 1706-1717.

BHATT, S. P., LUQMAN-ARAFATH, T. K. & GULERIA, R. (2007) Non-pharmacological management of hypertension. *Indian Journal of Medical Sciences,* 61, 616-624.

BLANCO, M., NOMBELA, F., CASTELLANOS, M., RODRIGUEZ-YANEZ, M., GARCIA-GIL, M., LEIRA, R., LIZASOAIN, I., SERENA, J., VIVANCOS, J., MORO, M. A., DAVALOS, A. & CASTILLO, J. (2007) Statin treatment withdrawal in ischemic stroke. A controlled randomized study. *Neurology,* 69, 904-10.

BLAND, M. (2000) *An introduction to medical statistics,* Oxford, Oxford University Press.

BLOCH, R. F. (1988) Interobserver agreement for the assessment of handicap in stroke patients (letter). *Stroke,* 19, 1448.

BLOOD PRESSURE IN ACUTE STROKE COLLABORATION (BASC) (2001) Interventions for deliberately altering blood pressure in acute stroke (Cochrane Review). *The Cochrane Library.* 4 ed. Oxford, Update Software.

BONITA, R. (1992) Epidemiology of stroke. *Lancet,* 339, 342-344.

BONITA, R., BEAGLEHOLE, R. & NORTH, J. D. K. (1984) Event, incidence and case fatality rates of cerebrovascular disease in Auckland, New Zealand. *American Journal of Epidemiology,* 120, 236-43.

BOWLING, A. (1995) *Measuring disease,* Buckingham, Open University Press.

BOWLING, A. (1997) *Measuring health. A review of quality of life measurement scales,* Buckingham, Open University Press.

BRODERICK, J. P., LU, M., KOTHARI, R., LEVINE, S. R., LYDEN, P. D., HALEY, E. C., BROTT, T. G., GROTTA, J., TILLEY, B. C., MARLER, J. R. & FRANKEL, M. (2000) Finding the most powerful measures of the effectiveness of tissue plasminogen activator in the NINDS tPA Stroke Trial. *Stroke,* 31, 2335-2341.

BROOKS, R. & WITH THE EUROQOL GROUP (1996) EuroQol: the current state of play. *Health Policy,* 37, 53-72.

BROTT, T., ADAMS, H. P., OLINGER, C. P., MARLER, J. R., BARSAN, W. G., BILLER, J., SPILKER, J., HOLLERAN, R., EBERLE, R., HERTZBERG, V., RORICK, M., MOOMAW, C. J. & WALKER, M. (1989) Measurements of acute cerebral infarction: a clinical examination scale. *Stroke,* 20, 864-870.

CANDELISE, L., ARITZU, E., CICCONE, A., RICCI, S., WARDLAW, J., TOGNONI, G., RONCAGLIONI, M. C., NEGRI, E., COLOMBO, F., BOCCARDI, E., DEGRANDI, C., SCIALFA, G., ARGENTINO, C., BERTELE, V., MAGGIONI, A. P., PERRONE, P., BARNETT, H. J. M., BOGOUSSLAVSKY, J., DELFAVERO, A., LOI, U., PETO, R., WARLOW, C., CANZI, S., COMPARETTI, S., CLERICI, F., PALUMBO, A., SGARONI, G., POLONARA, S., REGINELLI, R., CERAVOLO, M. G., PROVINCIALI, L., DELGOBBO, M., SCARPINO, O., BOTTACCHI, E., DALESSANDRO, G., DIGIOVANNI, M., BLANC, S., ROVEYAZ, L., RALLI, L., VANNI, D., REFI, C., FEDERICO, F., CONTE, C., INCHINGOLO, V., INSABATO, R., SALSA, F., LORIZIO, A., ROTTOLI, M. R., BRUNI, L., DEFANTI, C. A., FERA, L., CAMERLINGO, M., CASTO, L., CENSORI, B., MAMOLI, A., PORAZZI, D., GRAMPA, G., LASPINA, I., GIGLIA, L., AVENIA, V., GUELI, S., LOLLI, V., MIELE, V., SANTANGELO, M., COPPOLA, G., TRIANNI, G., MARRA, M., GRECO, E., TONTI, D., PRETOLANI, E., STELLIO, M., ARNABOLDI, M., CIOLA, R., DANIELI, G., REZZONICO, M., GUIDOTTI, M., PELLEGRINI, G., RAUDINO, F., DELFAVERO, C., FRATTINI, T., RICCARDI, T., LEVIMINZI, C., LOCATELLI, F., PASSERI, F., LOMBARDO, G., COCCO, F., PRATESI, M., SANTINI, S., CARDOPATRI, F., TAFANI, O., LANDINI, G. C., PIERAGNOLI, E., BELLESI, R., BAGNOLI, L., GHETTI, A., MARRAZZA, O. B., MENEGAZZO, P., SPOLVERI, S., CAPPELLETTI, C., CANDELIERE, G., et al. (1995) Randomized Controlled Trial of Streptokinase, Aspirin, and Combination of Both in Treatment of Acute Ischemic Stroke. *Lancet,* 346, 1509-1514.

CARANDANG, R., SESHADRI, S., BEISER, A., KELLY-HAYES, M., KASE, C. S., KANNEL, W. B. & WOLF, P. A. (2006) Trends in incidence, lifetime risk, severity, and 30-day mortality of stroke over the past 50 years. *Journal of the American Medical Association,* 296, 2939-2946.

CAST (CHINESE ACUTE STROKE TRIAL) COLLABORATIVE GROUP (1997) CAST: randomised placebo-controlled trial of early aspirin use in 20,000 patients with acute ischaemic stroke. *Lancet,* 349, 1641-1649.

CELANI, M. G., CANTISANI, T. A., RIGHETTI, E., SPIZZICHIONO, L., RICCI, S. & ON BEHALF OF THE ITALIAN INTERNATIONAL STROKE TRIAL (IST) COLLABORATORS (2002) Different measures for assessing stroke outcome. An analysis from the International Stroke Trial in Italy. *Stroke,* 33, 218-223.

CHEN, Z. M., SANDERCOCK, P. A. G., PAN, H., COUNSELL, C., COLLINS, R., LIU, L., XIE, J., WARLOW, C., PETO, R. & ON BEHLF OF THE CAST AND IST COLLABORATIVE GROUPS (2000) Indications for early aspirin use in acute ischemic stroke. *Stroke,* 31, 1240-1249.

CLARK, W. M., ALBERS, G. W., MADDEN, K. P. & HAMILTON, S. (2000) The rtPA (alteplase) 0-6 hour acute stroke trial, part A (A0276g). Results of a double-blind, placebo-controlled, multicenter study. *Stroke,* 31, 811-816.

CLARK, W. M., WARACH, S. J. & PETTIGREW, L. C. (1997) A randomised dose-response trial of citicoline in acute ischemic stroke patients. *Neurology,* 49, 671-678.

CLARK, W. M., WARACH, S. J., PETTIGREW, L. C., GAMMANS, R. E., SABOUNJIAN, L. A. & FOR THE CITICOLINE STROKE STUDY GROUP (1997) A randomized dose-response trial of citicoline in acute ischemic stroke patients. *Neurology,* 49, 671-678.

CLARK, W. M., WECHSLER, L. R., SABOUNJIAN, L. A. & SCHWIDERSKI, U. E. (2001) A phase III randomized efficacy trial of 2000mg citicoline in acute ischemic stroke patients. *Neurology,* 57, 1595-1602.

CLARK, W. M., WILLIAMS, B. J., SELZER, K. A., ZWEIFLER, R. M., SABOUNJIAN, L. A. & GAMMANS, R. E. (1999) A randomised efficacy trial of citicoline in patients with acute ischaemic stroke. *Stroke,* 30, 2592-2597.

CLARK, W. M., WISSMAN, S., ALBERS, G. W., JHAMANDAS, J. H., MADDEN, K. P. & HAMILTON, S. (1999) Recombinant tissue-type plasminogen activator (alteplase) for ischemic stroke 3 to 5 hours after symptom onset - The ATLANTIS study: A randomized controlled trial. *Journal of the American Medical Association,* 282, 2019-2026.

COLLIN, C., WADE, D. T., DAVIES, S. & HORNE, V. (1988) The Barthel Index: a reliability study. *International Disability Studies,* 10, 61-63.

COLLINS, P. (2002) Clinical cardiovascular studies of hormone replacement therapy. *American Journal of Cardiology,* 90, 30F-34F.

COLLINS, R., PETO, R. & MACMAHON, S. (1990) Blood pressure, stroke and coronary heart disease, part 2: short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. *Lancet,* 335, 827-838.

COMMITTEE FOR PROPRIETARY MEDICINAL PRODUCTS (CPMP) (2001) Points to Consider on Clinical Investigation of Medicinal Products for the Treatment of Acute Stroke.

CONOVER, W. J. (1971) *Practical nonparametric statistics,* New York, John Wiley & Sons.

COOPE, J. & WARRENDER, T. S. (1986) Randomised trial of treatment of hypertension in elderly patients in primary care. *British Medical Journal,* 293, 1145-1151.

CORR, S. & BAYER, A. (1995) Occupational therapy for stroke patients after hospital discharge - a randomised controlled trial. *Clinical Rehabilitation,* 9, 291-296.

COULL, A. J., LOVETT, J. K., ROTHWELL, P. M. & OXFORD VASCULAR STUDY (2004) Population based study of early risk stroke after transient ischaemic attack or minor stroke: implications for public education and organisation of services. *British Medical Journal,* 328, 326.

COUNSELL, C., DENNIS, M., MCDOWELL, M. & WARLOW, C. (2002) Predicting outcome after acute and subacute stroke. Development and validation of new prognostic models. *Stroke,* 33, 1041-1047.

D'OLHABERRIAGUE, L., LITVAN, I., MITSIAS, P. & MANSBACH, H. H. (1996) A reappraisal of reliability and validity studies in stroke. *Stroke,* 27, 2331-2336.

DAVIS, S. M., LEES, K. R., ALBERS, G. W., DIENER, H. C., MARKABI, S., KARLSSON, G., NORRIS, J. & FOR THE ASSIST INVESTIGATORS (2000) Selfotel in acute ischemic stroke. Possible neurotoxic effects of an NMDA antagonist. *Stroke,* 31, 347-354.

DE DEYN, P., DE REUCK, J., DEBERTDT, W., VLIETINCK, R. & ORGOGOZO, J. M. (1997) Treatment of acute ischemic stroke with piracetam. *Stroke,* 28, 2347-2352.

DE FREITAS, G. R. & BOGOUSSLAVSKY, J. (1997) What is the place of clinical assessment in acute stroke management? IN BOGOUSSLAVSKY, J. (Ed.) *Acute stroke treatment.* London, Martin Dunitz.

DE HAAN, R. N., LIMBURG, M., BOSSUYT, P., VAN DER MEULEN, J. & AARONSON, N. (1995) The clinical meaning of Rankin 'handicap' grades after stroke. *Stroke,* 26, No 11, 2027-2030.

DEL ZOPPO, G. J., HIGASHIDA, R. T., FURLAN, A. J., PESSIN, M. S., ROWLEY, H. A., GENT, M. & AND THE PROACT INVESTIGATORS (1998) PROACT:A phase II randomized trial of recombinant Pro-Urokinase by direct arterial delivery in acute middle cerebral artery stroke. *Stroke,* 29, 4-11.

DENNIS, M. S. (2004) The FOOD Trial 3 - Results of a multicentre RCT comparing feeding via a nasogastric tube (NG) and percutaneous endoscopic gastrostomy (PEG) in the first month after stroke (abstract). *Cerebrovascular Diseases,* 17 (Suppl 5), 1-125.

DEPARTMENT OF HEALTH (2007) *National Stroke Strategy,* London, Department of Health.

DI CARLO, A., LAMASSA, M., BALDERESCHI, M., PRACUCCI, G., BASILE, A. M., WOLFE, C. D., M., G., RUDD, A., GHETTI, A., INZITARI, D. & EUROPEAN BIOMED STUDY OF STROKE CARE GROUP (2003) Sex differences in the clinical presentation, resource use, and 3-month outcome of acute stroke in Europe. Data from a multicenter multinational hospital-based registry. *Stroke,* 34, 1114-1119.

DONABEDIAN, A. (1980) *Explorations in Quality Assessment and Monitoring Vol. 1. The Definition of Quality and Approaches to Its Assessment.* Ann Arbor, MI: Health Administration Press.

DONNAN, G. A., DAVIS, S. M., CHAMBERS, B. R., GATES, P. C., KANKEY, G. J., MCNEIL, J. J. & ROSEN, D. (1996) Streptokinase for acute ischemic stroke with relationship to time of administration: Australian Streptokinase (ASK) Trial Study Group. *Journal of the American Medical Association,* 276, 995-996.

DUNCAN, D. B. (1955) Multiple range and multiple F tests. *Biometrics* 11, 1-42.

EBRAHIM, S. & HARWOOD, R. (2003) *Stroke. Epidemiology, evidence and clinical practice,* Oxford, Oxford University Press.

EBRAHIM, S., NOURI, F. & BARER, D. (1985) Measuring Disability after a Stroke. *Journal of Epidemiology and Community Health,* 39, 86-89.

EFRON, B. & TIBSHIRANI, R. J. (1993) *An introduction to the Bootstrap,* New York, Chapman & Hall.

ELKINS, J. S., KHATABI, T., FUNG, L., ROOTENBERG, J. & JOHNSTON, S. C. (2006) Recruiting subjects for acute stroke trials. *Stroke,* 37, 123-128.

ELTING, J.-W., SULTER, G. A., KASTE, M., LEES, K. R., DIENER, H. C., HOMMEL, M., VERSAVEL, M., TEELKEN, A. W. & DE KEYSER, J. (2002) AMPA Antagonist ZK200775 in patients with acute ischemic stroke. Possible glial cell toxicity detected by monitoring of S-100B serum levels. *Stroke,* 33, 2813-2818.

ENLIMOMAB ACUTE STROKE TRIAL INVESTIGATORS (2001) Use of anti-ICAM-1 therapy in ischemic stroke. Results of the Enlimomab Acute Stroke Trial. *Neurology,* 57, 1428-1434.

FAGERBERG, B., CLAESSON, L., GOSMAN-HEDSTROM, G. & BLOMSTRAND, C. (2000) Effect of acute stroke unit care integrated with care continuum versus conventional treatment: a randomized 1-year study of elderly patients. *Stroke,* 31, 2578-2584.

FARRELL, B., GODWIN, J., RICHARDS, S. & WARLOW, C. (1991) The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: Final results. *Journal of Neurology, Neurosurgery and Psychiatry,* 54, 1044-1054.

FEIGIN, V. L., LAWES, C. M. M., BANNETT, D. A. & ANDERSON, C. S. (2003) Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. *The Lancet Neurology,* 2, 43-53.

FERREIRA-GONZÁLEZ, I., PERMANYER-MIRALDA, G., DOMINGO-SALVANY, A., BUSSE, J. W., HEELS-ANSDELL, D., MONTORI, V. M., AKL, E. A., BRYANT, D. M., ALONSO-COELLO, P., ALONSO, J., WORSTER, A., UPADHYE, S., JAESCHKE, R., SCHÜNEMANN, H. J., PACHECO-HUERGO, V., WU, P., MILLS, E. J. & GUYATT, G. H. (2007) Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *British Medical Journal,* 334, 786.

FLIGNER, M. A. & POLICELLO, G. E. (1981) Robust Rank procedures for the Behrens-Fisher problem. *Journal of the American Statistics Association,* 76, 162-168.

FURLAN, A., HIGASHIDA, R., WECHSLER, L., GENT, M., ROWLEY, H., KASE, C., PESSIN, M., AHUJA, A., CALLAHAN, F., CLARK, W. M., SILVER, F. & RIVERA, F. (1999) Intra-arterial prourokinase for acute ischemic stroke. The PROACT II study: a randomized trial. *Journal of the American Medical Association,* 282, 2003-2011.

FURLAN, A. J. (2002) Acute stroke trials: strengthening the underpowered. *Stroke,* 33, 1450-1451.

GABRIEL, S. R., CARMONA, L., ROQUE, M., SANCHEZ, G. L. & BONFILL, X. (2005) Hormone replacement therapy for preventing cardiovascular disease in post-menopausal women. *Cochrane Database of Systematic Reviews*, CD002229.

GARDNER, M. J. & ALTMAN, D. G. (Eds.) (1989) *Statistics with confidence,* London, British Medical Journal.

GENSTAT (2005) GenStat Release 8.1 (PC/Windows XP),. Lawes Agricultural Trust (Rothamsted Experimental Station).

GILBERTSON, L., LANGHORNE, P., WALKER, A., ALLEN, A. & MURRAY, G. D. (2000) Domiciliary occupational therapy for patients with stroke discharged from hospital: randomised controlled trial. *British Medical Journal,* 320, 603-606.

GRANGER, C. V., DEWIS, L. S., PETERS, N. C., SHERWOOD, C. C. & BARRETT, J. E. (1979) Stroke rehabilitation: Analysis of repeated Barthel Index measures. *Archives of Physical and Medical Rehabilitation,* 60, 14-17.

GRAY, L. J., SPRIGG, N., BATH, P. M., SORENSEN, P., LINDENSTROM, E., BOYSEN, G., DE DEYN, P. P., FRIIS, P., LEYS, D., MARTTILA, R., OLSSON, J.-E., O'NEILL, D., RINGELSTEIN, B., VAN DER SANDE, J.-J., TURPIE, A. G. G. & FOR THE TAIST INVESTIGATORS (2006) Significant variation in mortality and functional outcome after acute ischaemic stroke between western countries: data from the tinzaparin in acute ischaemic stroke trial (TAIST). *Journal of Neurology Neurosurgery and Psychiatry,* 77, 327-333.

GRAY, L. J., SPRIGG, N., BATH, P. M. W., BOYSEN, G., DE DEYN, P. P., LEYS, D., O'NEILL, D., RINGLESTEIN, E. B. & FOR THE TAIST INVESTIGATORS (2007) Sex differences in quality of life in stroke survivors. Data from the Tinzaparin in Acute Ischaemic Stroke Trial (TAIST). *Stroke,* 38, 2960.

GRAY, L. J., SPRIGG, N., BATH, P. M. W., CHRISTENSEN, H., DE DEYN, P. P., LEYS, D., O'NEILL, D., RINGLESTEIN, E. B. & THE TAIST INVESTIGATORS (2008) Significant variation in quality of life after acute ischaemic stroke between western countries: data from the 'Tinzaparin in Acute Ischaemic Stroke Trial' (TAIST). *In preparation.*

GRODSTEIN, F., STAMPFER, M. J., MANSON, J. E., COLDITZ, G. A., WILLET, W. C., ROSNER, B., SPEIZER, F. E. & HENNEKENS, C. H. (1996) Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. *New England Journal of Medicine,* 335, 453-461.

GROTTA, J. & THE US AND CANADIAN LUBELUZOLE ISCHEMIC STROKE STUDY GROUP (1997) Lubeluzole treatment of acute ischemic stroke. *Stroke,* 28, 2338-2346.

GUSEV, E. I., SKVORTSOVA, S. I., DAMBINOVA, S. A., RAEVSKIY, K. S., ALEKSEEV, A. A., BASHKATOVA, V. G., KOVALENKO, A. V., KUDRIN, V. S. & YAKOVLEVA, E. V. (2000) Neuroportective effects of glycine for therapy of acute ischaemic stroke. *Cerebrovascular Diseases,* 10, 49-60.

HACKE, W., ALBERS, G. W., AL-RAWI, Y., BOGOUSSLAVSKY, J., DAVALOS, A., ELIASZIW, M., FISCHER, M., FURLAN, A. J., KASTE, M., LEES, K. R., SOEHNGEN, M., WARACH, S. & FOR THE DIAS STUDY GROUP (2005) The Desmoteplase in acute ischemic stroke trial (DIAS). *Stroke,* 36, 66-73.

HACKE, W., BLUHMKI, E., STEINER, T., TATLISUMAK, T., MAHAGNE, M.-H., SACCHETTI, M.-L. & MEIER, D. (1998) Dichotomized efficacy end points and global end-point analysis applied to the ECASS intention to treat data set. Post hoc analysis of ECASS 1. *Stroke,* 29, 2073-2075.

HACKE, W., BLUHMKI, E., STEINER, T., TATLISUMAK, T., MAHAGNE, M. H., SACCHETTI, M. L. & MEIER, D. (1998) Dichotomized efficacy end points and global end-point analysis applied to the ECASS intention-to-treat data set - Post hoc analysis of ECASS I. *Stroke,* 29, 2073-2075.

HACKE, W., KASTE, M., FIESCHI, C., TONI, D., LESAFFRE, E., VON KUMMER, R., BOYSEN, G., BLUHMKI, E., HOXTER, G., MAHAGNE, M.-H., HENNERICI, M. & ECASS STUDY GROUP. (1995) Intravenous thrombolysis with tissue plasminogen activator for acute hemispheric stroke. The European Cooperative Acute Stroke Trial (ECASS). *Journal of the American Medical Association,* 274, 1017-1025.

HACKE, W., KASTE, M., FIESCHI, C., VON KUMMER, R., DAVALOS, A., MEIER, D., LARRUE, V., BLUHMKI, E., DAVIS, S., DONNAN, G., SCHNEIDER, D., DIEZ-TEJEDOR, E. & TROUILLAS, P. (1998) Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). *Lancet,* 352, 1245-1251.

HALEY, E. C. (1998) High-dose tirilazad for acute stroke (RANTTAS II). *Stroke,* 29, 1256-1257.

HALKES, P. H. A., GRAY, L. J., BATH, P. M. W., BOUSSER, M.-G., DIENER, H.-C., GUIRAUD-CHAUMEIL, B., YATSU, F., ALGRA, A. & ON BEHALF OF THE DIPYRIDAMOLE IN STROKE COLLABORATION (DISC) (2008) Dipyridamole plus aspirin in the secondary prevention after TIA or stroke of arterial origin: a meta-analysis by risk using individual patient data from randomised trials. *Journal of Neurology, Neurosurgery and Psychiatry.* DOI:10.1136/jnnp.2008.143875.

HANTSON, L. & DE KEYSER, J. (1994) Neurological scales in the assessment of cerebral infarction. *Cerebrovascular Diseases,* 4, 7-14.

HARBISON, J., HOSSAIN, O., JENKINSON, D., DAVIS, J., LOUW, S. J. & FORD, G. A. (2003) Diagnostic Accuracy of Stroke Referrals From Primary Care, Emergency Room Physicians, and Ambulance Staff Using the Face Arm Speech Test. *Stroke,* 34, 71.

HAUCK, W. W., ANDERSON, S. & MARCUS, S. M. (1998) Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials,* 19, 249-256.

HEART PROTECTION STUDY COLLABORATIVE GROUP (2002) MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20536 high-risk individuals: a randomised placebo-controlled trial. *Lancet,* 360, 7-22.

HERNANDEZ, A. V., EIJKEMANS, M. J. C. & STEYERBERG, E. W. (2006) Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? *Annals of Epidemiology,* 16, 41-48.

HERNANDEZ, A. V., STEYERBERG, E. W., BUTCHER, I., MUSHKUDIANI, N., TAYLOR, G. S., MURRAY, G. D., MARMAROU, A., CHOI, S. C., LU, J., HABBEMA, J. D. F. & MAAS, A. I. R. (2006) Adjustment for strong predictors of outcome in traumatic brain injury trials: 25% reduction in sample size requirements in the IMPACT study. *Journal of Neurotrauma,* 23, 1295-1303.

HERNANDEZ, A. V., STEYERBERG, E. W. & HABBEMA, J. D. F. (2004) Clinical trials with dichotomous end-points: covariate adjustment increases power and potentially reduces sample size. *Journal of Clinical Epidemiology,* 57, 454-460.

HOLLANDER, M. & WOLFE, D. A. (1999) *Nonparametric statistical methods,* New York, John Wiley & Sons inc.

HOTELLING, H. (1931) The generalization of Student's ratio. *Annals of Mathematics and Statistics,* 2, 360-378.

HUI, E., LUM, C. M., WOO, J., K.H., O. & KAY, R. L. C. (1995) Outcomes of elderly stroke patients. Day hospital versus conventional medical management. *Stroke,* 26, 1616-1619.

INDREDAVIK, B., BAKKE, F., SOLBERG, R., ROKSETH, R., HAAHEIM, L. L. & HOLME, I. (1991) Benefit of a stroke unit: A randomised controlled trial. *Stroke,* 22, 1026-1032.

INGRAM, D. D. & KLEINMAN, J. C. (1989) Empirical comparisons of proportional hazards and logistic regression models. *Statistics in Medicine,* 8, 525-538.

INTERNATIONAL STROKE TRIAL COLLABORATIVE GROUP (1997) The International Stroke Trial (IST); a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *Lancet,* 349, 1569-1581.

INTRAVENOUS MAGNESIUM EFFICACY IN STROKE (IMAGES) STUDY INVESTIGATORS (2004) Magnesium for acute stroke (Intravenous Magnesium Efficacy in Stroke trial): randomised controlled trial. *Lancet,* 363, 439.

JAGER, H. R. (2000) Diagnosis of stroke with advanced CT and MR imaging. *British Medical Bulletin,* 56, 318-333.

JAKOBSSON, U. (2004) Statistical presentation and analysis of ordinal data in nursing research. *Scandinavian Journal of Caring Sciences,* 18, 437-440.

JENNETT, B. & BOND, M. (1975) Assessment of outcome after severe brain damage. A practical scale. *Lancet,* i, 480-484.

JOHNSTON, K. C., CONNORS, A. F., WAGNER, D. P. & HALEY, E. C. (2004) Risk adjustment effect on stroke clinical trials. *Stroke,* 35, e43-e45.

JUBY, L. C., LINCOLN, N. B. & BERMAN, P. (1996) The effect of a stroke rehabilitation unit on functional and psychological outcome: A randomised controlled trial. *Cerebrovascular Diseases,* 6, 106-110.

JULIOUS, S. A., CAMPBELL, M. J., WALKER, S. J., GEORGE, S. L. & MACHIN, D. (2000) Sample size for cancer trials where health related quality of life is the primary outcome. *British Journal of Cancer,* 83, 959-963.

JUTTLER, E., SCHWAB, S., SCHMIEDEK, P., UNTERBERG, A., HENNERICI, M. G., WOITIK, J., WITTE, S., JENETZKY, E., HACKE, W. & FOR THE DESTINY STUDY GROUP (2007) Decompressive surgery for the treatment of malignant infarction of the middle cerebral artery (DESTINY): A randomized, controlled trial. *Stroke,* 38, 2518-2525.

KAKAR, P., GUNARATHNE, A. & LIP, G. Y. H. (2006) Stroke: ethnic differences do exist. *Expert Review of Neurotherapeutics,* 6, 1769-1771.

KALRA, L., DALE, P. & CROME, P. (1993) Improving stroke rehabilitation: a controlled study. *Stroke,* 24, 1462-1467.

KALRA, L. & EADE, J. (1995) Role of stroke rehabilitation units in managing severe disability after stroke. *Stroke,* 26, 2031-2034.

KALRA, L., EVANS, A., PEREZ, I., KNAPP, M., DONALDSON, N., SWIFT, C. & 1683 (2000) Alternative strategies for stroke care: a prospective randomised controlled trial. *Lancet,* 356(9233), 894-9.

KASTE, M., PALOMAKI, H. & SARNA, S. (1995) Where and how should elderly stroke patients be treated? A randomized trial. *Stroke.,* 26, 249-253.

KAY, R., WONG, K. S., YU, Y. L., CHAN, Y. W., TSOI, T. H., AHUJA, A. T., CHAN, F. L., FONG, K. Y., LAW, C. B., WONG, A. & WOO, J. (1995) Low-Molecular-Weight Heparin for the Treatment of Acute Ischemic Stroke. *New England Journal of Medicine,* 333, 1588-1593.

KEARNEY, P. M. & PRYOR, J. (2004) The International Classification of Functioning, Disability and Health (ICF) and nursing. *Journal of Advanced Nursing,* 46, 162-170.

KIDWELL, C., LIEBESKIND, D., STARKMAN, S. & SAVER, J. (2001) Trends in acute ischaemic stroke trials through the 20th century. *Stroke,* 32, 1349-1359.

LAMPL, Y., BOAZ, M., GILAD, R., LORBERBOYM, M., DABBY, R., RAPOPORT, A., ANCA-HERSHKOWITZ, M. & SADEH, M. (2007) Minocycline treatment in acute stroke. An open-label, evaluator-blinded study. *Neurology,* 69, 1404-1410.

LAVALLEY, M. P. & FELSON, D. T. (2002) Statistical presentation and analysis of ordered categorical outcome data in rheumatology journals. *Arthritis & Rheumatism-Arthritis Care & Research,* 47, 255-259.

LEES, K. R., ZIVIN, J. A., ASHWOOD, T., DAVALOS, A., DAVIS, S. M., DIENER, H.-C., GROTTA, J., LYDEN, P., SHUAIB, A., HARDEMARK, H.-G., WASIEWSKI, W. W. & FOR THE STROKE-ACUTE ISCHEMIC NXY TREATMENT (SAINT I) TRIAL INVESTIGATORS (2006) NXY-059 for acute ischemic stroke. *New England Journal of Medicine,* 354, 588-599.

LEFKOPOULOU, M. & RYAN, L. (1993) Global tests for multiple binary outcomes. *Biometrics,* 49, 975-88.

LINCOLN, N. B. & EDMANS, J. A. (1990) A re-evaluation of the Rivermead ADL scale for elderly patients with stroke. *Age and Ageing,* 19, 19-24.

LINDLEY, R. I., WADDELL, F., LIVINGSTONE, M., SANDERCOCK, P., DENNIS, M. S., SLATTERY, J., SMITH, B. & WARLOW, C. (1994) Can Simple Questions Assess Outcome after Stroke. *Cerebrovascular Diseases,* 4, 314-324.

LOGAN, P. A., AHERN, J., GLADMAN, J. R. F. & LINCOLN, N. B. (1997a) A randomised controlled trial of enhanced social service occupational therapy for stroke patients. *Clinical Rehabilitation,* 11, 107-113.

LUMLEY, T. (1996) Generalized estimating equations for ordinal data. *Biometrics,* 52, 354-361.

MAAS, A. I. R., MURRAY, G. D., HANNEY III, H., KASSEM, N., LEGRAND, V., MANGELUS, M., MUIZELAAR, J.-P., STOCCHETTI, N., KNOLLER, N. & ON BEHALF OF THE PHARMOS TBI INVESTIGATORS (2006) Efficacy and safety of dexanabinol in severe traumatic brain injury: results of a phase III randomised, placebo-controlled, clinical trial. *Lancet Neurology,* 5, 38-45.

MACLEOD, M. & SANDERCOCK, P. (2005) Can systematic reviews help animal experimental work? *Research Defence Society News,* Winter.

MACMAHON, S., PETO, R., CUTLER, J., COLLINS, R., SORLIE, P. & NEATON, J. (1990) Blood pressure, stroke and coronary heart disease. Part I: effects of prolonged differences in blood pressure - evidence from nine prospective observational studies corrected for the regression dilution bias. *Lancet,* 335, 765-774.

MAHONEY, F. I. & BARTHEL, D. W. (1965) Functional evaluation: The Barthel Index. *Maryland State Medical Journal*, 61-65.

MANT, J., HOBBS, F. D. R., FLETCHER, K., ROALFE, A., FITZMAURICE, D., LIP, G. Y. H., MURRAY, E. & ON BEHALF OF THE BAFTA INVESTIGATORS AND THE MIDLAND PRACTICES NETWORK (MIDREC) (2007) Warfarin versus aspirin for stroke prevention in an elderly community population with atrial fibrillation (the Birmingham Atrial Fibrillation Treatment of the Aged Study, BAFTA): a randomised controlled trial. *Lancet,* 370, 493-503.

MARMAROU, A., LU, J., BUTCHER, I., MCHUGH, G. S., MUSHKUDIANI, N. A., MURRAY, G. D., STEYERBERG, E. W. & MAAS, A. I. (2007) IMPACT database of traumatic brain injury: design and description. *Journal of Neurotrauma,* 24, 239-50.

MARMOT, M. G. & POULTER, N. R. (1992) Primary prevention of stroke. *Lancet,* 339, 344-347.

MAYER, S. A., BRUN, N. C., BEGTRUP, K., BRODERICK, J., DAVIS, S., DIRINGER, M. N., SKOLNICK, B. E., STEINER, T. & FOR THE RECOMBINANT ACTIVATED FACTOR VII INTRACEREBRAL HEMORRHAGE TRIAL INVESTIGATORS. (2005) Recombinant activated factor VII for acute intracerebral haemorrhage. *New England Journal of Medicine,* 352, 777-785.

MAYER, S. A., BRUN, N. C., BRODERICK, J., DIRINGER, M. N., DAVIS, S., SKOLNICK, B. E. & STEINER, T. (2007) The FAST Trial: Main Results. *European Stroke Conference.* Glasgow.

MENDELOW, A. D., GREGSON, B. A., FERNADES, H. M., MURRAY, G. D., TEASDALE, G. M., HOPE, T. D., KARIMI, A., SHAW, M. D. M., BARER, D. H. & FOR THE STICH INVESTIGATORS (2005) Early surgery versus initial conservative treatment in patients with spontaneour supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial. *Lancet,* 365, 387-97.

MILLIKAN, C. H., MCDOWELL, F. & EASTON, J. D. (1987) *Stroke,* Philadelphia, Lea and Febiger.

MORRIS, A. D., RITCHIE, C., GROSSET, D. G. & 1225 (1995) A pilot study of streptokinase for acute cerebral infarction. *Quarterly Journal of Medicine,* 88, 727-731.

MOSES, L. E., EMERSON, J. D. & HOSSEINI, H. (1984) Analyzing Data from Ordered Categories. *New England Journal of Medicine,* 311, 442-448.

MULTICENTER ACUTE STROKE TRIAL - EUROPE STUDY GROUP (1996) Thrombolytic therapy with streptokinase in acute ischaemic stroke. *New England Journal of Medicine,* 335, 145-150.

MULTICENTRE ACUTE STROKE TRIAL-ITALY (MAST-I) GROUP (1995) Randomised controlled trial of streptokinase, aspirin, and combination of both in treatment of acute ischaemic stroke. *Lancet,* 346, 1509-1514.

MURRAY, G. D., BARER, D., CHOI, S., FERNANDES, H., GREGSON, B., LEES, K. R., MAAS, A. I. R., MARMAROU, A., MENDELOW, A. D., STEYERBERG, E. W., TAYLOR, G. S., TEASDALE, G. M. & WEIR, C. J. (2005) Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. *Journal of Neurotrauma,* 22, 511-517.

NATIONAL AUDIT OFFICE (2005) *Reducing brain damage, faster access to better stroke care,* London, Stationary office.

NHS DIRECT (2001) What is a stroke or TIA? , NHS Direct.

NOETHER, G. (1987) Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association,* 82, 645-647.

NORTH AMERICAN SYMPTOMATIC CAROTID ENDARTERECTOMY TRIAL COLLABORATORS (1991) Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *New England Journal of Medicine,* 325, 445-453.

ORGOGOZO, J. M. (1995) TESS II. *Unpublished Work.*

PARKER, C. J., GLADMAN, J. R. F., DRUMMOND, A. E. R., DEWEY, M. E., LINCOLN, N. B., BARER, D., LOGAN, P. A. & RADFORD, K. A. (2001) A multicentre randomized controlled trial of leisure therapy and conventional occupational therapy after stroke. *Clinical Rehabilitation,* 15, 42-52.

PAYNE, R. (1993) SSIGNTEST *Genstat Reference Manual,* 3.

PETERS, G. R., HWANG, L.-J., MUSCH, B., BROSSE, D. M. & ORGOGOZO, J. M. (1996) Safety and efficacy of 6 mg/kg/day tirilizad mesylate in patients with acute ischemic stroke (TESS study). *Stroke,* 27, 195.

PLANT, J. D., GIOVANINI, J. N. & VILLARROEL, A. (2007) Frequency of appropriate and inappropriate presentation and analysis methods of ordered categorical data in the veterinary dermatology literature from January 2003 to June 2006. *Veterinary Dermatology* 18, 260-266.

POCOCK, S. J., ASSMANN, S. E., ENOS, L. E. & KASTEN, L. E. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trials reporting: current practice and problems. *Statistics in Medicine,* 21, 2917-2930.

RAAB, G. M., DAY, S. & SALES, J. (2000) How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials,* 21, 330-342.

RANKIN, J. (1957) Cerebral vascular accidents in patients over the age of 60. 2. Prognosis. *Scottish Medical Journal,* 2, 200-215.

ROBERTS, L. & COUNSELL, C. (1998) Assessment of clinical outcomes in acute stroke trials. *Stroke,* 29, 986-991.

RONNING, O. M. & GULDVOG, B. (1998) Stroke units versus general medical wards, 1: twelve-and eighteen month survival. *Stroke,* 29, 58-62.

ROTHER, J. (2008) Neuroprotection does not work! *Stroke,* 39, 523-524.

ROTHWELL, P. M., COULL, A. J., SILVER, L. E., FAIRHEAD, J. F., GILES, M. F., LOVELOCK, C. E., REDGRAVE, J. N. E., BULL, L. M., WELCH, S. J. V., CUTHBERTSON, F. C., BINNEY, L. E., GUTNIKOV, S. A., ANSLOW, P., BANNING, A. P., MANT, D., MEHTA, Z. & FOR THE OXFORD VASCULAR STUDY (2005) Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (Oxford Vascular Study) *Lancet,* 366, 1773-83.

ROTHWELL, P. M., GILES, M. F., CHANDRATHEVA, A., MARQUARDT, L., GERAGHTY, O., REDGRAVE, J. N., LOVELOCK, C. E., BINNEY, L. E., BULL, L. M., CUTHBERTSON, F. C., WELCH, S. J., BOSCH, S., ALEXANDER, F. C., SILVER, L. E., GUTNIKOV, S. A., MEHTA, Z. & ON BEHALF OF THE EARLY USE OF EXISTING PREVENTIVE STRATEGIES FOR STROKE (EXPRESS) STUDY (2007) Effect of urgent treatment of transient ischaemic attack and minor stroke on early recurrent stroke (EXPRESS study): a prospective population-based sequential comparison. *Lancet,* 370, 1432-42.

RUSYNIAK, D. E., KIRK, M. A., MAY, J. D., KAO, L. W., BRIZENDINE, E. J., WELCH, J. L., CORDELL, W. H. & ALONSO, R. J. (2003) Hyperbaric oxygen therapy in acute ischemic stroke. *Stroke,* 34, 571-574.

SACCO, R. L., BODEN-ALBALA, B., ABEL, G., LIN, I. F., ELKIND, M. & HAUSER, W. A. (2001) Race-ethnic disparities in the impact of stroke risk factors: the Northern Manhattan Stroke Study. *Stroke,* 32, 1725-1731.

SALPETER, S. R., WALSH, J. M., GREYBER, E., ORMISTON, T. M. & SALPETER, E. E. (2004) Mortality associated with hormone replacement therapy in younger and older women: a meta-analysis. *Journal of General Internal Medicine,* 19, 791-804.

SAMSA, G. P. & MATCHAR, D. B. (2001) Have randomized controlled trials of neuroprotective drugs been underpowered? An illustration of three statistical principles. *Stroke,* 32, 669-674.

SARE, G. M., GRAY, L. J. & BATH, P. M. W. (2008) Association between hormone replacement therapy and subsequent cerebrovascular, cardiovascular and thromboembolic disease: a meta analysis. *European Heart Journal,* DOI:10.1093/eurheartj/ehn299.

SARREL, P. M. (1996) Cardiovascular disease in women: implications of hormone replacement therapy. *International Journal of Fertility and Menopausal Studies*, 41, 90-93.

SAVER, J. L. (2004) Number needed to treat estimates incorporating effects over the entire range of clinical outcomes. *Archives of Neurology,* 61, 1066-1070.

SAVITZ, S. I., BENATAR, M., SAVER, J., GROTTA, J. & FISHER, M. (2008) Outcome measures in clinical trial design for acute ischaemic stroke: physicians' attitudes and choices [Abstract]. *International Stroke Conference.* New Orleans, Stroke.

SAVITZ, S. I., LEW, R., BLUHMKI, E., HACKE, W. & FISHER, M. (2007) Shift analysis versus dichotomization of the modified Rankin Scale outcome scores in the NINDS and ECASS-II trials. *Stroke,* 38, 3205-3212.

SAXENA, R., WIJNHOUD, A. D., CARTON, H., HACKE, W., KASTE, M., PRZYBELSKI, R. J., STERN, K. N. & KOUDSTAAL, P. J. (1999) Controlled safety study of a hemoglobin-based oxygen carrier, DCLHb, in acute ischemic stroke. *Stroke,* 30, 993-996.

SENA, E., VAN DER WORP, H. B., HOWELLS, D. & MACLEOD, M. (2007) How can we improve the pre-clinical development of drugs for stroke? *Trends in Neuroscience,* 30, 433-439.

SENN, S. J. (1989) Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine,* 8, 797-799.

SHAPIRO, S. S. & WILK, M. B. (1965) An analysis of variance test for normality (complete samples). *Biometrika,* 52, 591-611.

SHERMAN, D. G., ATKINSON, R. P., CHIPPENDALE, T., LEVIN, K. A., NG, K., FUTRELL, N., HSU, C. Y., LEVY, D. E. & FOR THE STAT PARTICIPANTS (2000) Intravenous ancrod for the treatment of acute ischemic stroke. The STAT study: A randomised controlled trial. *Journal of the American Medical Association,* 283, 2395-2403.

SHINTON, R. & BEEVERS, G. (1989) Meta-analysis of the relation between cigarette smoking and stroke. *British Medical Journal,* 298, 789-794.

SHUAIB, A., LEES, K. R., LYDEN, P., GROTTA, J., DAVALOS, A., DAVIS, S. M., DIENER, H.-C., ASHWOOD, T., WASIEWSKI, W. W., EMERIBE, U. & FOR SAINT II TRIAL INVESTIGATORS (2007) NXY-059 for the treatment of acute ischemic stroke. *New England Journal of Medicine,* 357, 562-71.

SIEGEL, S. & CASTELLAN, N. J. (1988) *Nonparametric statistics for the behavioural sciences,* Singapore, McGraw-Hill.

SIVENIUS, J., PYORALA, K., HEINONEN, O. P., SALONEN, J. T. & RIEKKINEN, P. (1985) The Significance of Intensity of Rehabilitation of Stroke - a Controlled Trial. *Stroke,* 16, 928-931.

SONG, F., JEROSCH-HEROLD, C., HOLLAND, R., DRACHLER MDE, L. & HARVEY, I. (2006) Statistical methods for analysing Barthel Scores in trials of post stroke interventions: a review and computer simulations. *Clinical Rehabilitation,* 20, 347-56.

SPRIGG, N., GRAY, L. J., BATH, P. M. W., LINDENSTROM, E., BOYSEN, G., DE DEYN, P. P., FRIIS, P., LEYS, D., MARTTILA, R., OLSSON, J.-E., O'NEILL, D., RINGLESTEIN, E. B., VAN DER SANDE, J.-J., TURPIE, A. G. G. & FOR THE TAIST INVESTIGATORS (2006) Relationship between outcome and baseline blood pressure and other haemodynamic measures in acute ischaemic stroke: data from the TAIST trial. *Journal of Hypertension,* 24, 1413-1417.

SPRIGG, N., GRAY, L. J., BATH, P. M. W., LINDENSTROM, E., BOYSEN, G., DE DEYN, P. P., FRIIS, P., LEYS, D., MARTTILA, R., OLSSON, J.-E., O'NEILL, D., RINGLESTEIN, E. B., VAN DER SANDE, J.-J., TURPIE, A. G. G. & FOR THE TAIST INVESTIGATORS (2007) Stroke severity, early recovery and outcome are each related with clinical classification of stroke: Data from the Tinzaparin in Acute Ischaemic Stroke Trial' (TAIST). *Journal of the Neurological Sciences,* 254, 54-59.

STEVENS, R. S. & AMBLER, N. R. (1982) The Dover Stroke Rehabilitation Unit: a randomised controlled trial of stroke management. IN ROSE, F. C. (Ed.) *Advances in Stroke Therapy.* New York, Raven Press.

STINGELE, R., BLUHMKI, E. & HACKE, W. (2001) Bootstrap statistics of ECASS II data: Just another post hoc analysis of a negative stroke trial? *Cerebrovascular Diseases,* 11, 30-33.

STOKES, M. E., DAVIS, C. S. & KOCH, G. G. (1995) *Categorical data analysis using SAS,* Cary, NC, SAS Institute.

STREINER, D. L. & NORMAN, G. R. (1995) *Health measurement scales. A practical guide to their development and use,* Oxford, Oxford Medical Publications.

STROKE UNIT TRIALISTS' COLLABORATION (1997) Collaborative systematic review of the randomised trials of organised inpatient (stroke unit) care after stroke. *British Medical Journal,* 314, 1151.

SULTER, G., STEEN, C. & DE KEYSER, J. (1999) Use of the Barthel Index and Modified Rankin Scale in acute stroke trials. *Stroke,* 30, 1538-1541.

TEASDALE, G. & JENNETT, B. (1974) Assessment of coma and impaired consciousness. A practical scale. *Lancet,* 2, 81-83.

THE CONSORT STATEMENT (1996) Improving the quality of reporting of randomized controlled trials. *Journal of the American Medical Association,* 276, 637-639.

THE EDARAVONE ACUTE BRAIN INFARCTION STUDY GROUP (CHAIR: EIICHI OTOMO MD) (2003) Effect of a Novel Free Radical Scavenger, Edaravone (MCI-186), on acute brain infarction. *Cerebrovascular Diseases,* 15, 222-229.

THE ENOS TRIAL INVESTIGATORS (2006) 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). *International Journal of Stroke,* 1, 245-249.

THE FOOD TRIAL COLLABORATION (2005) Effect of timing and method of enteral tube feeding for dysphagic stroke patients (FOOD): a multicentre randomised controlled trial. *Lancet,* 365, 764-772.

THE MULTICENTER ACUTE STROKE TRIAL - EUROPE STUDY GROUP (1996) Thrombolytic therapy with streptokinase in acute ischemic stroke. *New England Journal of Medicine,* 335, 145-150.

THE NATIONAL INSTITUTE OF NEUROLOGICAL DISORDERS AND STROKE RT-PA STROKE STUDY GROUP (1995) Tissue plasminogen activator for acute stroke. *New England Journal of Medicine,* 333, 1581-1587.

THE OPTIMISING ANALYSIS OF STROKE TRIALS (OAST) COLLABORATION (2007) Can we improve the statistical analysis of stroke trials? Statistical re-analysis of functional outcomes in stroke trials. *Stroke,* 38, 1911-1915.

THE OPTIMISING ANALYSIS OF STROKE TRIALS (OAST) COLLABORATION (2008) Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches. *International Journal of Stroke,* 3, 78-84.

THE RANTTAS INVESTIGATORS (1996) A randomized trial of tirilazad mesylate in patients with acute stroke (RANTTAS). *Stroke,* 27, 1453-1458.

THE STIPAS INVESTIGATORS (1994) Safety study of tirilazad mesylate in patients with acute ischemic stroke (STIPAS). *Stroke.,* 25, 418-423.

THE STROKE ASSOCIATION (2008) http://www.stroke.org.uk/information/what_is_a_stroke/brain_attack.html.

THE STROKE PREVENTION BY AGGRESSIVE REDUCTION IN CHOLESTEROL LEVELS (SPARCL) INVESTIGATORS (2006) High-dose atorvastatin after stroke or transient ischemic attack. *The New England Journal of Medicine,* 355, 549-59.

THE TIRILAZAD INTERNATIONAL STEERING COMMITTEE (2002) Tirilazad for acute ischaemic stroke (Cochrane Review). 3 ed. Oxford, Update Software.

THOMASSEN, L., WAJE-ANDREASSEN, U., NAESS, H., ELVIK, M.-K. & RUSSELL, D. (2005) Long term effect of intravenous thrombolytic therapy in acute stroke: responder analysis versus uniform analysis of excellent outcome. *Cerebrovascular Diseases,* 20, 470-474.

THORVALDSEN, P., ASPLUND, K., KUULASMAA, K., RAJAKANGAS, A. M., SCHROLL, M. & FOR THE WHO MONICA PROJECT (1995) Stroke incidence, case fatality, and mortality in the WHO MONICA Project. *Stroke,* 26, 361-367.

TILLEY, B. C., MARLER, J., GELLER, N. L., LU, M., LEGLER, J., BROTT, T., LYDEN, P., GROTTA, J. (1996) Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA stroke trial. *Stroke,* 27, 2136-2142.

VAHEDI, K., HOFMEIJER, J., JUETTLER, E., VICAUT, E., GEORGE, B., ALGRA, A., AMELINK, G. J., SCHMIEDECK, P., SCHWAB, S., ROTHWELL, P., BOUSSER, M.-G., VAN DE WORP, H. B., HACKE, W. & FOR THE DECIMAL DESTINY AND HAMLET INVESTIGATORS (2007) Early decompressive surgery in malignant infarction of the middle cerebral artery: a pooled analysis of three randomised controlled trials. *Lancet Neurology,* 6, 215-222.

VAN SWIETEN, J. C., KOUDSTAAL, P. J., VISSER, M. C., SCHOUTEN, H. J. A., VAN GIJN, J. & 1519 (1988) Interobserver agreement for the assessment of handicap in stroke patients. *Stroke,* 19, 604-607.

VITTINGHOFF, E. & MCCULLOCH, C. E. (2006) Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology,* 165, 710-718.

VUADENS, P. & BOGOUSSLAVSKY, J. (1998) Diagnosis as a guide to stroke therapy. *Lancet,* 352, 5-9.

WADE, D. T. (1992) *Measurement in neurological rehabilitation,* Oxford, Oxford University Press.

WADE, D. T. & LANGTON HEWER, R. (1987) Functional abilities after stroke: measurement, natural history and prognosis. *Journal of Neurology, Neurosurgery, and psychiatry,* 50, 177-182.

WAHLGREN, N. G., MACMAHON, D. G., DE KEYSER, J., INDREDAVIK, B., RYMAN, T. & INWEST STUDY GROUP (1994) Intravenous Nimodipine West European Stroke Trial (INWEST) of nimodipine in the treatment of acute ischaemic stroke. *Cerebrovascular Diseases,* 4, 204-210.

WALKER, M., GLADMAN, J., LINCOLN, N., SIEMONSMA, P., WHITELEY, T. & 1996 (1999) Occupational therapy for stroke patients not admitted to hospital: a randomised controlled trial. *The Lancet,* 354, 278-280.

WALKER, M. F., DRUMMOND, A. E. R. & LINCOLN, N. B. (1996) Evaluation of dressing practice for stroke patients after discharge from hospital: a crossover design study. *Clinical Rehabilitation,* 10, 23-31.

WALKER, M. F., LEONARDI-BEE, J., BATH, P., LANGHORNE, P., CORR, S., DRUMMOND, A., GILBERTSON, L., GLADMAN, J. R. F., JONGBLOED, L. & PARKER, C. (2004) An individual patient data meta-analysis of randomised controlled trials of community occupational therapy for stroke patients. *Stroke,* 35, 2226-2232.

WALTER, S. D. (2001) Number needed to treat (NNT): estimation of a measure of clinical benefit. *Statistics in Medicine,* 20, 3947-3962.

WALTERS, S. J. (2004) Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36. *Health and Quality of Life Outcomes,* 2.

WALTERS, S. J. & CAMPBELL, M. J. (2005) The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes. *Statistics in Medicine,* 24, 1075-1102.

WALTERS, S. J., CAMPBELL, M. J. & LALL, R. (2001) Design and analysis of trials with quality of life as on outcome: a practical guide. *Journal of Biopharmaceutical statistics,* 11, 155-176.

WARACH, S., PETTIGREW, C., DASHE, J. F., PULLICINO, P., LEFKOWITZ, D. M., SABOUNJIAN, L., HARNETT, D., SCHWIDERSKI, U., GAMMANS, R. & CITICOLINE 010 INVESTIGATORS (2000) Effect of citicoline on ischemic lesions as measured by diffusion-weighted magnetic resonance imaging. *Annals of Neurology,* 48, 713-721.

WARDLAW, J. M., SANDERCOCK, P. A. G., WARLOW, C. P. & LINDLEY, R. I. (2000) Trials of thrombolysis in acute ischemic stroke. Does the choice of primary outcome measure really matter? *Stroke,* 31, 1133-1135.

WARE, J. E., SNOW, K. K., KOSINSKI, M. & GANDEK, B. (1993) *SF-36 Health survey manual and interpretation guide,* Boston, MA:New England Medical Centre, The Health Institute.

WARLOW, C. (1998) Epidemiology of stroke. *Lancet,* 352, 1-4.

WARLOW, C. (2002) Advanced issues in the design and conduct of randomized clinical trials: the bigger the better? *Statistics in Medicine,* 21, 2797-2805.

WARLOW, C. P., DENNIS, M. S., VAN GIJN, J., HANKEY, G. J., SANDERCOCK, P. A. G., BAMFORD, J. M. & WARDLAW, J. (1996) *Stroke. A practical guide to management,* Oxford, Blackwell Science.

WARNER GARGANO, J., WEHNER, S. & REEVES, M. (2008) Sex differences in acute stroke care in a statewide stroke registry. *Stroke,* 39, 24-29.

WEAVER, C. S., LEONARDI-BEE, J., BATH-HEXALL, F. J. & BATH, P. M. W. (2004) Sample size calculations in acute stroke trials: A systematic review of their reporting, characteristics, and relationship with outcome. *Stroke,* 35, 1216-1224.

WEIR, C. J., BRADFORD, A. P. J. & LEES, K. R. (2003) The prognostic value of the components of the Glasgow Coma Scale following acute stroke. *QJM: An International Journal of Medicine,* 96, 67-74.

WEIR, C. J. & LEES, K. R. (2003) Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Statistics in Medicine.,* 22, 705-726.

WHITEHEAD, A., OMAR, R. Z., HIGGINS, J. P. T., SAVALUNY, E., TURNER, R. M. & THOMPSON, S. G. (2001) Meta-analysis of ordinal outcomes using individual patient data. *Statistics in Medicine,* 20, 2243-2260.

WHITEHEAD, J. (1993) Sample-Size Calculations for Ordered Categorical-Data *Statistics in Medicine,* 12, 2257-2271.

WHO MONICA PROJECT PRINCIPAL INVESTIGATORS (1988) The World Heath Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): A major international collaboration. *Journal of Clinical Epidemiology,* 41, 105-114.

WILCOXON, F. (1945) Individual comparisons by ranking methods. *Biometrics,* 1, 80-83.

WILKIN, D., HALLAM, L. & DOGGETT, M. A. (1993) *Measures of need and outcome for primary health care,* Oxford, Oxford University Press.

WILSON, J. T. L., HAREENDRAN, A., GRANT, M., BAIRD, T., SCHULZ, U. G. R., MUIR, K. W. & BONE, I. (2002) Improving the assessment of outcomes in stroke. Use of a structured interview to assign grades on the modified Rankin scale. *Stroke,* 33, 2243-2246.

WOLF, P. A. (1998) Prevention of stroke. *Lancet,* 352, 15-18.

WOLFE, C. D. A. (2000) The impact of stroke. *British Medical Bulletin,* 56, 275-286.

WOLFE, C. D. A., RUDD, A. G., HOWARD, R., COSHALL, C., STEWART, J., LAWRENCE, E., HAJAT, C. & HILLEN, T. (2002) Incidence and case fatality rates of stroke subtypes in a multiethnic population: the South London Stroke Register. *Journal of Neurology, Neurosurgery, and Psychiatry,* 72, 211-216.

WONG, K. S., CHEN, C., NG, P. W., TSOI, T. H., LAU, K. K., YEUNG, J., WONG, C. K., CHANG, C. M. & FISS-TRIS STUDY GROUP (2005) A randomized controlled study of low molecular weight heparin versus aspirin for the treatment of acute ischaemic stroke in patients with large artery occlusive disease. IN TRIALS, N. C. (Ed. *14th European Stroke Conference.* Bologna, Italy.

WORLD HEALTH ORGANISATION (2001) International Classification of Functioning, Disability and Health *World Health Organisation*.

WORLD HEALTH ORGANIZATION (1980) *International classification of impairments, disabilities and handicaps,* Geneva, WHO.

WREN, B. G. (1998) Megatrials of Hormonal Replacement Therapy. *Drugs & Aging,* 12, 343-348.

YAMAGUCHI, T., SANO, K., TAKAKURA, K., SAITO, I., SHINOHARA, Y., ASANO, T. & YASUHARA, H. (1998) Ebselen in acute ischemic stroke. *Stroke,* 29, 12-17.

YEO, D., FALEIRO, R. & LINCOLN, N. B. (1995) Barthel ADL Index: a comparison of administration methods. *Clinical Rehabilitation,* 9, 34-9.

YOUNG, F. B., LEES, K. R., WEIR, C. J. & FOR THE GAIN INTERNATIONAL STEERING COMMITTEE AND INVESTIGATORS (2005) Improving trial power through use of prognosis-adjusted end points. *Stroke,* 36, 597-601.

YOUNG, F. B., LEES, K. R., WEIR, C. J. & FOR THE GLYCERINE ANTAGONINST IN NEUROPROTECTION INTERNATIONAL TRIAL STEERING COMMITEE AND INVESTIGATORS (2003) Strengthening acute stroke trials through optimal use of disability end points. *Stroke,* 34, 2676-2680.

YOUNG, J. B. & FORSTER, A. (1992) The Bradford community stroke trial: results at six months. *British Medical Journal,* 304, 1085-9.

ZEC, R. F. & TRIVEDI, M. A. (2002) Effects of hormone replacement therapy on cognitive aging and dementia risk in postmenopausal women: a review of ongoing large-scale, long-term clinical trials. *Climacteric,* 5, 122-134.