Improving efficiency in Stroke Trials:

An exploration of methods to improve the use of the modified Rankin Scale in acute stroke trials

Kate Susan McArthur

MBChB, BSc (Med Sci), MRCP (Glasg)

Submitted in fulfilment of the requirements for the Degree

Of Doctor of Medicine (MD)

Institute of Cardiovascular and Medical Sciences

College of Medical, Veterinary and Life Sciences

University of Glasgow

December 2013

Abstract

The modified Rankin Scale (mRS) is the preferred outcome measure in stroke trials. Typically, mRS assessment is based on a clinician's rating of a patient interview and interobserver variability is common. Meta-analysis suggests an overall reliability of κ=0.46 but this may be less (κ=0.25) in multi-centre studies. Mandatory training in mRS assessment is employed in most trials to mitigate this but the problem persists. Variability in assigning outcomes may lead to endpoint misclassification increasing the challenge of accurately demonstrating a treatment effect. We aimed to assess the impact of endpoint misclassification on trial power and explore methods to improve the use of the mRS in acute stroke trials.

First we used the mRS outcome distributions of previous phase III randomised controlled trials (RCT) in stroke (NXY059 study and tPA NINDS study) to perform statistical simulations. We generated power estimates and sample sizes from simulated mRS studies under various combinations of sample size, mRS reliability and adjudication panel size. Simulations suggest that the potential benefit of improving mRS reliability from k 0.25 to k 0.5, k 0.7 or k 0.9 may allow a reduction in sample size of n= 386, n= 490 or n= 488 in a typical n=2000 RCT.

We then developed a method for providing group adjudication of mRS endpoints and examined the feasibility, reliability and validity of its use in a multicentre clinical trial. We conducted a "virtual" acute stroke trial across 14 UK sites. Local mRS interviews were scored as normal but also recorded to digital video camera. Video clips were uploaded via secure web portal for scoring by adjudication committee reviewers. We demonstrated excellent technical success rates with acceptability to both participants and investigators. 370 participants were included in our "virtual" acute stroke trial and 563 mRS video assessments were uploaded for central review. 96% (538/563) of study visits resulted in an adjudicated mRS score. At 30 and 90 days respectively, 57.5% (161/280) and 50.8% (131/258) of clips were misclassified. Agreement was measured using kappa statistics (κ/κ_w) and intraclass correlation coefficient. Agreement between the adjudication committee was very good (30

days κ_w 0.85 [95%CI 0.81-0.86], 90 days κ_w 0.86 [95% CI 0.82-0.88]) with no significant or systematic bias in mRS scoring in comparison to the local mRS. We demonstrated criterion and construct validity of centrally adjudicated mRS scores through comparison with the locally assigned mRS score and other measures known to affect stroke outcome including baseline NIHSS (bNIHSS), Systolic Blood Pressure (SBP), blood glucose and home time.

We studied our cohort of mRS video clips to identify any features predictive of variability in mRS scoring. Patient specific variables included participant age, pre stroke mRS, baseline stroke severity as graded by baseline NIHSS (bNIHSS) and presence of language disorder. Interview specific variables included length of interview, poor sound quality, location of the interview, use of a proxy or discussion of prior disability. At both 30 and 90 days only "interview length" was a significant predictor of agreement in mRS scoring.

Using a sample of mRS video clips in English and Mandarin, we conducted a pilot study to assess the effect of translation of mRS interviews on interobserver reliability. The interobserver reliability of the translated mRS assessments was similar to native language clips (Native (n=69) κ_w 0.91 [95%CI 0.86-0.99], Translated (n=89) κ_w 0.90 [95% CI 0.83-0.96]). We then incorporated a translation step into the central adjudication model using our existing web portal. Inter observer reliability seen in the modified clips (κ_w 0.85 [95% CI 0.74-0.95]) was similar to that seen in the original video files (κ_w 0.88 [95% CI 0.78-0.99]).

Finally we aimed to investigate the ability of raters to detect more subtle degrees of disability within mRS ranks through blinded assessment of pairs of clips with matching mRS grades. These pairs contained either two clips with full agreement in mRS grade at initial group review or one clip with full agreement and one clip where scores were skewed in the direction of "more" or "less" disability. Pairs were randomly assigned to multiple raters. We could not identify any reliable pattern in identification of the "less disabled" mRS clip. More sensitive grading of the mRS with "good" or "bad" forms of each grade is not reliable on the basis of this exploratory study. Perhaps alternative methods of converting the ordinal ranks of the mRS scale into a more continuous distribution should be investigated; such as the use of a mean mRS score following multiple mRS ratings.

Prior estimates of mRS reliability in multicentre studies are poor [κ=0.25]. The risks of endpoint misclassification affecting trial power are substantial. Simulations suggest that the effect of improving interobserver reliability and multiple mRS assessments may reduce study sample size by 25%, resulting in substantial ethical and financial benefits. Agreement between our adjudication committee was good [κ=0.59(95% CI:0.53-0.63), κ_w=0.86(95% CI:0.82-0.88)]. Central review may bring many additional potential benefits: “expert” review, quality control and improved blinding in complex trial design.

Central adjudication of mRS assessments is feasible, reliable and valid, including the use of translated mRS assessments. This model of outcome assessment has been incorporated into four ongoing large clinical trials: CLEAR-3, MISTIE-3, EUROHYP-1 and SITS-OPEN.

"When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science, whatever the matter may be."

Lord Kelvin, 1883

The University of Glasgow

Acknowledgements

I would like to thank my principal supervisor, Professor Kennedy Lees for the opportunity to undertake this research. His knowledge, enthusiasm, vision, guidance and support have been inspirational and invaluable. I would also like to thank my second supervisor Professor Matthew Walters for his support and advice. Without their time and encouragement this thesis would not have been possible.

I am indebted to the many people who in some way contributed to the work contained in this thesis, particularly the all of the CARS investigators who provided the participants and videos. To Professor Kennedy Lees, Professor Matthew Walters, Professor Peter Langhorne, Dr Jesse Dawson, Dr Terry Quinn and Dr Peter Higgins who expended time and energy scoring more Rankin assessments than I can count; without your time none of this work would have been possible and I am truly grateful. To Dr Paul Johnson, Dr Chris Weir and Dr Rachael Fulton who gave their statistical expertise and advice, I couldn't have completed any of the analysis without their guidance and patience. I would like to thank the Chief Scientist Office for their financial support in funding this project.

I would like to thank a number of members of staff from the Stroke Unit and Stroke Research Group at the Western Infirmary, Glasgow who have given me time, support and friendship since I began my post as clinical research fellow: Pamela MacKenzie, Belinda Manak, Elizabeth Colquhoun, Lesley MacDonald and Karen Shields.

Finally, to my family and friends who have given me so much encouragement and support, particularly during the writing phase. To Ewan, who began as boyfriend and has become husband and father during this period of research, I am grateful for your patience and unwavering support. To Liz and Moira, two wonderful Grannies who provided countless hours of childcare to help me achieve this thesis. Finally to Isla, this thesis is the reason that Mummy was at "work" so much, I hope you think it was worth it.

Declaration

The work described in this thesis was performed during my period as a Clinical Research Fellow in the University Institute of Cardiovascular and Medical Sciences at the Western Infirmary, Glasgow (August 2009 – August 2011).

I declare that I am the sole author of this thesis entitled "Improving efficiency in stroke trials: An exploration of methods to improve the use of the modified Rankin Scale in acute stroke trials". The work contained within this thesis has been the result of successful collaboration with a number of colleagues who are formally acknowledged below. This work has never been submitted previously for a higher degree.

Recruitment of Participants and completion of outcome assessment visits

The CARS investigators: See Appendix D

Assessment of mRS Videos

Review and scoring of mRS videos was undertaken by myself and the following colleagues; without their time and patience none of the work contained in this thesis would have been possible.

Chapter 4-6: CARS Adjudication Committee. Professor Kennedy R Lees, Professor Matthew R Walters, Professor Peter Langhorne, Dr Terence J Quinn, Dr Jesse Dawson, Dr Peter Higgins.

Chapter 7: Predictors of Variability in mRS. Margaret Yuan (Medical Student)

Chapter 8: mRS Translation Pilot. Professor Y Huang, Dr H Xing, Dr Wei Sun, Dr Weiping Sun and Dr Q Peng (Beijing, China). Professor Kennedy R Lees, Professor Matthew R Walters, Dr Terence J Quinn, Dr Jesse Dawson (Glasgow, UK). CARS Translation Sub Study. Professor Kennedy R Lees, Professor Matthew R Walters, Dr Terence J Quinn, Dr Jesse Dawson, Dr Peter Higgins, Dr Senthil Raghunathan, Dr Mary Joan MacLeod, Mrs Belinda Manak, Mrs

Lesley MacDonald, Miss Elizabeth Colquhoun, Mrs Angela Welsh, Mr Michael Keeling, Mrs Jacqueline Strover.

Chapter 9: mRS Pairs Study. Professor Kennedy Lees, Professor Matthew R Walters, Dr Terence J Quinn, Dr Jesse Dawson, Dr Peter Higgins, Mrs Belinda Manak, Mrs Lesley MacDonald, Miss Elizabeth Colquhoun, Miss Margaret Yuan.

Statistical Analysis

Chapter 2: The complex statistical modelling was performed by Dr Paul Johnson, previously of the Robertson Centre for Biostatistics, University of Glasgow with the guidance of Dr Chris Weir, MRC Hub for Trials Methodology Research, University of Edinburgh.

Chapter 5: Inter-observer reliability analyses and the statistical analysis of measurement error between observers were performed by Dr Paul Johnson, previously of the Robertson Centre for Biostatistics, University of Glasgow with the guidance of Dr Chris Weir, MRC Hub for Trials Methodology Research, University of Edinburgh.

Chapter 6 and 7: Ordinal Logistic Regression Analysis was performed with guidance from Dr Rachael Fulton, Institute of Cardiovascular and Medical Sciences, University of Glasgow.

Relevant Publications and Presentations

McArthur KS, Johnson PC, Quinn TJ, Higgins, P, Langhorne P, Walter MR, Weir CJ, Dawson J, Lees KR. Improving the Efficiency of Stroke Trials: Feasibility and Efficacy of Group Adjudication of Functional End Points. Stroke. 2013; doi: 10.1161/STROKEAHA.113.002266

Platform Presentations

McArthur KS, Shi YM, Fulton RL, Quinn TJ, Higgins P, Dawson J, Langhorne P, Walters MR, Lees KR on behalf of the CARS investigators. Factors associated with Inter-Observer Variability in the modified Rankin Scale (mRS) Assessment. European Stroke Conference (Nice) 2014

McArthur KS, Fulton RL, Quinn TJ, Higgins P, Dawson J, Langhorne P, Walters MR, Lees KR. Ranking withing Rankin: Can disability be graded within modified Rankin Scale (mRS) grades? European Stroke Conference (Nice) 2014

McArthur KS, Dawson J, Quinn TJ, Higgins P, Johnson P, Weir C, Langhorne P, Walters MR, Lees KR. Central Adjudication of modified Rankin Scale disability assessments in acute stroke trials. UK Stroke Forum (Glasgow). 2011

McArthur KS, Quinn TJ, Johnson PCD, Dawson J, Walters MR, Weir CJ, Lees KR. Beneficial effect of improving modified Rankin scale reliability on sample size for stroke trials. European Stroke Conference (Barcelona). 2010

Poster Presentations

McArthur KS, Xing H, Dawson J, Quinn TJ, Higgins P, Langhorne P, Walters MR, Yuang H, Lees KR. Translation and Central Adjudication of modified Rankin Scale assessments in acute stroke trials is feasible and reliable. European Stroke Conference (London). 2013.

Winner of the Young Investigator 2013 (European Stroke Conference)

McArthur KS, Xing H, Quinn T, Dawson J, Walters MR, Sun W, Peng Q, Higgins P, Huang Y, Lees KR. Reliabiltiy and validity of a translated modified Rankin scale assessment – A pilot study in Mandarin and English. International Stroke Conference (Los Angeles) 2011.

List of Abbreviations

AbESTT II	A Study of Effectiveness and Safety of Abciximab in Patients With Acute Ischemic Stroke
ADL	Activities of daily living
AF	Atrial Fibrillation
AVI	Audio Video Interleave (video file format)
AVS	AVS Video Converter Software©
BI	Barthel Index
bNIHSS	Baseline National Institutes of Health Stroke Score
CARS	Central Adjudication of Modified Rankin Scales in Acute Stroke Trial
CCF	Congestive Cardiac Failure
CD	Compact Disc
CI	Confidence Interval
CLASS	Clomethiazole Acute Stroke Study
CLEAR	Clot Lysis Evaluating Accelerated Resolution of Intraventricular Haemorrhage
CMH	Cochran Mantel Haenszel
CT	Computed Tomography
DALY	Disability Adjusted Life Years

DESTINY	Decompressive Surgery for the Treatment of Malignant Infarction of the Middle Cerebral Artery
DVD	Digital Versatile Disc
DVT	Deep Venous Thrombosis
ECASS	European Cooperative Acute Stroke Study
eCRF	Electronic Case Report Form
EMEA	European Medicines Agency
EuroHyp	European Hypothermia Trial
FAST-MAG	Field Administration of Stroke Therapy – Magnesium Trial
FDA	Food and Drug Administration
GAIN	Glycine Antagonist in Neuroprotection for Acute Stroke Study
GCP	Good Clinical Practice
GCS	Glasgow Coma Scale
GOS	Glasgow Outcome Scale
GOSE	Glasgow Outcome Scale (Extended)
ICC	Intraclass Correlation Coefficient
ICD	International Classification of Diseases
IMAGES	Intravenous Magnesium Efficacy in Acute Stroke Trial
IQR	Interquartile Range
IST	International Stroke Trial

IT	Information Technology
κ	Kappa statistic
κ_w	Quadratically weighted kappa statistic
KB	Kilobyte
LACS	Lacunar Infarct / Lacunar Stroke
LMWH	Low Molecular Weight Heparin
LOC	Loss of Consciousness
MB	Megabyte
MI	Myocardial Infarction
MISTIE	Minimally Invasive Surgery plus tPA for Intracerebral Haemorrhage Evacuation Trial
mNIHSS	Modified National Institutes of Health Stroke Scale
MPEG	Moving Picture Experts Group (video file format)
MR / MRI	Magnetic Resonance Imaging
mRS	modified Rankin Scale
MSE	Mean Squared Error
MTS	MPEG Transport Stream (video file format)
NHS	National Health Service
NIHSS	National Institutes of Health Stroke Scale
NINDS	National Institutes of Neurological Disorders and Stroke

NNH	Number Needed to Harm
NNT	Number Needed to Treat
NXY059	NXY059 for Acute Ischaemic Stroke Trial
OHS	Oxford Handicap Scale
OR	Odds Ratio
PACS	Partial Anterior Circulation Infarct
PC	Personal Computer
POCS	Posterior Circulation Infarct
PROBE	Prospective Randomised Open Blinded Endpoint Study
PTE	Pulmonary Thromboembolism
QALY	Quality Adjusted Life Year
RCB	Roberston Centre for Biostatistics, University of Glasgow
RCT	Randomised Controlled Trial
REC	Research Ethics Committee
REML	Restricted Maximal Likelihood Model
RFA	Rankin Focussed Assessment
ROC	Receiver Operating Characteristics
RS	Rankin Scale
SAE	Serious Adverse Event
SAINT	NXY059 for Acute Ischaemic Stroke Trial

SBP	Systolic Blood Pressure
SD	Standard Deviation
SE	Standard Error
SIS	Stroke Impact Scale
smRSq	Simplified modified Rankin Scale Questionnaire
STITCH	Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial.
TACS	Total Anterior Circulation Infarct
TIA	Transient Ischaemic Attack
TOAST	Trial of Org 10172 in Acute Stroke Treatment, Classification of Ischaemic Stroke
tPA	Tissue Plasminogen Activator
UK	United Kingdom
UK-TIA	United Kingdom – Transient Ischaemic Attack Study
UN	United Nations
US / USA	United States (of America)
USB	Universal Standard Bus
USS	Ultrasound Scan
VISTA	Virtual International Stroke Archive

WHO	World Health Organisation
WHO-GBD	World Health Organisation – Global Burden of Disease
WHO-ICF	World Health Organisation – International Classification of Disease
WMV	Windows Media File (video file format)

Contents

List of Figures

List of Tables

Chapter 1

Background and Introduction

1.1. Introduction

Modern medicine must rely upon robust evidence to verify the safety, quality and efficacy of new therapeutic strategies. The burden of delivering evidence lies with the scientists and physicians who believe that their hypotheses may translate into new treatments with potential to transform patient outcomes. This journey, from theory to change in clinical practice is challenging; academically, logistically and financially.

Whether the aim is to investigate an entirely novel intervention or to use an existing drug in a new clinical context the process of gathering evidence must be meticulously planned, closely regulated and peer reviewed. From early laboratory work involving animal or tissue samples to large phase III randomised controlled trials in the target patient population this process is necessary to deliver a change in clinical practice that provides safer, more effective and more accessible therapies.

The time and money invested by academic and commercial institutions in clinical research is substantial. Only a tiny percentage of novel compounds make it to market, many being abandoned in the early stages of development. It is estimated that discovery and development of each new pharmaceutical agent costs an average of $800 million US dollars and takes between nine and twelve years to gain the necessary regulatory approvals[1]. These compounds must generate enough income to cover their own development costs and recover the expenditure lost in early investigation of abandoned compounds. The financial burden continues to rise, attributable to a multitude of factors including increasing

regulatory requirements, complex trial design requiring large sample sizes, training and initiation of multiple international sites and increasing difficulty recruiting participants to clinical trials.

To maintain the important advances in evidence based medicine we have achieved in recent decades we must find a way to make clinical research more efficient. This is a particular challenge in the field of acute stroke due to the heterogeneous nature of its aetiology, clinical presentation and possible outcomes.

1.2. Evidence Based Medicine in Acute Stroke

Acute stroke is a major cause of death and disability in the developed world[2, 3] and its consequences place a substantial burden on healthcare resources and economic productivity[4]. The number and quality of clinical trials in the field of acute stroke has increased considerably in recent decades despite an acknowledgement that the funding available to stroke research is significantly less than other disease states such as cancer and coronary heart disease[5]. Improved understanding of the basic pathophysiological processes in stroke has led to marked changes in contemporary trial design, methodology and statistical analysis techniques.

1.2.1. Trial Design

Where the investigational product or procedure allows, a model phase III clinical trial should include several components in its design to ensure that results are valid, robust and reliable. These generic requirements are applicable to any field of medicine. The primary objective of the trial and its hypothesis must be clearly and prospectively stated, together with any proposed subgroup analyses. Acknowledging that there are likely to be systematic, cultural and demographic variables, where practical, studies should be multicentre to ensure that results are reproducible and generalisable internationally. Participants must be carefully selected according to predefined criteria (inclusion / exclusion) and followed up prospectively at specific, standardised time points. Treatment allocation should be randomly allocated and blinded to both participant and investigator (double blind) until the point of trial closure and data analysis. The number of participants necessary to demonstrate a

realistic and clinically relevant treatment effect must be determined in advance through the use of an appropriate power calculation. The outcome measure must be relevant, appropriate, valid and reliable. The statistical analysis plan must be developed and documented prior to data collection to avoid retrospective data manipulation.

1.2.2. Early improvements in acute stroke trial design

A 20 fold increase in the number of stroke trials was seen between 1950 and 1999. The total number of participants enrolled has increased markedly; mean (median) sample size in 1950s research was 38 (26) in comparison to 661 (113) in the 1990s. The integration of each of the "ideal" trial components in published stroke research in the second half of the twentieth century has gradually improved[6]. Prior to 1970 no published studies were multicentre, by the 1990's 68% met this standard. The proportion of published studies which were randomised (75% to 99%) and double blind (38% to 83%) also improved in the same time period. In the first decade of the 21st century more than 125 acute stroke trials successfully provided evidence to change practice (thrombolytic therapy in an extended time window, hemicraniectomy in select patients with malignant infarction and coiling in subarachnoid haemorrhage secondary to aneurysmal bleeds)[7].

Additional changes in trial design specific to stroke medicine are also notable. Our understanding of the underlying pathophysiology of stroke disease has improved participant selection. The clinical syndromes of stroke disease are well recognised and understood; however, the pathophysiological basis for each event may be quite different. Stroke subtype and severity are important prognostic factors relevant to trial design and selection of participants[8]. Recruitment criteria may often predefine a severity restriction to prevent the inclusion of participants who might attenuate the detection of a treatment effect in other subgroups. For example inclusion of either a very severely affected participant who is unlikely to survive or a very mildly affected participant who is almost certain to recover well is not informative.

The timing of participant recruitment in acute stroke trials is crucial and time windows have progressively decreased[6]. The majority of interventional or pharmacological treatments aim

either to recanalise occluded vessels thus reperfusing damaged tissue or to protect vulnerable brain tissue at risk of permanent damage. Following brain ischaemia there is a complex cascade of events culminating in complete tissue death. The time window for salvage of vulnerable tissue, known as penumbra, is short and subsequently very early rescue treatment is likely to yield greater clinical benefit. For this reason the time window from stroke onset to enrolment in stroke trials should be limited and tailored to the hypothesis of each individual agent.

1.2.3. Lessons from previous Acute Stroke Trials

Several large acute stroke trial programmes have been conducted with negative results, the lessons from which we can use in the planning and design of future studies. The early ECASS studies (I[9] and II[10]) were underpowered and perhaps chose an unfortunate combination of drug dose, timing and endpoints without the benefit of current knowledge. Following publication of the NINDS tPA stroke study, thrombolysis was considered safe and effective, a study with more complex design[11]. The Lubeluzole trials were conceived to demonstrate an effect on mortality seen in early studies[12] that was not reproduced in later trials[13, 14]. A negative trial result with a trend towards significance in certain groups has previously prompted trialists to chase post hoc analysis, such as in the CLASS study programme[15, 16]. Consistent efforts were made to improve trial design in the last decades of the twentieth century. Landmark neuroprotectant studies (GAIN[17] and SAINT[18] trials) pioneered best practice in terms of trial design, sample size and statistical approaches but unfortunately the choice of drug and pre-clinical studies were flawed. Novel compounds showed promise in initial pre-clinical studies but the translation of these results into large scale phase three clinical trials was not possible, leading to negative results. The responsible factors are likely to be multifactorial, however lessons can be learned from imperfect methodologies to optimise future trial design.

1.2.3.1. Translation of animal models to clinical studies

It is very likely that the initial promise shown in pre-clinical studies using animal models provided ambitious and inaccurate estimates of the expected treatment effect of

neuroprotectant agents. Often the number of animals used in each study is small, encouraging meta-analysis of published studies to generate an estimate of treatment effect. There is substantial publication bias in preclinical studies. Any publication bias may lead to fundamental flaws in the generation of this estimated treatment effect, rendering it unrealistically high. The impact of this error is critical to the planning and design of the next phase of clinical research; the potential to introduce considerable error in power and sample size calculations for clinical studies is large.

The relevance of animal model findings to clinical stroke in human subjects is questionable[19]. Investigating a new compound in animal models allows standardisation of many factors which clinical trialists can never hope to equal. The heterogeneity and complexity of stroke in human subjects is incomparable to animal infarct models where stroke location, severity and infarct size can be closely controlled. Time to treatment in human subjects is limited by patient presentation and clinical services, where in animal studies drug delivery is precisely timed and progressively delayed until a beneficial effect is no longer seen. Control of confounding factors that may affect outcome is possible in pre-clinical studies, both environmental (temperature / blood pressure) and physiological (sex / age / co-morbid illness).

One opinion of how to optimise the translation of laboratory experiments to the bedside is to try and better emulate the animal experiment in our patient population[20]. However, to match a clinical trial to the animal model would necessitate tighter standardisation of participants enrolled and shorten the time to treatment, potentially resulting in unfeasibly slow participant enrolment and limiting the generalisability of results. It has also been advocated that initial trials in animal models should instead be designed to emulate clinical trials: multicentre, with randomisation, blinding and central review of results where appropriate. In either case it is important to acknowledge the limitations in matching animal experiments to human subjects when designing clinical studies.

1.2.3.2. Trial design errors

With the value of hindsight, the design of several large neuroprotectant trials led to almost inevitable failure. In the early 21st century stroke academics collectively reassessed the design issues that may have resulted in this disappointment[21].

Several common defects were identified, many of which arose from inaccurate translation of the animal studies to trial design as discussed above. The patient population was often unhelpful, for example enrolment of participants with lacunar infarcts (affecting purely white matter) in the study of drugs that are known to be active only at grey matter synaptic terminals. Small sample sizes were known to be insensitive to detection of the modest treatment effect expected. The timing of drug administration and dosages were considered too late and too small, leading to inadequate drug delivery to the penumbral tissue. Finally the choice of trial endpoint and statistical analytical techniques were criticised.

1.3. Outcome measures in stroke trials

The choice of outcome measure is fundamental to the design of any clinical trial. The objective in a randomised controlled clinical trial is to assess the effect of an intervention or treatment, either positive or negative. In order to do this a method of measuring and documenting the outcome of participants must be chosen that is relevant to the physician, the patient and the disease process.

In many areas of clinical medicine the appropriate outcome measure is clear. For example, in a trial of a new antihypertensive agent the trialists might choose to use blood pressure measurements or vascular events; in a study examining the effect of a new chemotherapy agent mortality rate or periods of remission may be the obvious choice. These “hard” endpoints are easy to conceptualise and relatively straightforward to measure and record.

Unfortunately, stroke disease is a more difficult and complex area to study. Outcomes are variable, ranging from complete recovery to various degrees of disability and death. Death is typically considered the worst possible outcome in clinical trials; this is not necessarily the case in stroke medicine. To keep patients alive at the expense of severe disability and loss of

quality of life would be considered by many to be a worse outcome[22]. Stroke trials now favour a measure of functional recovery as the primary endpoint. While a functional assessment is theoretically attractive, in practice it adds a layer of complexity to the trial. Functional outcomes are varied and subjective and it can be difficult to reduce the complex qualitative experience of stroke survival to a numerical value. Measurement and documentation of these more subjective outcomes is much more problematic and can severely compromise clinical trials where assessment is not reliable, robust and reproducible.

1.3.1. Early Stroke Outcome Research

Accurate and meaningful assessment of patients' progress after a stroke event is important in the context of an interventional stroke trial but also in observational research and registry data collection. Measurement of outcome is necessary to monitor the effect of stroke disease on the wider population, to allow an educated estimate of likely prognosis and to enable rational health care planning at a population level. This was recognised in the 1980s and began a movement to study stroke outcome measures as an entity[23].

The complex nature of stroke and heterogeneity of patient outcomes was noted as a challenge to creation of the "ideal" stroke outcome measure[24, 25]. Recognition of the components of a useful stroke outcome measure were detailed in a 1990 task force statement[26]. Broad recommendations to be considered in the study of stroke outcome measures were detailed: 1) Outcome must be measured in conjunction with the time from ictus, 2) The site and side of the stroke lesion should be specified, 3) Imaging should be used to classify stroke events and 4) outcome studies should be limited to patients with a first stroke in recognition of the introduction of confounding by prior disability. It was recommended that instruments must be simple to apply in a short space of time, encompass a description of activities of daily living (ADLs) together with an assessment of cognitive function, speech and communication, emotional wellbeing and social functioning[27].

These general principles have become the basis of a large body of literature pertaining to the use of various stroke outcome measures. In the past 20 years terminology has been

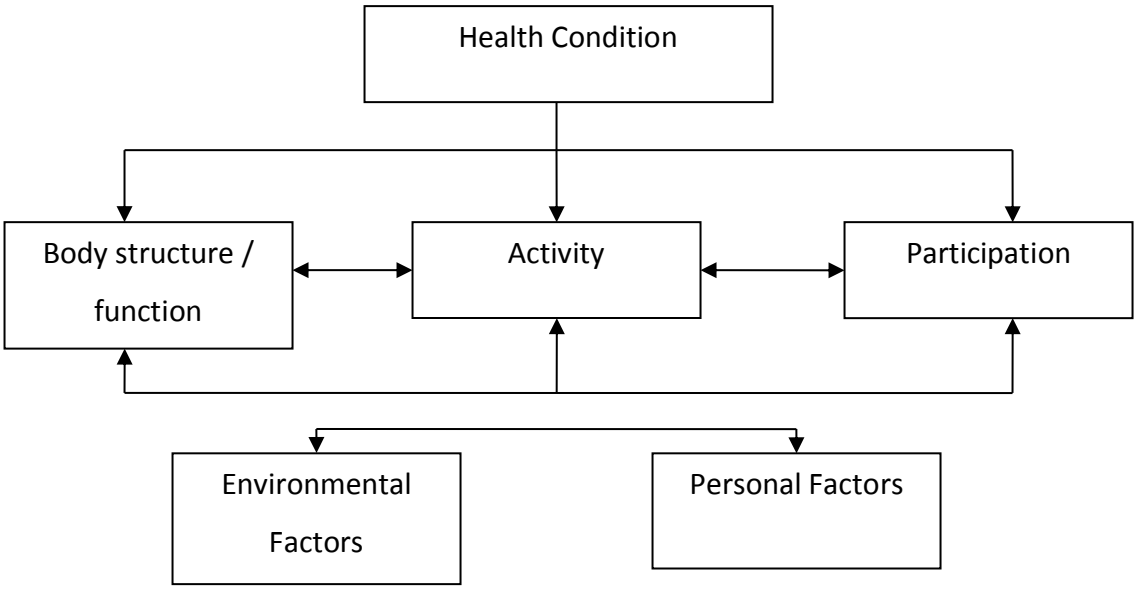
standardised and clinimetric properties examined to help guide the use of outcome scales in trial design.

1.3.2. Functional Outcome

Within the realms of "functional outcome" there is varied terminology to describe the degree to which residual symptoms affect a patient. The terminology has been regularly updated; the most recent WHO framework for describing function and limitation is found in the WHO International Classification of Functioning, Disability and Health (ICF). This is a multidimensional framework in which three levels of human functioning are described. – the body (or body part), the whole person and the whole person in the context of his/her position and role in society (Figure 1).

An outcome measure may assess function at any of these three levels: *problems in body structure/function* (formerly impairment) - signs of an underlying pathological process, *functional activity* (formerly disability) - a limitation in execution of a particular task as a result of this problem or *societal participation* (formerly handicap) - the social effect of that impairment in overall quality of life and social role[28].

There are multiple stroke outcome scales designed to measure function at each of these levels. A sample of the extensive list of outcome measures relevant to stroke medicine and their relevance to the WHO ICF are summarised in Table 1. Opinion regarding the best level at which to measure function after stroke is divided – should we be interested only in the physical deficit or symptom, or is it of more relevance to the patient to determine how that symptom affects their ability to return to their previous activity level and fulfil their roles and responsibilities within society? Certainly the former is easier to measure with less subjectivity but it could be argued that this is not as useful or meaningful a measure as the latter.



Body structure / function: anatomical parts and physiological function of body systems.

Impairments are problems with body structure/function

Activity: execution of a task or action by an individual

Participation: involvement of an individual in life situations

Environmental and Personal Factors: physical, social and cultural environment which affects how individuals live and conduct their lives

***Activity Limitation**:* difficulties an individual may experience in performing activities

***Participation Restriction**:* problems an individual may experience in involvement in life situations

Figure 1 Interactions among components of the WHO ICF

Table 1 Common scales used in measurement of stroke outcome by ICF stratification

Scales that measure disorder of:		
Body Structure / Function *(Impairments)*	**Limitation to Activity** *(Disability)*	**Limitation to Participation** *(Handicap)*
Glasgow Coma Scale (GCS)	Barthel Index	London Handicap scale
Mini-Mental State Examination (MMSE)	Nottingham ADL Scale	Medical Outcomes Study Short Form – 36
Geriatric Depression Scale	Glasgow Outcome Scale	Nottingham Health Profile
Hospital Anxiety and Depression Scale	Katz ADL Scale	
Modified Ashworth Test	10 metre walk test	
	Timed get-up-and-go test	
National Institute of Health Stroke Scale	International Stroke Trial (IST) Simple Questions	Stroke Adapted Sickness Impact Profile
Fugl Meyer Assessment	Modified Rankin Scale	Stroke Impact Scale
Orgogozo Stroke Scale	Oxford Handicap Scale	Stroke Specific Quality of Life
Canadian Neurological Scale	Hamrin Activities Index	Frenchay Activities Index
Scandinavian Stroke Scale	Adams Disability Method	
Toronto Stroke Scale		
(Modified) Mathew Scale		
European Stroke Scale	**General Neurological Scale**	
Frenchay Aphasia Screening Test	**Stroke Specific Scale**	

The choice of outcome scale is entirely dependent upon the clinical application. The scale chosen for use in a busy outpatient clinical setting will differ from that chosen in a detailed research protocol. The relative strengths and limitations of various stroke outcome scales in the context of clinical trials are debated. The "ideal" outcome measure is unlikely ever to exist. However, various instruments are preferred in contemporary stroke research and these merit further discussion[29]. (See section 1.3.4) The appropriateness and effectiveness of

a scale can be measured by describing its clinimetric properties. Before discussing the common stroke outcome measures in detail, an explanation of these clinimetric properties is necessary.

1.3.3. Clinimetric Properties

Scientifically valid research requires quantified comparisons between subjects. This requires translation of what often begins as a qualitative assessment, originating in a conversation, interview or physical examination, into quantitative data. After collection of these clinical and personal details they must be documented in a format that is easily comparable to other subjects. For many years clinical data were considered too "soft" to meet the standard of data quality demanded by scientists and clinical trials[30-32]. A description of the clinimetric properties of an outcome measure allows trialists to demonstrate that their assessment method and study design is robust and reliable. Practical and logistical issues must be considered together with the statistical clinimetric measurements to ensure optimum participation and follow up.

Clinimetrics is the methodological discipline that focuses on quality of clinical measurements. Outcome scales are traditionally assessed in terms of validity, reliability and responsiveness. Other important metrics include feasibility; acceptability and cost benefit.

1.3.3.1. Validity

Some attributes, such as height or weight are easily definable. Less tangible concepts such as functional ability require a description of the validity of the measurement[33]. A scale is considered valid if it bears a strong relationship to the attribute that it aims to measure[34]. The concept of validity includes several key components.

Content Validity and **Face validity** are relatively subjective terms which are used where an intuitive or consensus opinion would suggest that the content of the test *appears* to measure the attribute being studied. These may be considered where a committee of experts aim to develop a novel measure; however, a more objective test of its validity would be desirable before use in formal research.

Criterion Validity refers to a tests ability to match a criterion that is known to represent the attribute being studied (a "gold standard"). Where no such "gold standard" is available, measures of validity based upon other surrogate outcome measures or predictors are necessary. **Convergent Validity** refers to a tests relationship with other outcome measures. This is often used in studies examining new scales, allowing comparison with a variety of other established scales. **Construct Validity** refers to a tests relationship to other accepted indicators of measuring the desired attribute. This can be used in developing a novel outcome assessment tool by correlating the test result with other known prognostic indicators (for example with stroke subtype or blood pressure measurement in stroke outcome).

1.3.3.2. Reliability

Reliability is an important measure of a scales ability to be used clinically or for research purposes. The purpose of translating clinical data into scale format is to allow comparisons to be made with other subjects, however, without assurance of adequate reliability the content of the test is jeopardised by variability. Consistency of scoring is particularly important in scales with multiple items where the opportunity for disagreement is significant. In the development of scoring systems attempts should be made to ensure there are a minimal number of potential grades per modality tested with simple and unambiguous definitions[35].

Reliability can be measured in two ways – **intra-observer reliability:** reproducibility of results when tested repeatedly by the same assessor (also known as test-retest reliability) and **inter-observer reliability**: reproducibility of results when tested by two or more separate observers.

A measure of reliability estimates the degree of random error that is introduced in scoring. This error can be quantified using correlation statistics such as the kappa statistic (κ) or intraclass correlation coefficient (ICC).

1.3.3.3. Measures of Inter-observer reliability

The reliability of two or more observers rating an outcome can be measured using κ, κ_w, or ICC, depending on whether the outcome is categorical, ordinal or continuous, respectively. All three measures are expected to range from 0 to 1 and have the same interpretation: a value of 0 represents observers guessing randomly while 1 indicates perfect agreement. When comparing two different methods of measurement a Bland Altman plot can be employed, this is a graphical representation of the difference between two methods.

Kappa (κ) κ measures reliability as agreement adjusted for chance. A kappa statistic is a measure of agreement between raters, beyond that which is expected by chance alone. It can be calculated to quantify agreement between raters for nominal and categorical data[36]. A kappa statistic of zero indicates agreement equal to that which would be expected by chance, a kappa statistic of one indicates perfect agreement. A negative kappa statistic would indicate active disagreement of a similar magnitude. Commonly accepted thresholds for kappa reliability statistics are present in the literature. A score of 0-0.2 is considered poor; 0.21-0.4 fair; 0.41-0.6 moderate; 0.61-0.8 excellent and 0.81-1.00 excellent. A kappa statistic of 0.6 or above is considered necessary for clinical use. κ is less useful for ordinal outcomes such as mRS because near agreement and strong disagreement have the same impact on κ.

Weighted Kappa (κ_w) Weighted κ, κ_w, penalises near agreement less severely than strong disgreement, and is therefore more appropriate for ordinal outcomes. In this situation, a disagreement of more than one level on the scale (in either direction) would be considered more significant than disagreement across adjacent levels. Here a quadratic weight can be applied to the degree of disagreement to generate a weighted kappa statistic (κ_w)[37] The most common method of calculating κ_w is Fleiss-Cohen (quadratic) weighting. In this thesis, κ_w always implies Fleiss-Cohen κ_w.

Intraclass Correlation Coefficient (ICC) Inter-observer reliability for continuous outcomes can be measured by the intraclass correlation coefficient, which measures the proportion of variance in the outcome due to differences between subjects. The remaining variance (1 –

ICC) is due to differences between observers. Fleiss and Cohen[38] showed that for ordinal outcomes ICC and κ_w are equivalent.

Bland Altman Plot The aim of a Bland Altman plot is to demonstrate if two methods of measurement agree sufficiently for them to be used interchangeably or for one method of measurement to supersede an existing gold standard. Primarily designed to assess the difference between continuous measurements, a Bland Altman plot can also be used for categorical data[39]. The plot is of the average measurement against the difference between measurements[40].

1.3.3.4. Responsiveness

The responsiveness of a scale describes its sensitivity to detect change over time within subjects in response to meaningful changes in clinical status[34]. A stroke scale should be able to detect changes in functional ability as the patient progress through rehabilitation and recovery.

Where a scale is used to measure the magnitude of a therapeutic treatment effect, responsiveness is an important characteristic of the scale. The quantification of a scale's responsiveness can be achieved by comparison to another external criterion using correlation or receiver operator characteristics (ROC) analysis[41].

Responsiveness is also sometimes described by use of a "minimal clinically important difference" or" minimal clinically important change"[42]. This concept attempts to quantify the degree of change required in a scale that is associated with a patient perception of benefit. However, there are as many as nine reported methods of generating these estimates[42] and the lack of standardisation in development of these figures has undermined the utility of such a concept in clinimetrics[43]. The minimal degree of change that is significant will vary according to the trial. However even clinically modest improvements in function can have substantial meaning to patients and be important at a population level. Increasing responsiveness is by its nature often at the cost of increasing complexity.

The responsiveness of many scales is restricted by floor or ceiling effects. This describes the limits of a scale's ability to detect change, beyond which no further improvement (or deterioration) can be described. Where a subject achieves "full marks" in a score relative to the specific scale items in question, yet has further recovery potential which could positively impact their function and quality of life, the scale is limited by a ceiling effect (and vice versa). Where a scale suffers from floor or ceiling effects the measure is not useful to assess or document further potential improvement or deterioration which may be attained by the next assessment in a study protocol.

1.3.3.5. Acceptability

Acceptability refers to the burden of administering the scale. An ideal scale will be quick to administer with minimal distress to participants and minimal disruption to other activities.

1.3.3.6. Feasibility

Feasibility also refers to the burden of administering a scale, however in this context it is considered in terms of the burden on assessors. Pressure of time, expense and disruption to other clinical care duties are particularly relevant in a clinical context. However, even within the realms of research, similar pressures are present and the choice of outcome scale must not be allowed to discourage trial enrolment or progression through undue effort placed upon assessors.

1.3.3.7. Interpretability

The scores generated should be meaningful and comparable to previous studies. Clearly in the generation of a novel outcome measure this is less relevant; however in choosing an outcome measure for a trial where a number of tools are used it is important to ensure that the choice of primary outcome measure will generate data that is meaningful in the context of existing literature. The landmark thrombolysis trials were comparable in an important pooled analysis due to the virtue of similar outcome measures facilitated through prospective collaboration[44].

A recent analysis of outcome measures in contemporary stroke research has demonstrated use of 47 different outcome scales[29]. However, a significant majority of studies used one of a smaller group of favoured tools. These tools are preferred because of increasing clinimetric data to support their use, both objectively (valid, reliable and responsive tools) but also subjectively; the practicality (acceptability, feasibility and interpretability) of the chosen tool(s) is crucial to the success of a trial.

1.3.4. Commonly used stroke outcome measures

The widespread development of stroke scales in recent decades has led to much confusion in the literature. There are scales designed to classify stroke syndromes (TOAST Classification, Oxford Classification, ICD 9/10, Physicians Health Study Stroke Subtypes etc.), scales developed to quantify stroke deficit and neurological examination (NIHSS, Canadian Neurological Scale, European Stroke Scale, Mathew Scale etc.) and scales developed to measure function and quality of life (Barthel Index, modified Rankin Scale, Stroke Impact Scale, Functional Independence Measure etc.). These scales have all been used and continue to be used in stroke trials; however a preference for certain measures has been noted[29].

1.3.4.1. National Institutes of Health Stroke Scale (NIHSS)

The NIHSS is a measurement of physical impairment following stroke and is a common marker of stroke severity, allowing simple quantification of clinical examination findings. It was originally described in 1989 using components of several previous stroke scales (the Toronto Stroke Scale, the Cincinnati Stroke Scale, the Oxbury Initial Severity Scale and the Edinburgh-2 Coma Scale)[45] and in its current form it assesses fifteen items with a 3 to 4 point scale (level of consciousness / extraoccular eye movements / visual fields / facial muscle function / arm and leg motor function / sensory disturbance / ataxia / language / dysarthria and inattention). (Figure 2) Construct validity of the NIHSS as a marker of stroke severity is favourable when compared to infarct volume on CT scanning[45]. Moderate to substantial inter-rater reliability has consistently been demonstrated for the complete scale in assessors of diverse backgrounds[45, 46, 46-48]. However, some specific items are noted to be less reliable; ataxia, facial weakness, dysarthria and level of consciousness; exhibiting unacceptable

agreement by kappa statistic[46]. A modified version of NIHSS (mNIHSS) was proposed in 2001, excluding these more contentious items. This version of the scale demonstrated slightly increased reliability but appeared to convey no advantage in statistical modelling[49]. Reliability and validity of the mNIHSS has also been demonstrated in a prospective sample[50] but the use of this version of the scale has not been adopted routinely in trial design.

In its current form the NIHSS is frequently used as a research tool and in clinical practice. Measurement of NIHSS was included in 27.8% of recent trial procedures and it was used as a primary outcome measure in 11.9%[29]. In routine clinical practice it is also used commonly to aid treatment decisions and provide prognostic information at baseline[51, 52]. Widespread experience and knowledge of the scale is one of its great strengths in both stroke physicians and non specialists.

However, the NIHSS does have accepted limitations. As a scale used to document physical impairments there is an acknowledged ceiling effect in its administration (see section 1.3.3.3). The scale items favour left hemisphere events; of 42 possible marks there are seven directly related to language function where only two measure a degree of neglect / inattention. A study examining CT infarct volume has questioned the validity of NIHSS on this basis; the median volume of right hemisphere infarcts was found to be consistently higher than left hemisphere infarcts for the same NIHSS score[53]. Common clinical findings in posterior cerebral circulation events are also poorly reflected in the NIHSS. The development of hemisphere specific or stroke syndrome specific versions of NIHSS have been proposed[54] but have not been adopted.

Overall, the NIHSS is a useful tool for acute clinical assessment and is acknowledged to accurately document stroke severity and predict outcome[45, 51]. For this reason it is widely accepted in stroke research. However, its usefulness as a functional outcome measure is questionable; recovery of a physical deficit may not translate to meaningful functional recovery.

1a. Level of Consciousness (LOC)	0= Alert; keenly responsive 1= Not alert; but arousable by minor stimulation 2= Not alert; requires repeated stimulation to attend 3= Unresponsive; reflex movements only	
1b. LOC Questions Ask patient the month and his/her age	0= Answers both questions correctly 1= Answers one question correctly 2= Answers neither question correctly	
1c. LOC Commands Open and close eyes Grip and release non paretic hand	0= Performs both tasks correctly 1= Performs one task correctly 2= Performs neither task correctly	
2. Best Gaze Horizontal movements only	0= Normal 1= Partial gaze palsy 2= Forced deviation not overcome by oculocephalic manoeuvre	
3. Visual Fields	0= Normal 1= Partial Hemianopia 2= Complete Hemianopia 3= Bilateral Hemianopia (blind including cortical blindness)	
4. Facial Palsy	0= Normal 1= Minor Paralysis (flattened nasolabial fold, asymmetry on smiling) 2= Partial Paralysis (total or near total paralysis of lower face) 3= Complete Paralysis of one or both sides, absence of facial movement in the upper and lower face	
5. Motor Function – Arm	0= Normal; holds limb 90 (or 45) degrees for 10 seconds without drift 1= Drift; limb holds 90 (or 45) degrees but drifts down before full 10 seconds but does not hit bed or other support 2= Some effort against gravity 3= No effort against gravity; limb falls 4= No movement UN= Untestable; joint fused or amputated	
6. Motor Function – Leg	0= Normal; leg holds 30 degree position for 5 seconds without drift 1= Drift; leg falls by end of the 5 second period but does not hit bed 2= Some effort against gravity 3= No effort against gravity; limb falls 4= No movement UN= Untestable; joint fused or amputated	
7. Limb Ataxia	0= No ataxia 1= Present in one limb 2= Present in two limbs UN= Untestable; joint fused or amputated	
8. Sensory	0= Normal; no sensory loss 1= Mild to moderate sensory loss, aware of touch 2= Severe to total sensory loss	
9. Best Language	0= No aphasia 1= Mild to moderate aphasia; loss of fluency or comprehension 2= Severe aphasia; fragmented communication 3= Mute, global aphasia; no useable speech or auditory comprehension	
10. Dysarthria	0= Normal 1= Mild to moderate dysarthria; slurring of words, at worst can be understood with some difficulty 2= Severe dysarthria; near unintelligible or unable to speak (out of proportion to aphasia) UN= Untestable due to intubation or physical barrier	
11. Extinction and Inattention	0= No abnormality 1= Inattention or extinction to bilateral simultaneous stimulation in one sensory modality (visual, tactile, auditory, spatial or personal) 2= Profound hemi-inattention or extinction to one or more modality	
Total Score		

Figure 2 The National Institutes of Health Stroke Scale (NIHSS)[45]

1.3.4.2. The Barthel Index of Activities of Daily Living

The Barthel Index (BI) has been in common use for many years and is a familiar instrument in all areas of rehabilitation medicine. First described in the literature in 1965[55] the scale had been in clinical use locally for almost a decade. Initially proposed as a simple measure to quantify independence it was used as a score to document improvement in rehabilitation. The original scale describes ten tasks in the areas of personal care and mobility, scoring 0 to 100 with 5 point increments. (Figure 3) As a general rehabilitation scale it does not contain any stroke specific domains such as communication / cognition.

There are several variations of the original scale which are collectively described in the literature as the BI. Although the variations on the original scale are largely similar they include modifications such as change in scale items[56] or definition[57] / reordering of scale items[58]. These changes have potential to alter the clinimetric properties of the scale raising question over their validity in a research setting.

The methods used to collect BI data also vary substantially across the literature. The original scale was designed to be administered through direct observation of scale tasks. However, validation of a scale for one purpose does not ensure that its validity will translate into a different clinical context or situation[59]. For ease of data collection centres have administered the scale via self reporting[60], telephone interview[61] or postal questionnaire[62]. It must be recognised that a change in scale administration can alter its performance.

The reliability of the BI has consistently been reported as good when administered in a general rehabilitation population[63]. Individual studies examining reliability in stroke patients are reassuring in their estimates[64-66]. A recent systematic review and meta-analysis found overall excellent inter-observer variability[67], however there were discernible differences in the included studies; limited by small sample size, heterogeneous study population and diverse methodology. However, in an elderly cohort, systematic review of BI reliability found only fair to moderate interobserver agreement by kappa statistic in individual items, postulating that the use of this as a tool may be limited in an elderly population where cognitive impairment is prevalent[68].

Feeding	0= Unable 5= Needs help (cutting / spreading) 10= Independent
Transfers Bed to chair (or wheelchair) and back	0= Unable, no sitting balance 5= Major help (of one or two people), can sit 10= Minor help (verbal or physical) 15= Independent
Grooming Face/Hair/Teeth/Shaving	0= Needs help 5= Independent, implements provided
Toilet Use Handling clothes, on and off toilet, wipe, flush	0= Dependent 5= Needs some help but can do something alone 10= Independent
Bathing	0= Dependent 5= Independent
Mobility On surface level	0= Immobile or < 50 yards 5= Wheelchair independent, including corners, >50 yards 10= walks with help of one person (verbal or physical), >50 yards 15= Independent (but may use aid), > 50 yards
Stairs	0= Unable 5= Needs help (verbal, physical or carrying aid) 10= Independent
Dressing Includes buttons, zips, laces etc.	0= Dependent 5= Needs help but can do ≈50% unaided 10= Independent, including fastenings
Bowels	0= Incontinent (or requires enemas) 5= Occasional accident 10= Continent
Bladder	0= Incontinent (or catheterised) 5= Occasional accident 10= Continent
Total Score	

Figure 3 The Barthel Index (BI)[55]

The BI certainly has strengths as a stroke outcome measure which explains its use in several previous landmark stroke trials, including the thrombolysis trials. It has been used as an instrument in 40.5% of recent publications and was primary outcome measure in 7.9%. It is familiar to a wide range of clinicians both in clinical practice and as a research tool, aiding its use in multicentre trials. It has good intra and inter observer reliability and is a useful predictor of length of stay, discharge destination[69] and of the likelihood of a patients return to independent community living[70, 71]. It also generates a standardised, numerical score which is easily comparable at various time points, aiding in statistical analysis and manipulation.

However, there are widely acknowledged limitations in the use of BI as a functional outcome measure, which may explain increasing reluctance in contemporary trials to use it as a primary outcome measure. Heterogeneity in the use of various versions of the BI and diversity in methodology of administration adds complexity when interpreting and comparing trial results. As an ADL score the BI is not designed to assess many areas of neurological deficit or disability commonly reported in a stroke population (disorders of cognition, language, vision etc.) which may have significant impact upon function and quality of life. The primary limitations seen in the BI are the substantial "floor" and "ceiling" effects in scoring subjects at the extremes of the scale[72-74] together with a U-shaped distribution in clinical practice. These limit the responsiveness of the BI, hindering the ability of the scale to detect meaningful change over time[75] and requiring a large change shift in outcome in order for this to be detected statistically.

1.3.4.3. The modified Rankin Scale / Oxford Handicap Scale

The modified Rankin Scale is the most commonly used outcome measure in contemporary stroke research, used in 64% of recent stroke trials and as a primary outcome measure in 26%[29]. An ordinal, hierarchical scale, it is used to measure disability across seven ranks. (Figure 4) There have been several variations of the scale since its inception in the 1950's (including the alternatively named Oxford Handicap Scale - Figure 5) and it has been widely adopted in a standard form for use in trials. A detailed discussion of the modified Rankin Scale, its clinimetric properties and variations is found in section 1.4.

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

Figure 4 The modified Rankin Scale

0	No symptoms
1	Minor symptoms that do not interfere with lifestyle
2	Minor handicap, symptoms that lead to some restriction in lifestyle but do not interfere with the patient's capacity to look after himself
3	Moderate handicap, symptoms that significantly restrict lifestyle and prevent totally independent existence
4	Moderately severe handicap, symptoms that clearly prevent independent existence though not needing constant attention
5	Severe handicap, totally dependent patient requiring constant attention night and day

Figure 5 The Oxford Handicap Scale (OHS)[76]

1.3.4.4. Stroke Impact Scale

As a measure of stroke outcome across many domains, the Stroke Impact Scale (SIS) was developed in response to the criticism that the above common scales do not directly or adequately assess stroke survivors in a holistic or comprehensive manner. The SIS was developed using a patient centred approach involving feedback from patients, carers and therapists rather than trial investigators in the description of domains. The scale includes measurement of changes in emotion, cognition, communication and social role / participation in addition to the expected physical attributes tested after stroke. The scale was originally developed in 1999[77] and was modified in 2003 to its current version[78] although it

continues to be refined. Despite its careful description and development to ensure validity and reliability, the SIS has not been adopted widely in the stroke literature and its generalisability is questionable[79]. It was used in only 2.4% of recent stroke trials and was not utilised as a primary outcome measure in any[29].

1.3.5. Global Statistic Outcome Measures

Because each stroke outcome scale is recognised to measure important but often very different parameters in stroke recovery a global statistical outcome measure has been proposed in some large trials to better describe the spectrum of outcome after intervention. Mathematical techniques can be used to combine several outcome measures, each of which measures a different but important domain in stroke recovery, such as neurological deficits, independent function and quality of life. The NINDS tPA Stroke trial, a landmark thrombolysis trial[11, 80], utilised a global statistic incorporating the NIHSS, mRS, BI and Glasgow Outcome Scale (another hierarchical ranked scale used to measure outcome after brain injury – frequently traumatic head injury) as a primary outcome. As a successful method for the NINDS trial, a retrospective analysis of the ECASS I dataset using the NINDS statistical methodology altered the outcome of the intention to treat analysis from a negative study to one with a positive result. Clearly retrospective data analysis such as this has limitations and is methodologically flawed, however it highlights the importance of choosing the optimal outcome measure in trial results[81].

With the advent of more complex statistical analysis techniques the use of global statistics may be more widely adopted. Trial power may be enhanced with the use of a global statistic through reduction of the random variation seen in a single scale. However, there are disadvantages and the use of a global statistic and their use in place of a more recognisable primary outcome measure is discouraged by regulatory bodies. By mixing conceptually distinct recovery descriptors, ultimately such an approach provides an abstract result that is less intuitive than a single well defined outcome. It must be recognised that any global outcome measure will be limited by the weakest performing tool involved and the resulting statistics are less interpretable than a single outcome measure. There are also statistical complexities in analysis. A global outcome may be able to provide an estimate of statistical

significance, but is less able to provide a meaningful measure of effect size. Furthermore, analysis often includes the combination of several ordinal scales. In this situation global statistics are often analysed with multiple dichotomisation points with recognised statistical inefficiencies[82]. (See section 1.5)

1.4. The modified Rankin Scale

Originally described by Dr John Rankin in 1957 the Rankin Scale (RS)[83] was designed to provide a method of describing the recovery of a group of stroke patients at discharge or transfer from hospital. Dr Rankin established an archetypal specialist "stroke unit"[84] and the scale was designed to evaluate the effectiveness of this intervention.

1.4.1. The Modified Rankin Scale (mRS)

The original RS scale was modified in 1988 as part of the UK-TIA Aspirin Study[85] to what is now accepted as the modified Rankin Scale (mRS). (Figure 4) Changes to the original scale included the addition of an extra category (Grade 0) and an alteration of the wording of grades 0, 1 and 2 to better accommodate disturbances of language and cognitive function[86].

1.4.1.1. The Oxfordshire Handicap Scale (OHS)

Further modification was proposed in 1989 by the Oxfordshire Community Stroke Project with the alternative title of Oxfordshire Handicap Scale (OHS)[76]. In response to controversy surrounding the use of the functional terms in the mRS[87], the OHS was reworded to standardise the use of the term "handicap" in place of "disability" and include the term "lifestyle" in scoring. (Figure 5) In addition to this change in semantics, the focus of the mRS on mobility was reduced by removing the ability to walk as an explicit grading criterion[88]. The OHS has been used in few studies[89-91] and has not been adopted widely in the stroke literature.

The mRS in its 1988 format[86] (Figure 4) is widely accepted as the standard version of the scale. It is an ordinal hierarchical scale with grades from zero (no symptoms) to five (severe disability); an extra score of 6 is often added in clinical trials to signify death. It is important

to note that some studies describe the use of the mRS but cite alternative versions of the scale. It is important to clarify the specific instrument used in order to be confident of comparable trial results.

1.4.2. Clinimetric properties of mRS

The validity, reliability and responsiveness of the mRS are described in a broad literature. Although acknowledged as an imperfect scale, it has been demonstrated to meet the requirements of an outcome measure for use in randomised clinical trials and has been recommended for its brevity, simplicity and ease of interpretation in clinical trials[89]. The long term predictive value the mRS makes it an attractive scale for use in RCT's. mRS data collected at 90 days are representative of the likely long term functional outcome[92].

1.4.2.1. Validity

The mRS has been shown to correlate well with several other markers of stroke severity and outcome and can be considered a valid measurement of functional outcome. Convergent Validity has been shown with other outcome scales such as the NIHSS and BI[72, 93] and with economic indices such as length of hospital stay and cost of patient care[94]. Construct validity has also been reported via association with arguably more objective and direct measures of stroke severity such as infarct volume[95, 96] or recanalisation of affected vessels following tPA therapy[97].

1.4.2.2. Reliability

Reliability has been reported across a wide and somewhat fragmented literature. Caution must be exercised in interpreting results of these studies as they are heterogeneous in methodology: patient selection, timescale after stroke, sample size, assessor number, background and training and number of centres vary throughout.

In most cases reliability studies have been conducted using highly trained and motivated individuals in a single centre[98]. Systematic review has found wide estimates of inter-observer variability with overall good agreement (κ=0.61)[99]. Estimates range from κ=0.25[100] to κ=0.72[101] for a standard mRS interview. Only one mRS reliability study has a design that

attempts to emulate an active RCT; a multicentre study with several investigators of different clinical backgrounds[100]. The reliability of the traditional mRS in this study was of concern (κ=0.25) in the context of large multicentre studies. Substantial heterogeneity has been reported in mRS reliability amongst a large cohort of international observers[102]. Where our closest estimate of mRS reliability in a multicentre RCT shows substantial disagreement (κ=0.25) we might expect that this could be amplified in an international RCT.

1.4.2.3. Responsiveness

As a scale with only six potential categories, the mRS requires meaningful clinical improvement or deterioration to move between each point on the scale. This attribute makes it a poor tool for measuring change over short periods of time. However its advantage in clinical trials for the assessment of longer term follow up (e.g. at study completion) is recognised. A change in mRS score in response to treatment is very likely to be associated with a clinically meaningful change as perceived by both the patient and the investigator[72, 103]. No minimal clinically important change is reported for the mRS.

The mRS scale distributes disability states meaningfully, in comparison with the NIHSS and Barthel Index which are troubled by floor and ceiling effects. This is useful clinically and statistically. Figure 6[104] demonstrates the 90 day outcome measures for the NIHSS, BI and mRS. It is clear that the information provided by the mRS and distribution of mRS outcomes will allow more robust statistical analysis and interpretation.

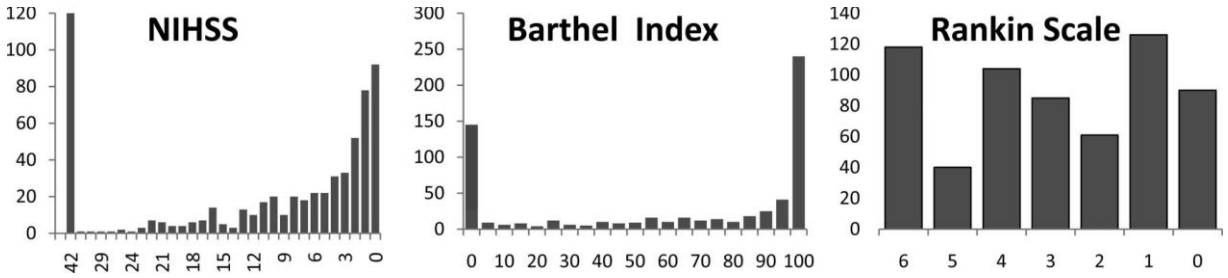


Figure 6 Typical Distribution of outcome with the NIHSS, BI and mRS at 90 days. [Final 90-day outcome scores in the 2 NINDS tissue-type plasminogen activator trials[104]].

1.4.3. Challenges in the use of mRS

Although it is the favoured outcome scale for use in contemporary stroke research, the mRS is acknowledged as an imperfect tool for several reasons. The perfect outcome measure is unlikely ever to exist and hence trialists must accept the limitations in the available tools and ensure that they are optimised in their administration.

The mRS has been criticised due to the broad nature of each category and large potential change in function that is contained within each rank. The broad categories of the mRS are considered by some as subjective, ambiguous and poorly defined[105]. In contrast, the "global" nature of mRS is also considered one of its strengths, particularly through reduction in the floor and ceiling effects that limit many outcome measures. As a simple, time efficient measure it should be performed within around fifteen minutes and can be administered by a variety of health care professionals with similar validity and reliability in scoring[106].

The broad nature of the mRS may be to its advantage; however there is an argument that it lacks specificity, without explicit measurement of certain domains. The wording of the mRS may place undue reliance on predominantly physical attributes such as ambulation and continence. There is no direct focus within the assessment in areas such as cognition, communication, language disorders or other common post stroke syndromes (fatigue / pain / mood disorder) that can affect motivation and patient perception of function. The "global" nature of the assessment allows implicit inclusion of these domains; an experienced assessor should consider physical and non physical characteristics in assessment of overall function and participation[103].

The influence of co-morbid conditions must be considered in assigning a score; arthritis, respiratory or cardiovascular disease, depression and many other confounding illnesses are common in the stroke population and can have an important impact upon the functional domains that are affected by stroke disease. Co-morbid illness is known to affect stroke survivors' quality of life[107] but the mRS should be used to assess the effect of stroke related limitations. Again, an experienced assessor should be able to disentangle the aetiology of a

patient's limitations but the uncertainty that these factors bring to the allocation of mRS scoring may limit its utility.

The timing of mRS administration is controversial. It is accepted that most stroke recovery occurs in the early phase of rehabilitation, however improvement can continue at a slower rate for up to six or twelve months. Administration of the mRS early in the recovery phase may be considered unreliable for outcome classification[108]. Prolonged admission is common after stroke for intensive rehabilitation and therapy. The effectiveness of the mRS in patient populations who have not returned home or had an opportunity to return to many aspects of previous function is questionable. The standard time point of mRS assessment in clinical trials is 90 days post ictus, a period recognised as a predictor of longer term outcome[92]. This prognostic information is useful clinically in planning rehabilitation goals or likely discharge destination; importantly in research it allows trial follow up to be completed within a reasonable timeframe.

Across each boundary of the mRS there are controversies, due to both the inherent ambiguity of the scale and the potential misinterpretation of its wording by assessors. For example, the boundary between mRS 3 and mRS 4 largely relies on the ability to walk unaided. The use of aids and adaptations (such as a walking frame or cane) should not merit the term "assistance"; rather these are aids to allow independence[88]. However, if the assistance of another person, physical or supervisory, is necessary then this would warrant a score of four. A similar dilemma exists in the use of a wheelchair; this aid can permit functional mobility without the ability to walk independently. Many wheelchair users would disagree that this modification leads to significant disability questioning the appropriateness of an mRS of four (Figure 4). Often a patient is capable of performing particular tasks but chooses not to because "assistance" is offered and this help is accepted, for example in the context of a caring relative who wishes to save the subject the time and energy required to complete tasks. In this case it is arguable whether they should they be scored based upon their potential capability or based upon their true function. The wording of the mRS does not clearly specify whether a subject should be scored on the basis of absolute loss of activity or chosen loss of activity for the sake of convenience or increased effort. Examples of areas of

controversy are present at each level of the scale but are more prevalent in the middle ranges where classification of "good" or "bad" outcome often rests. Areas of controversy result in potential for variability in scoring; they will be discussed further in later chapters.

Health related quality of life measures are important after stroke but are infrequently incorporated in trial design or used as a primary or secondary outcome[109]. For this reason, translation of the functional ability with each rank on the mRS scale to a single index value for health status is necessary to assess the "value" of treatment in terms of evaluating health care provision and in health economics. For this to be possible, a disease specific instrument such as the mRS must be expressed in a form comparable to universal generic health outcomes. In the United Kingdom, generic outcomes provide utility values from which health related quality of life measures, such as a Quality Adjusted Life Year (QALY) must be derived. It is a challenge to express the mRS outcome scale in this format but there is some evidence to suggest that it is possible and valid.

The mRS has been successfully mapped to EuroQol-5D[110], a generic health outcome measure commonly used by health economists in evaluation of new treatment strategies. An alternative method of expressing utility values is through the use of Disability Adjusted Life Year (DALY). This measure is generated from WHO Global Burden of Disease (WHO-GBD) disability weights, a numeric value which reflects the severity of disease or disability on a scale from 0 (perfect health) to 1 (equivalent to death). This spectrum of disability is intuitively similar to the mRS scale. The mRS has been described using this spectrum by generating disability weights for each rank of the mRS scale[111].

Despite these challenges and pitfalls, the mRS is a well recognised scale which has been well studied with an established and acceptable clinimetric literature. Its ease of administration and global nature is considered its great asset and in experienced hands it is used with confidence in large clinical trials[93].

1.4.4. Variants of the standard mRS

The standard mRS interview was developed as a face-to-face interview between assessor and patient. Several variants of the standard mRS interview have been proposed in an attempt to simplify administration, improve accessibility or improve reliability.

1.4.4.1. Premorbid mRS score

Traditional statistical analyses in acute stroke trials measure the proportion of participants achieving a "good" functional outcome, by whatever criterion that is defined in each trial. This concept is complicated by the inclusion of participants with considerable pre-stroke functional limitation. For this reason a pre-stroke mRS is often used as a measure of disability at entry to contemporary stroke trials. In clinical practice, a national audit programme is collecting data on outcomes at 6 months and one year, corrected for pre-stroke mRS[112]. In some centres, pre-stroke mRS is used as a measure of functional independence, upon which a decision to administer or withhold thrombolytic therapy may be based.

The concept of pre-stroke mRS is difficult to define, as by its very definition, the modified Rankin scale describes the presence or absence of symptoms and limitations after stroke. However, the descriptions in terms of symptoms, limitations and the need for assistance are extrapolated to include all aetiologies of disability in this context. The validity and reliability of a pre-stroke mRS is moderate with some concerns raised regarding its use as an entry criteria to stroke trials[113]. A re-written mRS with the same structure but alternative wording regarding symptoms and limitations has been proposed.

1.4.4.2. Acute mRS assessment

Assessment of mRS in the acute phase is common in stroke trials but has been poorly studied in the hospital setting, where most participants are in the immediate phase after stroke. The wording of the scale concentrates on functional limitations and the need for assistance with activity. In the hospital setting this help is available but not always required and participants have often not had the opportunity to assess their abilities with extended ADLs such as shopping, cooking or managing their finances. The reliability of the mRS in the acute phase

has been shown to be good, with acceptable inter-rater agreement. However, the validity of the scale in this context, particularly with reference to the relevance and interpretation it is questionable[114].

1.4.4.3. Remote mRS assessment

As there is no physical examination component to the mRS, remote assessment (via telephone or postal questionnaire) is an attractive prospect in circumstances where direct assessment is not possible or to limit the time and expense involved in patient visits.

A postal based mRS questionnaire approach has been used in clinical trials. There are no data comparing postal versions of the mRS with traditional face-to-face assessment. A study compared differing postal mRS questionnaires and telephone assessment. The authors concluded that telephone follow up may be preferable as response rate from the postal questionnaire was suboptimal[115]. Studies are available describing the properties of postal versions of alternative stroke assessment measures from which we can extrapolate. A postal version of the Glasgow Outcome Scale (GOS) in head injury patients was found to have good reliability in comparison to telephone assessment[116]. However, limitations in data collection due to the high risk of non-responders were highlighted as a disadvantage in a study of a postal version of the Stroke Impact Scale (SIS)[117].

There are more data pertaining to telephone versions of stroke outcome assessment. ADL based assessment such as the BI has been demonstrated to have good reliability by telephone[118, 119]. More severely affected patients are likely to yield less reliable results by telephone[61] which may be a reflection of a patient's propensity to overstate their ability or an indication of the important nonverbal cues that are gained by visualising a patient. Telephone mRS assessments have been assessed for both validity and reliability[120], often in the form of a structured interview. Results are conflicting with inter-rater reliability reported from κ=0.30[101] to κ=0.78[121]. The largest study of telephone mRS reports a weighted kappa statistic rather than a standard kappa statistic (κ_w=0.71). This makes direct comparison with other studies difficult but certainly reflects a significant degree of disagreement[122]. At best we must interpret the result of a telephone mRS with caution.

Derivation of the mRS from outpatient case records has been proposed as an alternative method of collecting follow up data. Case record extraction has been shown to be reliable for impairment scales such as NIHSS[123]. However, a prospective study found poor reliability in case record derived mRS in comparison to standard mRS (κ=0.34) and between observers (κ=0.33)[124].

1.4.4.4. Proxy mRS assessment

In some circumstances the mRS cannot be administered through direct patient interview. Communication or cognitive impairments are common and when severe can preclude a standard mRS assessment. The properties of a proxy mRS assessment have been assessed in a cohort of patients[125]. Family members, nurses and physiotherapists were used as proxy respondents and mRS scores were compared to direct patient assessment. Only moderate agreement between proxy and patient assessments was found (κ=0.4), encouraging caution in the interpretation of proxy assessments. The greatest variability was found in therapist assessments. There is evidence to suggest that systematic differences are present in the perception of overall function between different groups of respondents. Although not directly studies of the mRS, extrapolation from other stroke assessment scales suggests that proxy respondents may be less reliable. Patients are prone to self-report better function than their carers[126]. Proxies systematically report more dysfunction in aspects of quality of life measures than direct patient assessment, thought to be affected by proxy perception of burden[127]. Therapists report better functional scores than carers[128]. This is perhaps a reflection of improved effort on the part of a patient when undergoing therapy assessment in comparison to home performance or that the nature of therapist training is in looking for rehabilitation potential. Systematic review of proxy responses to stroke assessments found that ADL measures were more suited to a proxy assessment and that less agreement is seen in quality of life measures[129]. No mRS studies were available for inclusion in this review but given the nature of the mRS assessment lies somewhere between an ADL impairment scale and global quality of life measure we must consider this when using proxy mRS assessments.

1.4.4.5. Structured mRS assessment

Although it can be considered an inherent advantage of the mRS assessment, the open and global nature of the standard interview, risking subjective interpretation, increases potential for inter-observer variability. An experienced assessor should gather the appropriate information to grade patients successfully using a standard, unstructured mRS interview. However, in an attempt to widen the utility of the mRS and encourage more standardised results, a structured interview has been proposed.

Several versions of a structured mRS assessment are available, most of which use a checklist format through various ADL's and activities[66, 100, 101, 105, 130-133]. This is most appropriate for assessing mRS at the lower end of the scale (mRS 3-5) where basic ADL's are an important element in scoring. The translation of ADL scoring (BI) to mRS scores has been shown to be an effective method of mRS assessment[65], however, this is less effective at the top end of the mRS beyond the ceiling threshold of the BI.

Reliability estimates of the structured interview are conflicting[101, 105]. In a multicentre trial with assessors of various healthcare backgrounds the use of a structured interview was found to be highly reliable[100]. However, the structured interview has been criticised as a complicated tool to implement. A Rankin Focused Assessment (RFA), a four page standardised assessment form is reported to be less cumbersome. Designed as part of the FAST-MAG trial protocol it has been shown to have excellent inter-observer reliability[133]. It is important to note that although RFA scores showed substantial agreement between raters, this study was restricted to one centre and no comparison was made with mRS scores generated in a traditional manner to confirm validity.

Even more focused mRS assessment has been described using a questionnaire methodology. Using the OHS version of the mRS scale a two question assessment (1. Do you feel that you have made a complete recovery from your stroke? and 2. Do you require help from another person for everyday activities?) was found to be sensitive in predicting BI and OHS scores of 20 or 0 respectively (i.e no symptoms or limitations)[134]. In an era of dichotomised mRS outcomes as the common analysis technique (see section 5.3) this was considered an

advantage in terms of time and efficiency. However, current trial design would not support this method. A more recent questionnaire based mRS assessment (Simplified modified Rankin Questionnaire smRSq) has been described using three questions each generating a binary yes / no response[132]. This generates mRS results across the full ordinal scale and has been shown in a single study to have similar reliability to a standard mRS assessment between two raters[132] and has been validated in comparison to stroke severity[135], stroke size[136] and quality of life measures[137]. The validity of the smRSq has been demonstrated in a study comparing its use to the standard mRS and the NIHSS in a cohort of Chinese stroke patients[138].

1.4.5. Adjudicated mRS outcomes

Central adjudication of trial end points is routinely used in a variety of settings but has rarely been used in stroke. This is straightforward where the endpoint is based upon the assessment of paper records or images; however, adjudicated endpoints of participant function are more problematic.

Group adjudication of mRS outcomes has been employed in some trial designs, including telephone mRS in the IMAGES[139] and IST 1[140] trials. In the HAMLET trial a written summary of the mRS assessment was provided for central review[141]. However, as discussed in section 1.4.4.3, the reliability of remote mRS is debated.

In order to evaluate mRS outcomes remotely in a format as close to current assessments as possible there are numerous barriers to overcome. First, this would require capture of the assessment, including both visual and audio components, in a format that can be subsequently reviewed 'off-line'. Further, most trials are international, mutli-cultural and multilingual; meaning methods of handling these complexities including translation would need to be considered and should be demonstrated not to influence scoring. Pilot data suggest that video based mRS assessment is valid and reliable[142].

1.5. Improving statistical analysis in acute stroke trials

The failure of many previous large RCTs in acute stroke has been attributed in part to trial design and statistical analysis. The detection of what is likely to be a small treatment effect requires statistically efficient methods to optimise the likelihood of demonstrating improved outcome with a realistic sample size.

1.5.1. Choice of scale

As previously discussed, the selection of outcome measure is crucial. It must be a reliable, robust and valid measure of what the treatment in question aims to achieve. The most common outcome measure is the mRS or its incorporation in a global endpoint statistic. Varying estimates of inter-rater reliability in the mRS have the potential to affect trial results. Most contemporary phase III trials are conducted internationally, with the potential for up to hundreds of investigators; therefore the scope for amplification of even small degrees of inter-observer variability is of concern. Endpoint misclassification, through incorrect or variable administration of the mRS, has the potential to introduce type II statistical error and compromise trial results through loss of power. In a trial of pneumococcal vaccine the recording of the endpoint was erroneously assumed to be infallible. It was demonstrated that a modest misclassification in the reported cause of death reduced the trial power by 40%[143]. Analysis of data from a neurotrauma studies, with similar challenges to that of stroke recovery, confirmed that endpoint misclassification could have a significant effect upon the magnitude of treatment effect. With flawless classification the effect size was 7.5% (p=0.039) but this lost statistical significance and fell to 6% (p=0.102) and 4.5% (p=0.228) where there was 10% and 20% misclassification respectively[144]. A pattern of misclassification upwards, downwards or in both directions can attenuate the treatment effect found in clinical trials[145].

1.5.2. Training and Certification

The requirement for large studies in stroke research demands international collaboration, multicentre trials and a considerable number of investigators to contribute data for analysis. Although all stroke scales are acknowledged to be imperfect, the consistent application of

even an imperfect tool is crucial in the collection of accurate, comparable and generalisable data[146].

The importance of investigator training was acknowledged when the mRS was originally described although at this stage was considered impractical in the context of a multicentre trial[86]. However, with technological advances this situation has changed.

Training in application of the NIHSS was developed for the TOAST trial in 1994 and was found to be feasible, effective and inexpensive[147]. Improved reliability in scoring using a video training package was reported in the TOAST investigators (n=162)[46]. An early analysis of reliability using the original video NIHSS training package in a large cohort (n=7405) found less consistent results in scoring[148]. Subsequently the training package was updated to DVD technology[48] and again to its current format as an online package. This has been robustly validated with good reliability in a large cohort from multiple venues (n=8214)[149] and there is evidence to support its use in translated form[150].

A similar training resource is available to improve the application of the mRS in clinical trials[151]. Again this began as an instructional DVD with subsequent conversion to a web based package accessible internationally. This has also been validated with good reliability in a large cohort (n=2942) from several countries[102]. As a less structured tool than the NIHSS no studies to date have investigated the validity and reliability of a translated mRS assessment.

Consistency in application of stroke scales requires a degree of quality control. This is best achieved through formal training and certification of investigators and a compulsory training process is now ubiquitous amongst contemporary stroke research where the NIHSS or the mRS are used.

1.5.3. Statistical Analysis Techniques

Following collection of outcome data, the analysis technique chosen to interpret it can significantly alter trial results. The relative advantages and disadvantages of different analysis techniques have been much debated in the stroke literature in recent years.

1.5.3.1. Dichotomised Analysis

A popular method of analysing disability endpoints has been to dichotomise outcome data, separating participants into those with "favourable" and "non-favourable" outcomes and comparing proportions. With mRS a variety of cut offs have been used to define "favourable" outcome status. There is no consensus as to the level of mRS that best represents acceptable recovery and choice has been partially dependent on the expected benefit of the therapy and the baseline disability of the cohort. For example, in a trial of intervention in the often fatal condition of malignant middle cerebral artery infarction a "good" outcome was defined as mRS 3 (moderate disability), while in trials of thrombolysis, where better functional outcomes are expected, use of mRS 0-1 (no significant disability) defines treatment success. Dichotomisation of outcome data is attractive because it allows easy interpretation into "good" and "bad" outcome, simplifying statistical analysis and interpretation of results. Placing the threshold between "good" and "bad" outcome across one health state is less meaningful to participants and their families. A traditional dichotomisation point on the mRS would be 0-2 vs. 3-6, meaning that the transition from death to being able to walk on ones own would not be considered a successful treatment. In the Kansas City Stroke Study(n=459) it was noted that 65% of patients crossed one or more boundaries on the mRS by three month follow up. Dichotomisation of the mRS at 0-1 or 0-2 resulted in only 15% or 42% of participants respectively being considered as having a "good" outcome[152]. The optimal cut off for dichotomisation of the mRS scale is debated[153] and often chosen arbitrarily.

Dichotomisation (or Trichotomisation) is now widely accepted as a suboptimal method of analysing acute stroke trials and is recognised as statistically inefficient[154]. Collection of data on an ordinal scale such as the mRS with a view to collapsing it into a binary outcome discards potentially useful information, risking a loss of statistical power.

Of more concern is the possibility that we may fail to detect the harmful effects of treatment. Where the dichotomised threshold is set high (mRS 1 or 2) but the treatment (e.g. thrombolytic therapy) is in fact harmful to patients with more severe stroke the analysis will not detect a transition from moderate disability (mRS3) to severe disability or death (mRS

5/6). Perhaps dichotomising with a definition of "poor" outcome such as death or mRs above a certain cut off would be safer and more meaningful in this respect[155].

1.5.3.2. Responder Analysis - Prognosis adjusted endpoints

Stroke is a heterogeneous disease with a wide variety of outcomes. Setting an arbitrary and identical target deemed to signify "recovery" or "good outcome" for this varied population will almost certainly miss important clinical effects in a large trial. The principle of using prognosis adjusted end points, also known as responder analysis or a sliding dichotomy, encourages judgement of a treatment effect in response to the severity of the initial stroke insult[156, 157]. This allows trialists to individualise end points to reflect the baseline characteristics and prognosis as assessed by factors such as age, stroke severity, classification, baseline NIHSS, blood pressure etc.

This approach has been demonstrated to be effective in several types of study. Retrospective analyses of previous trial datasets (Tirilizad Head Injury Trials[157], NINDS tPA trials[158], Trial of Org 10172 in Acute Stroke Treatment TOAST trial and ECASS I and ECASS II[159]) statistical simulations using previous trial datasets (Glycine Antagonist in Neuroprotection GAIN International Trial[160]) and prospective phase III clinical trials (AbESTT-II[161] and STITCH[162]).

1.5.3.3. Shift Analysis – Ordinal Analysis

Analysis of change in outcome distribution across the full ordinal range of a scale is an attractive prospect. This approach allows analysis of all functional transitions, both beneficial and harmful. A shift analysis technique is being employed in ongoing large phase III trials and has been used in analysis of previous RCT's (Erythropoetin in Acute Stroke[163], SAINT[18], FAST-MAG[164] and IST3[165]).

The advantage of shift analysis is the enhancement in statistical power that is gained by using all available information[154]. This is particularly important where the hypothesised treatment effect is seen uniformly across the spectrum of stroke severity. It may be even more important to use all available data where there are no assumptions or prior knowledge of where in the stroke severity spectrum a treatment effect may be of benefit[166]. A

collaboration of stroke academics re-analysed the datasets of 47 trials using techniques designed to examine the unprocessed ordinal data. In every case they found that shift analysis techniques performed better, were more efficient and reliable[167]. Stroke trial power can be substantially enhanced by the use of an ordinal shift analysis technique with test statistics such as the Cochran-Mantel-Haenszel Test[104].

1.5.4. Non-Expert Interpretation of Trial Analysis

A change in mRS score at the top end of the scale is most often considered by physicians to represent a clinical or functional success[168]. However, economic analysis demonstrates benefit with a shift across each level in mRS score. A success in economic term is measured in terms of length of stay and institutional care[94]. Information about both ends of the spectrum is important in overall risk/benefit reporting of a new product after clinical trials.

The use of shift analysis techniques is mathematically and computationally more complex than those using dichotomised data. This can be dealt with using most modern statistical computer packages but the important challenge is in converting this spectrum of non binary outcomes into terms that are easily understood and interpreted by clinicians and patients. In this situation the number needed to treat (NNT) or number needed to harm (NNH) is traditionally used. NNT and NNH are widely used and statistically valid measures of a treatment effect, the resulting figure is clinically meaningful and can be used to indicate to patients and their families how many patients would require treatment in order for one to have a beneficial or harmful outcome.

Dichotomised analysis allows easy calculation of the NNT for each patient to cross the threshold to a “good” outcome. The calculation of NNT in a shift analysis is more complex; a transition across each grade in the scale may be associated with different magnitudes of clinical change (the scale is not linear). The degree of change across the entire population may not be uniform. In some cases a small number of patients improve considerably, in others the treatment effect may improve outcome to a modest degree for a greater proportion of patients[166]. A method of estimating NNT in an ordinal analysis is available[169]

and allows the calculation of a figure that represents the NNT in order for one patient to improve by one or more point on the mRS scale.

Overall there has been a marked change in attitudes to statistical techniques used in contemporary stroke research since the start of the twenty-first century. A greater understanding and appreciation of the various methods and their relative merits, even amongst non academic clinicians, has permitted this change in ideology[168].

1.6. Optimising Acute Stroke Trials

Improvements in acute stroke trials have been ongoing for many years. Coherent international collaboration, well designed protocols, appropriate training and certification of study investigators, adequate sample sizes and shortened time to treatment are a selection of the advances that are now considered ubiquitous in contemporary stroke research. There is no one best approach for statistical analysis in all acute stroke trials and it is crucial that each study is designed with appreciation of the population of patients to be enrolled, the realistic effect size and therefore the necessary sample size[156]. Despite meticulous planning, an inefficient outcome measure and statistical strategy can jeopardise trial outcome. It is increasingly accepted by contemporary trialists that the choice of outcome measure and statistical analysis plan may be paramount to a study's success[104].

1.7. Research questions

This programme of research and thesis aims to investigate an alternative method of collecting outcome data and strategies to optimise analysis of these data.

We present data exploring complementary themes, generated by the original research questions for the CARS study (Central Adjudication of modified Rankin Scale Assessments in Acute Stroke Trials):

- What are the expected benefits in trial power and study sample size with improved reliability of mRS and/or multiple mRS assessments?

- Does Central Adjudication of video recordings of mRS assessments:
    a. Provide a feasible method of measuring outcome in a multicentre trial setting?
    b. Offer a more accurate measure of outcome?
    c. Allow measurement of more subtle effect on outcome through grading of outcomes within mRS categories?

Chapter 2

Statistical Simulations: a study of the potential benefits of improved mRS reliability on study sample size.

2.1. Introduction

The global and unstructured nature of the mRs is a great advantage; without relying on individual activities of daily living there are no floor or ceiling effects in its application which are common to structured intsruments. Typically mRS assessment is based on a clinician's rating of a patient interview. This subjectivity contributes to inter-observer variability, which can be considerable. In a recent systematic review and meta-analysis, the overall reliability seen in the traditional mRS interview was κ 0.46 (κ_w 0.9)[99]. Few studies have quantified the extent and impact of this in a multicentre clinical trial but its reliability in this context is estimated to be poor (κ 0.25)[100]. Mandatory training in mRS assessment is employed in most trials to mitigate this[151] but several issues persist in international randomised controlled trials; the large number of assessors and cultural differences in activity, treatment and rehabilitation of stroke patients. Further, bias is possible with use of the mRs where observers are not or cannot be blinded to treatment allocation. Traditional dichotomised methods of outcome analysis disregard important differences between adjacent mRs groups[94]. There are a number of ways in which our use of the mRs could be improved. Inter-

observer variability is associated with endpoint misclassification which in turn can affect trial power[143] and treatment effect size[144, 145].

We aimed to describe the effect of varying magnitudes of mRS inter-observer variability and multiple mRS scores on required sample size for an acute stroke study using statistical modelling techniques.

2.2. Methods

We performed simulations to demonstrate the effect of increasing mRS reliability, considering both dichotomised and ordinal analysis approaches and using multiple raters to assign mRS scores. To do this we generated power estimates from simulated mRS studies under different combinations of sample size (N), effect size (δ), reliability (unweighted κ and quadratically weighted κ_w), dichotomised mRS outcomes, adjudication panel size (N_{adj}) and method of summarising mRS across adjudicators (mode, mean and median). In order to test the robustness of our simulation findings across a variety of settings, we based the proportions in each mRS category and the underlying disability distribution on 3 different scenarios; 1) the tPA (NINDS 0-3hrs) study[11], 2) the NXY059 study[18] and 3) derived from the standard normal distribution. A fixed treatment effect (i.e. not based on the therapeutic intervention in the above mentioned studies) was chosen to give approximately 80% power across N at the lowest common variable (κ/κ_w or N_{adj}). This treatment effect was subtracted from the disability value of the patients assigned to the treatment group. Power was defined as the proportion of simulated data sets where $P < 0.05$ in a test of equal mRS proportions between placebo and treatment groups (Mann-Whitney U test with continuity correction[170]). From each power estimate a proportion of the original sample size (the sample size needed at the lowest (κ/κ_w or N_{adj}) required was generated.

2.2.1. The effect of increasing mRS reliability

To assess the effect of improving reliability, the degree of inter-observer error in assigning disability on mRS was varied. A range of reliability based on previous studies of mRS reliability were used. Kappa (κ) / weighted kappa [κ_w] of (0.25[100])/[0.74], (0.5[86])/[0.92],

(0.7^{99})/[0.96], (0.9)/[0.99] and (1.0)/[1.0] was simulated by adding varying amounts of random statistical noise. Each combination of N, δ and κ/κ_w was simulated 100,000 times. Software used was R (version 2.13.0 for Unix). Parameter combinations that yielded power estimates >99% were deemed uninformative and removed.

2.2.2. The effect of using dichotomised outcomes

To assess the effect of dichotomised mRS outcome at different levels the above reliability simulations were performed using mRS dichotomised at 0-1, 0-2 and 0-3.

2.2.3. The effect of using multiple scores

To assess the benefit of multiple mRS assessments, simulations were performed using summary statistics (mode/mean/median) of N_{adj} = 1,2,4 and 9. Each combination of N, δ and N_{adj} was simulated 10,000 times. Reliability (κ / κ_w) was represented by the agreement in modal mRS between two independent panels of size N_{adj}. When calculating the mode, ties were resolved randomly. Parameter combinations that yielded power estimates >95% were deemed uninformative and removed.

2.3. Results

All planned simulations were successfully completed. Results are described for a typical phase III RCT of n=2000 using the tPA (NINDS 0-3hrs) dataset. Results for each mRS distribution [tPA (NINDS 0-3 hrs) dataset, NXY059 dataset and the hypothetical normal distribution] are displayed in tabulated form.

2.3.1. The effect of increasing mRS reliability

Simulations using the using the tPA (NINDS 0-3hrs) dataset suggest that improving reliability in mRS scoring from κ (κ_w) 0.25 (0.74) to 0.5 (0.92), 0.7 (0.96), 0.9 (0.99) up to hypothetical perfect agreement at κ 1.0 (1.0) could reduce sample size by n=386, n=490, n=488 and n=484 respectively. (Tables 2-4). There is a plateau seen at near perfect agreement beyond which there is no further potential reduction in sample size. This is seen in both the tPA (NINDS 0-3hrs dataset and the NXY059 dataset but not in the standard normal distribution.

2.3.2. The effect of using dichotomised outcomes

Simulations using the tPA (NINDS 0-3hrs) dataset suggest that the use of a dichotomised outcome at mRS 0-1, 0-2 or 0-3 increases the sample size required by n=843, n=230 or n=884 at baseline reliability of κ (κ_w) 0.25 (0.74). (Tables 2-4)

2.3.3. The effect of using multiple scores

Simulations using the tPA (NINDS 0-3hrs) dataset suggest that a mean mRS score of 2, 4 or 9 adjudicators can reduce sample size by n=54, n=172 and n=318 respectively. The use of the mode or median of multiple mRS assessments does not convey similar benefits. The modal mRS of up to 9 adjudicators will reduce sample size by a maximum of n=18. A median mRS performs slightly better with maximal reduction in sample size of n=60. (Tables 5-7)

Table 2 - Sample size simulations using tPA (NINDS 0-3hrs) study dataset: effect of increasing reliability in mRS and the use of dichotomised outcomes

Total Sample Size (N)	Effect Size (δ/SD)	Power (SE) [Proportion of N required to match power at κ=0.25][b] {Proportion of N required to match power at κ=0.25 with dichotomised outcomes (Fisher exact test)}[c]				
		κ = 0.25 κ_w = 0.74	κ = 0.50 κ_w = 0.92	κ = 0.70 κ_w = 0.96	κ = 0.9 κ_w = 0.99	κ = 1.0 κ_w = 1.0
50	0.83[a]	0.621 (0.001) [100.0%][b] {0-1 vs 2-6 115.4% 0-2 vs 3-6 157.7% 0-3 vs 4-6 638.5%}[c]	0.761 (0.001) [76.9%] {0-1 vs 2-6 80.8% 0-2 vs 3-6 134.6% 0-3 vs 4-6 850.0%}	0.792 (0.001) [69.2%] {0-1 vs 2-6 69.2% 0-2 vs 3-6 134.6% 0-3 vs 4-6 888.5%}	0.797 (0.001) [69.2%] 0-1 vs 2-6 65.4% 0-2 vs 3-6 134.6% 0-3 vs 4-6 896.2%}	0.797 (0.001) [69.2%] 0-1 vs 2-6 65.4% 0-2 vs 3-6 134.6% 0-3 vs 4-6 903.8%}
200	0.38[a]	0.650 (0.001) [100.0%] {0-1 vs 2-6 125.5% 0-2 vs 3-6 118.6% 0-3 vs 4-6 229.4%}	0.767(0.001) [76.5%] {0-1 vs 2-6 89.2% 0-2 vs 3-6 89.2% 0-3 vs 4-6 222.5%}	0.800 (0.001) [70.6%] {0-1 vs 2-6 73.5% 0-2 vs 3-6 87.3% 0-3 vs 4-6 219.6%}	0.800 (0.001) [70.6%] {0-1 vs 2-6 60.8% 0-2 vs 3-6 85.3% 0-3 vs 4-6 218.6%}	0.800 (0.001) [70.6%] {0-1 vs 2-6 59.8% 0-2 vs 3-6 85.3% 0-3 vs 4-6 218.6%}
1000	0.17[a]	0.675 (0.001) [100.0%] {0-1 vs 2-6 136.3% 0-2 vs 3-6 113.8% 0-3 vs 4-6 156.1%}	0.774 (0.001) [79.6%] {0-1 vs 2-6 107.4% 0-2 vs 3-6 85.4% 0-3 vs 4-6 130.3%}	0.800 (0.001) [74.5%] {0-1 vs 2-6 88.0% 0-2 vs 3-6 83.8% 0-3 vs 4-6 125.3%}	0.800 (0.001) [74.5%] {0-1 vs 2-6 69.1% 0-2 vs 3-6 82.4% 0-3 vs 4-6 124.4%}	0.798 (0.001) [74.7%] {0-1 vs 2-6 68.1% 0-2 vs 3-6 82.4% 0-3 vs 3-6 124.2%}
2000	0.12[a]	0.685 (0.001) [100.0%] {0-1 vs 2-6 141.7% 0-2 vs 3-6 115.2% 0-3 vs 3-6 144.2%}	0.776 (0.001) [80.7%] {0-1 vs 2-6 113.8% 0-2 vs 3-6 87.4% 0-3 vs 4-6 117.5%}	0.802(0.001) [75.5%] {0-1 vs 2-6 93.1% 0-2 vs 3-6 85.3% 0-3 vs 4-6 112.2%}	0.802 (0.001) [75.6%] {0-1 vs 2-6 71.5% 0-2 vs 3-6 84.2% 0-3 vs 4-6 110.7%}	0.801 (0.001) [75.8%] {0-1 vs 2-6 70.6% 0-2 vs 3-6 84.1% 0-3 vs 3-6 110.9%}

Power estimates from simulated mRS studies under different combinations of sample size (N), effect size (δ) and reliability (unweighted kappa κ and quadratically weighted κ_w). The mRS proportions and underlying disability distribution were based on the tPA (NINDS 0.3hrs) study. Each combination of N, δ and κ was simulated 1e+05 times. Power is defined as the proportion of simulated data sets where P<0.05 in a test of equal mRS proportions between placebo and treatments groups (Mann Whitney U test with continuity correction). Parameter combinations that yielded power estimates >99% were deemed uninformative and removed.

[a]Effect size chosen to give approximately 80% power across N at κ = 1.0 (not related to the effect size seen in the original trial)
[b]For each example sample size (n=50, 200, 1000 and 2000) the proportion of that sample size required at the baseline level of interobserver reliability (κ=0.25) is noted as 100%; the proportion of n that is required with improved interobserver reliability is noted in each column as a percentage of the original sample size
[c]Proportion of n required (expressed as percentage of original sample size at baseline interobserver reliability (κ=0.25) with use of dichotomised endpoint (0-1 vs 2-6 / 0-2 vs 3-6 / 0-3 vs 4-6)

Table 3 - Sample size simulations using NXY059 study dataset: effect of increasing reliability in mRS and the use of dichotomised outcomes

Total Sample Size (N)	Effect Size (δ/SD)	Power (SE) [Proportion of N required to match power at κ=0.25] [b] {Proportion of N required to match power at κ=0.25 with dichotomised outcomes (Fisher exact test)} [c]				
		κ = 0.25 κ_w = 0.74	κ = 0.50 κ_w = 0.92	κ = 0.70 κ_w = 0.96	κ = 0.9 κ_w = 0.99	κ = 1.0 κ_w = 1.0
50	0.80[a]	0.622 (0.001) [100.0%] [b] {0-1 vs 2-6 130.8% 0-2 vs 3-6 165.4% 0-3 vs 3-6 273.1%} [c]	0.787 (0.001) [76.9%] {0-1 vs 2-6 92.3% 0-2 vs 3-6 123.1% 0-3 vs 3-6 238.5%}	0.808 (0.001) [73.1%] {0-1 vs 2-6 84.6% 0-2 vs 3-6 115.4% 0-3 vs 3-6 234.6%}	0.804 (0.001) [76.9%] {0-1 vs 2-6 80.8% 0-2 vs 3-6 115.4% 0-3 vs 3-6 234.6%}	0.799 (0.001) [76.9%] {0-1 vs 2-6 80.8% 0-2 vs 3-6 115.4% 0-3 vs 3-6 234.6%}
200	0.38[a]	0.660 (0.001) [100.0%] {0-1 vs 2-6 125.5% 0-2 vs 3-6 131.4% 0-3 vs 3-6 164.7%}	0.787(0.001) [74.5%] {0-1 vs 2-6 86.3% 0-2 vs 3-6 90.2% 0-3 vs 3-6 125.5%}	0.811 (0.001) [70.6%] {0-1 vs 2-6 80.4% 0-2 vs 3-6 83.3% 0-3 vs 3-6 118.6%}	0.807 (0.001) [70.6%] {0-1 vs 2-6 77.5% 0-2 vs 3-6 81.4% 0-3 vs 3-6 116.7%}	0.804 (0.001) [71.6%] {0-1 vs 2-6 77.5% 0-2 vs 3-6 81.4% 0-3 vs 3-6 116.7%}
1000	0.17[a]	0.654 (0.001) [100.0%] {0-1 vs 2-6 127.2% 0-2 vs 3-6 119.6% 0-3 vs 3-6 138.0%}	0.783 (0.001) [74.2%] {0-1 vs 2-6 90.0% 0-2 vs 3-6 81.0% 0-3 vs 3-6 99.2%}	0.808 (0.001) [69.8%] {0-1 vs 2-6 83.4% 0-2 vs 3-6 74.6% 0-3 vs 3-6 93.2%}	0.806 (0.001) [70.0%] {0-1 vs 2-6 81.6% 0-2 vs 3-6 73.0% 0-3 vs 3-6 91.4%}	0.804 (0.001) [70.6%] {0-1 vs 2-6 81.6% 0-2 vs 3-6 72.8% 0-3 vs 3-6 91.2%}
2000	0.12[a]	0.655 (0.001) [100.0%] {0-1 vs 2-6 128.3% 0-2 vs 3-6 118.3% 0-3 vs 3-6 131.8%}	0.781 (0.001) [74.4%] {0-1 vs 2-6 90.9% 0-2 vs 3-6 80.6% 0-3 vs 3-6 95.0%}	0.806(0.001) [70.1%] {0-1 vs 2-6 84.1% 0-2 vs 3-6 73.9% 0-3 vs 3-6 89.2%}	0.805 (0.001) [70.0%] {0-1 vs 2-6 82.1% 0-2 vs 3-6 72.1% 0-3 vs 3-6 87.3%}	0.803 (0.001) [70.5%] {0-1 vs 2-6 81.8% 0-2 vs 3-6 72.0% 0-3 vs 3-6 87.1%}

Power estimates from simulated mRS studies under different combinations of sample size (N), effect size (δ) and reliability (unweighted kappa κ and quadratically weighted κ_w). The mRS proportions and underlying disability distribution were based on the NXY059 study. Each combination of N, δ and κ was simulated 1e+05 times. Power is defined as the proportion of simulated data sets where P<0.05 in a test of equal mRS proportions between placebo and treatments groups (Mann Whitney U test with continuity correction). Parameter combinations that yielded power estimates >99% were deemed uninformative and removed.

[a]Effect size chosen to give approximately 80% power across N at κ = 1.0 (not related to the effect size seen in the original trial)
[b]For each example sample size (n=50, 200, 1000 and 2000) the proportion of that sample size required at the baseline level of interobserver reliability (κ=0.25) is noted as 100%; the proportion of n that is required with improved interobserver reliability is noted in each column as a percentage of the original sample size
[c]Proportion of n required (expressed as percentage of original sample size at baseline interobserver reliability (κ=0.25) with use of dichotomised endpoint (0-1 vs 2-6 / 0-2 vs 3-6 / 0-3 vs 4-6)

Table 4 - Sample size simulations using the standard normal distribution: effect of increasing reliability in mRS and the use of dichotomised outcomes

Total Sample Size (N)	Effect Size (δ/SD)	Power (SE) [Proportion of N required to match power at κ=0.25] [b] {Proportion of N required to match power at κ=0.25 with dichotomised outcomes (Fisher exact test)} [c]				
		κ = 0.25 κ_w = 0.74	κ = 0.50 κ_w = 0.92	κ = 0.70 κ_w = 0.96	κ = 0.9 κ_w = 0.99	κ = 1.0 κ_w = 1.0
50	0.86[a]	0.646 (0.001) [100.0%] [b] {0-1 vs 2-6 153.8% 0-2 vs 3-6 176.9% 0-3 vs 3-6 238.5%} [c]	0.760 (0.001) [80.8%] {0-1 vs 2-6 119.2% 0-2 vs 3-6 142.3% 0-3 vs 3-6 192.3%}	0.781 (0.001) [76.9%] {0-1 vs 2-6 115.4% 0-2 vs 3-6 134.6% 0-3 vs 3-6 184.6%}	0.790 (0.001) [76.9%] {0-1 vs 2-6 111.5% 0-2 vs 3-6 134.6% 0-3 vs 3-6 180.8%}	0.790 (0.001) [76.9%] {0-1 vs 2-6 111.5% 0-2 vs 3-6 130.8% 0-3 vs 3-6 180.8%}
200	0.42[a]	0.667 (0.001) [100.0%] {0-1 vs 2-6 156.4% 0-2 vs 3-6 156.4% 0-3 vs 3-6 181.2%}	0.775(0.001) [78.2%] {0-1 vs 2-6 122.8% 0-2 vs 3-6 122.8% 0-3 vs 3-6 145.5%}	0.794 (0.001) [75.2%] {0-1 vs 2-6 116.8% 0-2 vs 3-6 116.8% 0-3 vs 3-6 140.6%}	0.801 (0.001) [74.3%] {0-1 vs 2-6 114.9% 0-2 vs 3-6 115.8% 0-3 vs 3-6 138.6%}	0.802 (0.001) [73.3%] {0-1 vs 2-6 114.9% 0-2 vs 3-6 114.9% 0-3 vs 3-6 138.6%}
1000	0.19[a]	0.669 (0.001) [100.0%] {0-1 vs 2-6 154.7% 0-2 vs 3-6 145.5% 0-3 vs 3-6 161.3%}	0.779 (0.001) [77.4%] {0-1 vs 2-6 120.4% 0-2 vs 3-6 113.6% 0-3 vs 3-6 127.7%}	0.798 (0.001) [73.9%] {0-1 vs 2-6 115.0% 0-2 vs 3-6 108.6% 0-3 vs 3-6 122.8%}	0.805 (0.001) [72.7%] {0-1 vs 2-6 113.6% 0-2 vs 3-6 106.6% 0 – 3 121.0%}	0.806 (0.001) [72.5%] {0-1 vs 2-6 113.6% 0-2 vs 3-6 106.4% 0-3 vs 3-6 120.4%}
2000	0.13[a]	0.676 (0.001) [100.0%] {0-1 vs 2-6 152.1% 0-2 vs 3-6 145.2% 0-3 vs 3-6 156.5%}	0.786 (0.001) [77.3%] {0-1 vs 2-6 120.0% 0-2 vs 3-6 112.9% 0-3 vs 3-6 124.4%}	0.804 (0.001) [73.9%] {0-1 vs 2-6 115.1% 0-2 vs 3-6 108.0% 0-3 vs 3-6 119.3%}	0.811 (0.001) [72.6%] {0-1 vs 2-6 113.2% 0-2 vs 3-6 106.2% 0-3 vs 3-6 117.7%}	0.812 (0.001) [72.5%] {0-1 vs 2-6 112.9% 0-2 vs 3-6 106.1% 0-3 vs 3-6 117.8%}

Power estimates from simulated mRS studies under different combinations of sample size (N), effect size (δ) and reliability (unweighted kappa κ and quadratically weighted κ_w). The mRS proportions and underlying disability distribution were derived from the standard normal distribution. Each combination of N, δ and κ was simulated 1e+05 times. Power is defined as the proportion of simulated data sets where P<0.05 in a test of equal mRS proportions between placebo and treatments groups (Mann Whitney U test with continuity correction). Parameter combinations that yielded power estimates >99% were deemed uninformative and removed.

[a]Effect size chosen to give approximately 80% power across N at κ = 1.0

[b]For each example sample size (n=50, 200, 1000 and 2000) the proportion of that sample size required at the baseline level of interobserver reliability (κ=0.25) is noted as 100%; the proportion of n that is required with improved interobserver reliability is noted in each column as a percentage of the original sample size

[c]Proportion of n required (expressed as percentage of original sample size at baseline interobserver reliability (κ=0.25) with use of dichotomised endpoint (0-1 vs 2-6 / 0-2 vs 3-6 / 0-3 vs 4-6)

Table 5 - Sample size simulations using tPA (NINDS 0-3hrs) study dataset: effect of multiple scores (mode / mean / median)

Total Sample Size (N)	Effect Size (δ/SD)	Summary Statistic	Power (SE) [Proportion of N required to match power at N_{adj} = 1] [b]			
			N_{adj} = 1 κ = 0.7 κ_w = 0.96	N_{adj} = 2 κ = 0.7 κ_w = 0.96	N_{adj} = 4 κ = 0.78 κ_w = 0.97	N_{adj} = 9 κ = 0.86 κ_w = 0.98
50	0.81[a]	Mode	0.789 (0.004) [100.0%] [b]	0.778 (0.004) [100.0%]	0.786 (0.004) [96.4%]	0.786 (0.004) [96.4%]
		Mean	0.789 (0.004) [100.0%]	0.808 (0.004) [92.9%]	0.840 (0.004) [85.7%]	0.868 (0.004) [82.1%]
		Median	0.789 (0.004) [100.0%]	0.808 (0.004) [92.9%]	0.804 (0.004) [92.9%]	0.786 (0.004) [96.4%]
200	0.40[a]	Mode	0.838 (0.004) [100.0%]	0.834 (0.004) [101.0%]	0.842 (0.004) [99.0%]	0.841 (0.004) [99.0%]
		Mean	0.838 (0.004) [100.0%]	0.849 (0.004) [96.2%]	0.874 (0.004) [90.4%]	0.901 (0.004) [83.7%]
		Median	0.838 (0.004) [100.0%]	0.849 (0.004) [96.2%]	0.852 (0.004) [96.2%]	0.841 (0.004) [99.0%]
1000	0.18[a]	Mode	0.821 (0.004) [100.0%]	0.820 (0.004) [99.8%]	0.828 (0.004) [98.6%]	0.823 (0.004) [99.0%]
		Mean	0.821 (0.004) [100.0%]	0.829 (0.004) [97.0%]	0.856 (0.004) [90.9%]	0.881 (0.004) [83.9%]
		Median	0.821 (0.004) [100.0%]	0.829 (0.004) [97.0%]	0.832 (0.004) [96.4%]	0.823 (0.004) [99.0%]
2000	0.13[a]	Mode	0.806 (0.004) [100.0%]	0.804 (0.004) [100.0%]	0.807 (0.004) [99.1%]	0.811 (0.004) [99.2%]
		Mean	0.806 (0.004) [100.0%]	0.815 (0.004) [97.3%]	0.840 (0.004) [91.4%]	0.867 (0.004) [84.1%]
		Median	0.806 (0.004) [100.0%]	0.815 (0.004) [97.3%]	0.815 (0.004) [97.0%]	0.811 (0.004) [99.2%]

Power estimates from simulated mRS studies under different combinations of sample size (N), effect size (δ), adjudication panel size (N_{adj}) and associated reliability (unweighted κ and quadratically weighted κ_w), and method of summarising mRS across adjudicators (mode, mean and median). The mRS proportions and underlying disability distribution were based on the tPA (NINDS 0-3hrs) study. Each combination of N, δ and N_{adj} was simulated 10,000 times. Power is defined as the proportion of simulated data sets where P < 0.05 in a test of equal mRS proportions between placebo and treatment groups (Man-Whitney U test with continuity correction). Reliability represents the agreement in modal mRS between two independent panels of size N_{adj}. When calculating the mode, ties were resolved randomly. Parameter combinations that yielded power estimates >95% were deemed uninformative and removed.

[a]Effect size chosen to give approximately 80% power across N at N_{adj} = 1(not related to the effect size seen in the original trial)

[b]For each example sample size (n=50, 200, 1000 and 2000) the proportion of that sample size required at the baseline level of interobserver reliability (κ=0.25) is noted as 100%; the proportion of n that is required with improved interobserver reliability is noted in each column as a percentage of the original sample size

Table 6 - Sample size simulations using NXY059 study dataset: effect of multiple scores (mode / mean / median)

Total Sample Size (N)	Effect Size (δ/SD)	Summary Statistic	Power (SE) [Proportion of N required to match power at N_{adj} = 1] [b]			
			N_{adj} = 1 κ = 0.7 κ_w = 0.96	N_{adj} = 2 κ = 0.7 κ_w = 0.96	N_{adj} = 4 κ = 0.78 κ_w = 0.97	N_{adj} = 9 κ = 0.86 κ_w = 0.98
50	0.81[a]	Mode	0.810 (0.004) [100.0%] [b]	0.810 (0.004) [100.0%]	0.811 (0.004) [100.0%]	0.808 (0.004) [103.7%]
		Mean	0.810 (0.004) [100.0%]	0.816 (0.004) [100.0%]	0.825 (0.004) [96.3%]	0.836 (0.004) [96.3%]
		Median	0.810 (0.004) [100.0%]	0.816 (0.004) [100.0%]	0.811 (0.004) [100.0%]	0.808 (0.004) [103.7%]
200	0.40[a]	Mode	0.841 (0.004) [100.0%]	0.845 (0.004) [101.0%]	0.843 (0.004) [100.0%]	0.839 (0.004) [100.0%]
		Mean	0.841 (0.004) [100.0%]	0.842 (0.004) [96.2%]	0.843 (0.004) [99.0%]	0.851 (0.004) [97.1%]
		Median	0.841 (0.004) [100.0%]	0.842 (0.004) [96.2%]	0.840 (0.004) [100.0%]	0.839 (0.004) [100.0%]
1000	0.18[a]	Mode	0.838 (0.004) [100.0%]	0.839 (0.004) [99.8%]	0.840 (0.004) [99.6%]	0.840 (0.004) [100.0%]
		Mean	0.838 (0.004) [100.0%]	0.836 (0.004) [100.4%]	0.841 (0.004) [99.6%]	0.848 (0.004) [97.8%]
		Median	0.838 (0.004) [100.0%]	0.836 (0.004) [100.4%]	0.838 (0.004) [100.2%]	0.840 (0.004) [100.0%]
2000	0.13[a]	Mode	0.841 (0.004) [100.0%]	0.840 (0.004) [99.9%]	0.844 (0.004) [99.6%]	0.842 (0.004) [99.9%]
		Mean	0.841 (0.004) [100.0%]	0.840 (0.004) [100.2%]	0.842 (0.004) [99.5%]	0.849 (0.004) [97.6%]
		Median	0.841 (0.004) [100.0%]	0.840 (0.004) [100.2%]	0.839 (0.004) [100.2%]	0.843 (0.004) [99.9%]

Power estimates from simulated mRS studies under different combinations of sample size (N), effect size (δ), adjudication panel size (N_{adj}) and associated reliability (unweighted κ and quadratically weighted κ_w), and method of summarising mRS across adjudicators (mode, mean and median). The mRS proportions and underlying disability distribution were based on the NZY059 study. Each combination of N, δ and N_{adj} was simulated 10,000 times. Power is defined as the proportion of simulated data sets where P < 0.05 in a test of equal mRS proportions between placebo and treatment groups (Man-Whitney U test with continuity correction). Reliability represents the agreement in modal mRS between two independent panels of size $N_{adj.}$ When calculating the mode, ties were resolved randomly. Parameter combinations that yielded power estimates >95% were deemed uninformative and removed.

[a]Effect size chosen to give approximately 80% power across N at N_{adj} = 1 (not related to the effect size seen in the original trial)

[b]For each example sample size (n=50, 200, 1000 and 2000) the proportion of that sample size required at the baseline level of interobserver reliability (κ=0.25) is noted as 100%; the proportion of n that is required with improved interobserver reliability is noted in each column as a percentage of the original sample size

Table 7 - Sample size simulations using the standard normal distribution: effect of multiple scores (mode / mean / median)

Total Sample Size (N)	Effect Size (δ/SD)	Summary Statistic	Power (SE) [Proportion of N required to match power at N_{adj} = 1] [b]			
			N_{adj} = 1 κ = 0.7 κ_w = 0.96	N_{adj} = 2 κ = 0.7 κ_w = 0.96	N_{adj} = 4 κ = 0.78 κ_w = 0.97	N_{adj} = 9 κ = 0.86 κ_w = 0.98
50	0.81[a]	Mode	0.736 (0.004) [100.0%] [b]	0.739 (0.004) [100.0%]	0.741 (0.004) [100.0%]	0.745 (0.004) [96.3%]
		Mean	0.736 (0.004) [100.0%]	0.748 (0.004) [96.3%]	0.756 (0.004) [96.3%]	0.758 (0.004) [96.3%]
		Median	0.736 (0.004) [100.0%]	0.748 (0.004) [96.3%]	0.746 (0.004) [96.3%]	0.746 (0.004) [96.3%]
200	0.40[a]	Mode	0.752 (0.004) [100.0%]	0.752 (0.004) [100.0%]	0.760 (0.004) [99.0%]	0.762 (0.004) [98.0%]
		Mean	0.752 (0.004) [100.0%]	0.760 (0.004) [98.0%]	0.769 (0.004) [96.1%]	0.773 (0.004) [95.1%]
		Median	0.752 (0.004) [100.0%]	0.760 (0.004) [98.0%]	0.762 (0.004) [98.0%]	0.762 (0.004) [98.0%]
1000	0.18[a]	Mode	0.743 (0.004) [100.0%]	0.747 (0.004) [100.0%]	0.748 (0.004) [99.0%]	0.753 (0.004) [98.0%]
		Mean	0.743 (0.004) [100.0%]	0.757 (0.004) [97.6%]	0.761 (0.004) [96.2%]	0.763 (0.004) [95.8%]
		Median	0.743 (0.004) [100.0%]	0.757 (0.004) [97.6%]	0.756 (0.004) [97.8%]	0.753 (0.004) [98.0%]
2000	0.13[a]	Mode	0.755 (0.004) [100.0%]	0.756 (0.004) [100.1%]	0.763 (0.004) [99.0%]	0.765 (0.004) [98.4%]
		Mean	0.755 (0.004) [100.0%]	0.765 (0.004) [97.8%]	0.773 (0.004) [96.4%]	0.777 (0.004) [95.5%]
		Median	0.755 (0.004) [100.0%]	0.765 (0.004) [97.8%]	0.766 (0.004) [97.7%]	0.765 (0.004) [98.4%]

Power estimates from simulated mRS studies under different combinations of sample size (N), effect size (δ), adjudication panel size (N_{adj}) and associated reliability (unweighted κ and quadratically weighted κ_w), and method of summarising mRS across adjudicators (mode, mean and median). The mRS proportions and underlying disability distribution were derived from the standard normal distribution. Each combination of N, δ and N_{adj} was simulated 10,000 times. Power is defined as the proportion of simulated data sets where P < 0.05 in a test of equal mRS proportions between placebo and treatment groups (Man-Whitney U test with continuity correction). Reliability represents the agreement in modal mRS between two independent panels of size $N_{adj.}$ When calculating the mode, ties were resolved randomly. Parameter combinations that yielded power estimates >95% were deemed uninformative and removed.

[a]Effect size chosen to give approximately 80% power across N at N_{adj} = 1

[b]For each example sample size (n=50, 200, 1000 and 2000) the proportion of that sample size required at the baseline level of interobserver reliability (κ=0.25) is noted as 100%; the proportion of n that is required with improved interobserver reliability is noted in each column as a percentage of the original sample size

2.4. Conclusions

We found significant potential for reducing required trial sample size and / or increasing trial power through improving reliability of mRs assessments. A reduction in sample size of 20% to 25% was seen in each simulation with improvement of mRS reliability from baseline κ 0.25 to κ 0.5 or κ 0.7.

Previous studies assessing the inter-rater reliability of the mRS have predominantly been conducted in small, single centre studies with highly motivated individuals. Inter-observer reliability in a large scale clinical trial with the associated challenges; multicentre / multicultural / assessors from different professional backgrounds and level of experience; is likely to be poorer.

Simulations using real life mRS distributions from previous phase three clinical trials suggest that improving inter-rater reliability in mRS assessment may have positive effects on sample size. The use of multiple mRS assessments has similar beneficial effects in simulation. Interestingly we see a plateau in this improvement at near perfect agreement (κ (κ_w) 0.9 (0.99) and κ 1.0). This is not seen in the simulations using a hypothetical normal distribution. We propose that this is due to a statistical phenomenon; stochastic resonance[171]: small levels of variability or noise are present in the "real life" mRS distributions that are not present in the smooth normal distribution which lowers the agreement threshold at extremes of the scale.

The potential loss of valuable mRS information where a dichotomised mRS outcome is used is well documented[154]. Our simulations quantify the increase in sample size that would be required with dichotomisation at mRS boundaries 0-1, 0-2 and 0-3. The optimal cut off for dichotomisation is debated and often chosen arbitrarily[153] and our data further support the assertion that dichotomised mRS outcomes offer a sub optimal method of analysing acute stroke data and is statistically inefficient.

Studies with a sample size larger than necessary are economically and ethically unjustified. The potential reduction in sample size seen in these simulations may have a substantial

positive financial implication but may also deliver important ethical benefits. We must involve as few participants in research as necessary to provide the evidence required to confirm or refute our hypotheses. In a standard phase III RCT of n=2000 the potential saving might run to millions of pounds if the required sample size is cut by 25%.

We have demonstrated through simulations that there are benefits to be reaped if the use of the mRS as an outcome measure can be optimised. There are strengths and weaknesses to this analysis. We used two trial datasets to assess generalisability and found broad agreement between results. However, although based on real life trial datasets, simulations are no comparison to real life data. They allow us to speculate that there may be improvements in required sample size but it remains to be seen if these benefits might translate into real life economic and ethical savings in the future.

Chapter 3

The methodology, design and conduct of the CARS trial: Central Adjudication of modified Rankin Scale disability assessments in acute stroke trials.

3.1. Introduction

The importance of a robust functional outcome measure in acute stroke trials is evident. As we have discussed, the use of consistent and reproducible outcome scales has become a mandatory element of good trial design[6], the most commonly used outcome measure being the modified Rankin scale (mRS)[29]

For many years, aided by the development of fast and easy data transfer through secure electronic means, trialists in all medical fields have preferred to include electronic data collection and remote adjudication of their chosen study outcome where possible. Central reading of electrocardiograms, radiological images or mortality data is common practice in randomised controlled trials (RCTs) and is encouraged where the data is in a format that is

easily transferred to an independent central adjudicator or adjudication committee. This conveys several benefits including standardisation, quality control and reliable blinding. There is also a reduction in the number of trial investigators responsible for assigning outcome, likely to reduce the variability in scoring (reduced endpoint misclassification) and limit the chance of the study suffering the effects of type II error in data analysis. Despite the trend towards the use of a functional outcome as the primary endpoint in acute stroke trials, the concept of central adjudication of functional outcome has not been explored. We sought to investigate the feasibility, acceptability, validity and reliability of central adjudication of modified Rankin scale (mRS) assessments in a multicentre trial setting.

3.2. The CARS study - Central Adjudication of modified Rankin Scale disability assessments in acute stroke trials.

A "virtual" multicentre acute stroke trial was performed to assess the practicalities of using centrally adjudicated outcomes as a primary trial endpoint. The study was designed and conducted to emulate a typical contemporary acute stroke trial but there was no randomisation or intervention. We aimed to clarify the extent of possible endpoint misclassification and establish the validity and reliability of mRS outcomes assigned centrally.

Ethical approval for all study procedures was granted by Scotland A Research Ethics Committee (08/MRE00/72) and Essex 2 Research Ethics Committee (08/H0203/147) for study sites in Scotland and England/Wales respectively. The study was funded by the Chief Scientist Office (CSO reference number CBZ/4/595) and was supported by the UK Stroke Research Network.

3.3. Primary Research Questions

Our primary research questions were as follows;

Does central adjudication of video recordings of mRS assessments:

1. Provide a feasible method of measuring outcome in a multicentre trial setting?

2. Offer a more accurate measure of outcome?

i.e. Does outcome from central adjudication correlate better than on site raters' assessments with factors known to influence outcome (such as baseline NIHSS, glucose and blood pressure)

3. Allow measurement of more subtle effects on outcome?

i.e. Through grading of outcomes within mRS categories?

3.4. Trial Design

3.4.1. Study Population

Participants with a diagnosis of acute stroke of any aetiology who presented within 48 hours of ictus and with a demonstrable deficit on the National Institutes of Health Stroke Scale (NIHSS) were considered for inclusion in the study. Participants who had prior disability were excluded from the study in line with standard practice for acute stroke treatment trials.

Inclusion and exclusion criteria are detailed in Figure 7.

Inclusion Criteria:	1. Diagnosis of Acute Stroke (ischaemic or haemorrhagic) 2. Recruitment within 48 hours of stroke onset *3.* Demonstrable deficit on NIHSS (greater than 1)
Exclusion Criteria:	1. Pre-morbid mRS score of ≥3

Figure 7 CARS study Inclusion and Exclusion Criteria

We aimed to recruit a minimum of 300 participants. This sample size was chosen as it reflects the minimum number of participants likely to be required in a phase III acute stroke trial of a reperfusion strategy and could provide sufficient video assessments to enable us to evaluate

the technique and its impact upon trial design. We expected over 240 final assessments after accounting for mortality and withdrawals.

3.4.2. Consent Procedure

Participants were identified by a member of the treating clinical team as soon as possible after admission to the stroke unit. Suitable participants and/or their nearest relative were approached by a member of the clinical team with a brief description of the study. Where an interest in participation was noted, a research team member explained the study fully and obtained informed consent. Consent was sought as soon as possible after the stroke event in order to assess the practicalities of recruitment to a study of this nature. However, as no study specific procedures were involved within the initial days of admission, participants and relatives were given up to 48 hours within which to decide.

Research trials in acute stroke are frequently specifically designed to assist adults who are unable to consent for themselves. This is because the inherent nature of stroke disease often affects communication or cognitive abilities. For this reason, where individuals were unable to consent for themselves, assent from a relative / welfare guardian was sought. In practical terms the early procedures were indistinguishable from normal clinical care and only the later study procedures were considered more intrusive. If a participant subsequently regained the ability to provide their own consent they were asked to do so at future follow up visits and were given the opportunity to discontinue participation in the study if desired.

3.4.3. Study Centres

The study was co-ordinated from the University of Glasgow, based in the Institute of Cardiovascular and Medical Sciences, Western Infirmary. Fourteen centres were involved in the study, six in Scotland and eight in England/Wales. The study was adopted by the UK Stroke Research Network. Study centres are detailed in Figure 8.

Scotland Scotland A REC (08/MRE00/72)	Aberdeen Royal Infirmary
	Glasgow Royal Infirmary
	Ninewells Hospital, Dundee
	Southern General Hospital, Glasgow
	Stobhill Hospital, Glasgow
	Western Infirmary, Glasgow
England / Wales Essex 2 REC (08/H0203/147)	Countess of Chester Hospital, Chester
	Cumberland Infirmary, Carlisle
	Harrogate District General Hospital
	Leeds General Infirmary
	Royal Devon and Exeter Hospital
	Royal Glamorgan Hospital
	Wakefield (Pinderfields) Hospital
	York Hospital

Figure 8 CARS study centres

3.4.4. Study Investigators

Investigators at each site included medical staff, research nursing staff and allied health professionals involved in the care of acute stroke patients. A delegation log of investigators was held at each site and copied to the co-ordinating centre. All investigators were trained and certified in the use of the mRS and NIHSS assessment through trainingcampus©, an

internet based training and certification resource[172, 173]. Investigators were visited at each centre by a member of the co-ordinating study team to initiate the site and provide training in the study procedures and equipment. Written guidance was prepared and provided to all investigators. (See Appendix 1)

3.4.5. Study Procedures

Baseline data were collected at the time of recruitment including demographics, stroke subtype, NIHSS, medications, blood pressure, blood results and imaging results. No intervention or change to normal routine clinical care occurred during the study. Endpoint assessments were performed as they would be in an interventional trial (mRS, NIHSS, Serious Adverse Events (SAE), medications and home time). Two endpoint assessment visits were carried out at 30 and 90 days following the stroke event. The sequence of study procedures is outlined in Figure 9. Data were recorded locally on paper case report forms and also entered into an electronic case report form (eCRF) for remote access by the co-ordinating centre.

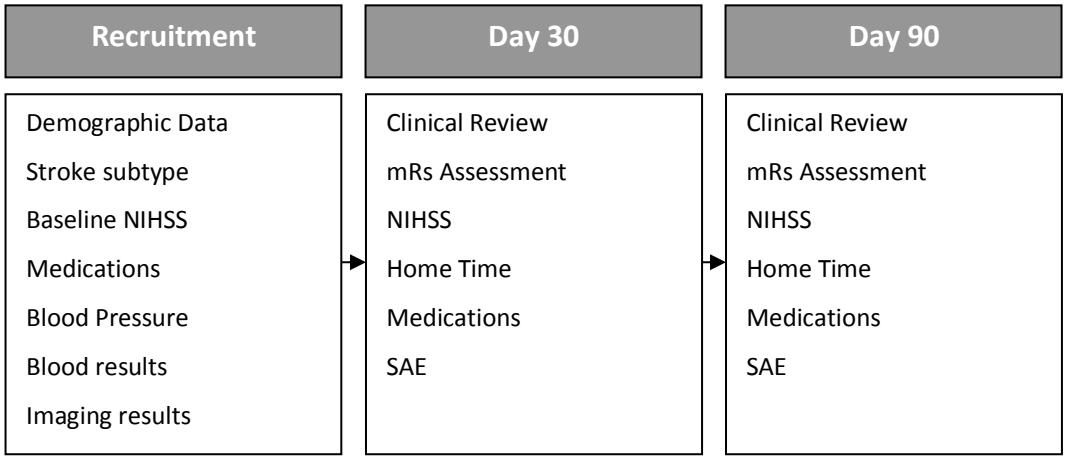


Figure 9 - CARS study procedure flowchart

3.4.5.1. Baseline Factors know to affect stroke outcome

Baseline stroke severity is well recognised as a predictor of functional outcome after stroke. Measurement of stroke severity using the NIHSS is the most useful marker to predict poor functional outcome[174, 175]. Hypertension is common in the stroke population and often diagnosed only after presentation with a stroke event. Regardless of whether hypertension has been previously diagnosed or is found only at the time of admission, there is a well documented association between baseline blood pressure and poor functional outcome[176, 177]. Diabetes is a recognised risk factor for stroke and similarly raised admission blood glucose level, whether previously diagnosed as diabetes / glucose intolerance or an incidental finding is associated with poor functional outcome[178-180].

We collected these data at baseline in order to compare local and centrally adjudicated mRS as an outcome measure relative to these factors previously known to be associated with functional outcome.

3.4.5.2. The modified Rankin scale assessment

The mRS assessment was performed in standard fashion according to normal practice at each centre. Assessments were performed in hospital wards, outpatient clinics or in participants' homes depending upon their location at the time of follow up. The mRS was recorded using a digital video camera. Where a participant was unable to be involved in an interview, due to significant disability or communication difficulties, a proxy was used. The chosen proxy was a relative, friend or health care provider with a good knowledge of the participant's functional capabilities.

The mRS video clip contained an image of the participant and/or their proxy. The assessor was not seen in the clip but their voice was audible. Anonymity was otherwise preserved. The recommended position of the camera in relation to the participant and assessor is demonstrated in Figure 10.

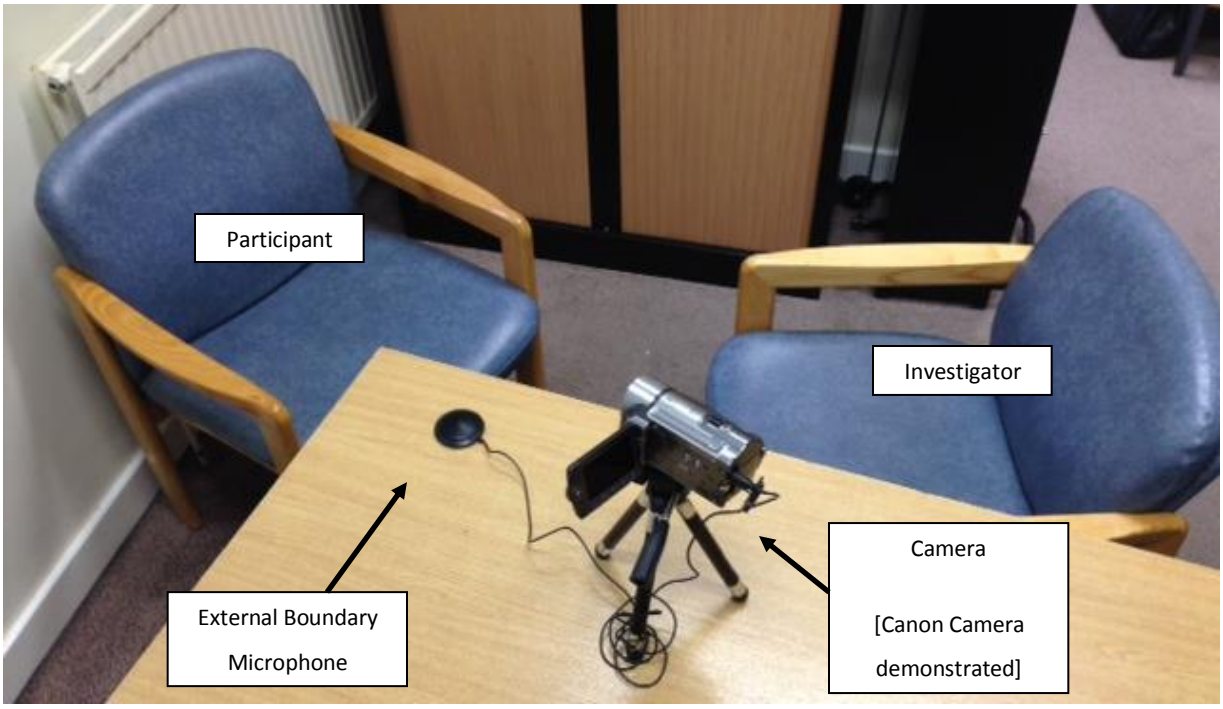


Figure 10 - Recommended position of participants, camera and microphone for mRS video assessment

The local investigator assigned a mRS score for the patient and entered this into the eCRF (the standard / local mRS score). Where possible the investigator remained constant across the follow-up period for a given patient, although we recognised that this was not possible in all cases, particularly in smaller sites with fewer investigators. In centres where it was standard practice to perform a structured mRS interview, an area was made available in the eCRF for entry of this score. The recorded mRS interview was uploaded to the eCRF through an internet based portal. These assessments were then reviewed by the co-ordinating centre according to a standard procedure, resulting in entry of the adjudicated mRS score (see section 3.4.6.2).

3.4.5.3. The NIHSS assessment

NIHSS assessments were performed in the standard fashion at each site. All investigators were trained and certified in the administration of this assessment tool.

3.4.5.4. Home Time Assessment

As a relatively novel but previously validated stroke outcome measure[181, 182] we included assessment of "Home Time" as a further comparative outcome measure. Home time is a measurement of the duration of time that a patient lives independently in the community following a stroke event. This is more useful than a simple measure of inpatient stay as shorter hospital stay does not always reflect a good outcome and in many cases may be associated with a poor outcome, for example death or early transfer to a long term institutional care facility. We collected Home Time data at 30 and 90 day follow up visits.

3.4.5.5. Recording of Serious Adverse Events

As an observational trial, recording of adverse events was not a legal requirement in line with good clinical practice (GCP) guidelines. However we aimed to co-ordinate the study in a similar manner to a clinical intervention trial and therefore in order emulate the practicalities of this as closely as possible we included collection of serious adverse events (SAE) in our case report form. An additional aim of SAE reporting was to help us to better identify individuals where the video technique was unsuitable or problematic. A serious adverse event was defined as an event that: a) results in death, b) is life threatening, c) requires hospitalisation or prolongation of existing hospitalisation, d) results in persistent or significant disability or incapacity or e) consists of a congenital anomaly or birth defect.

3.4.5.6. Study Withdrawal / Completion

Study completion was indicated to the co-ordinating centre via the eCRF after the 90 day visit. Where a participant chose not to return for follow up they were withdrawn from the study. The date of last contact with the participant and a reason for withdrawal (where available) was detailed in the eCRF. In this case participants were asked if they were happy for the data collected prior to withdrawal to be used in analysis. If they chose to withdraw consent entirely their data was not used in future analysis - this was clearly documented in the eCRF.

3.4.6. Review of Video mRS assessments

The review of mRS video assessments was co-ordinated through the web based portal by the study outcomes manager.

3.4.6.1. Handling of mRS clips

After initial upload of a video clip through the CARS web portal (see section 3.4.6.3) the study outcomes manager was notified by automated email. The clip was assessed for technical adequacy and anonymity.

If the clip was deemed inadequate, either because of technical difficulties (e.g. poor sound quality) or deficiencies in the mRS interview (e.g. inadequate information / questioning of participant) a repeat was requested from the local investigator by automated email. Any participant-identifying information was removed by the outcomes manager using Windows Movie Maker© (Microsoft, USA) video editing software. No other editing of video clips was performed. The nature of any editing performed and the reasons for this was recorded in the eCRF.

3.4.6.2. The CARS Endpoint Committee

The CARS endpoint committee was composed of seven experienced mRS assessors. All were physicians with a clinical and academic interest in stroke medicine (three professors, two clinical lecturers and two clinical research fellows).

Adequate mRS video assessments were distributed by the outcomes manager to a minimum of four endpoint committee members for independent, blinded mRS scoring. Video assessments were scored fully blinded to all other participant information. Where these scores agreed, the collective endpoint committee mRS score was assigned for that clip. Where there was a disagreement, the clip was “misclassified” and forwarded to the full endpoint committee for group review and consensus scoring where possible. The mRS clip review protocol is summarised in Figure 11.

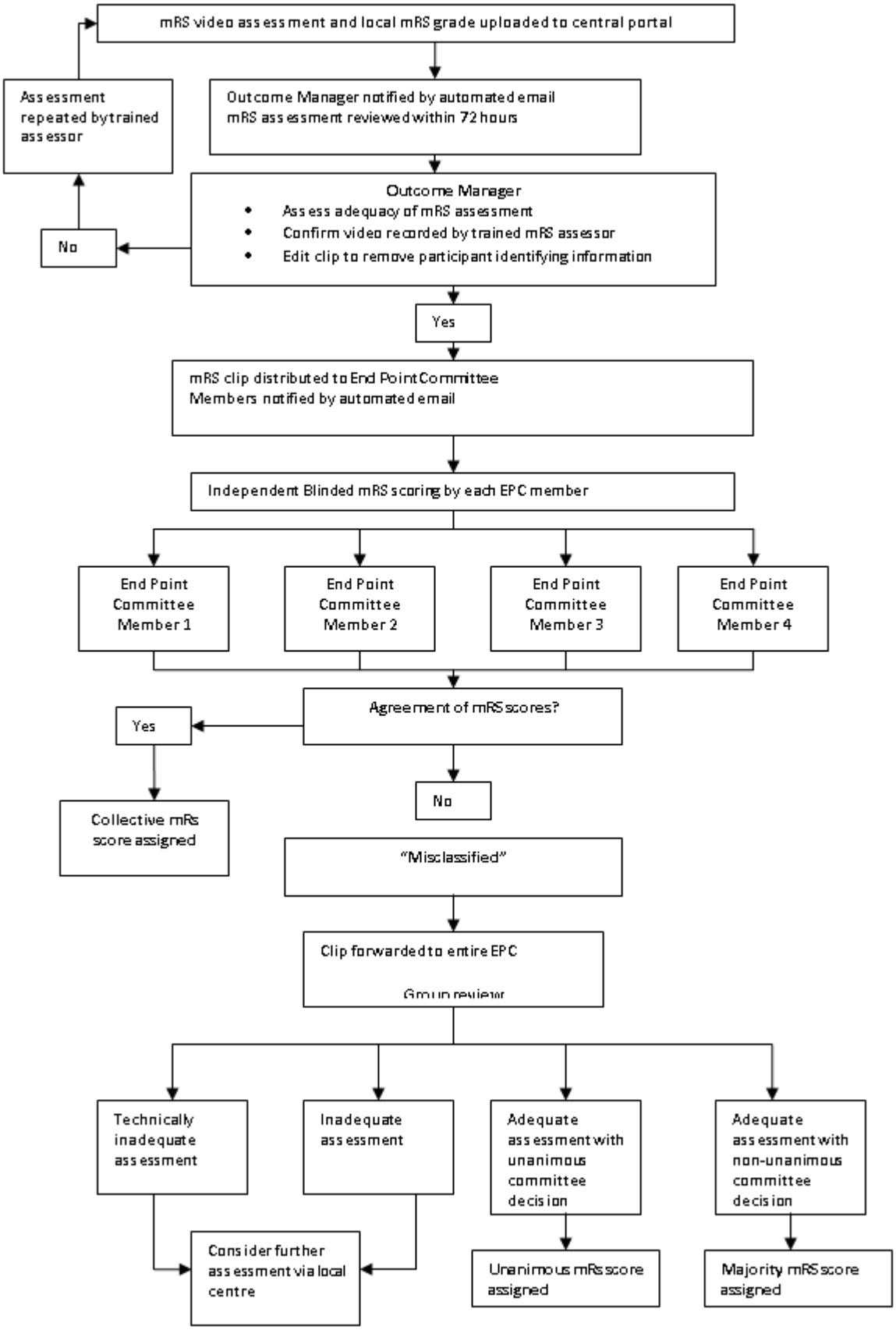


Figure 11 mRS Video review process

3.5. Technical Specifications

3.5.1. Video Equipment

The mRS assessment was recorded using a high definition video camera. Two camera systems were used during the study.

Initially a Canon© high definition camera was used (Canon HF10©, Canon HF100©, Canon HF200©). This system required the use of an external omni-directional condenser boundary microphone (ATR97, Audio-technica©. Frequency response 50-1500Hz). A portable desk tripod was used to mount the video camera (Hama Photo Traveller Compact Tripod©). The camera files were transferred to PC using an external USB cable. The total cost of this equipment was £700. (See Figure 12)

Figure 12 Canon© Camera Equipment

Later in the study we began to trial the use of a FLIP© Mino camera. This smaller and more portable model included an in built microphone and USB connection allowing greater ease of file transfer. The total cost of this equipment was £130. (See figure 13)

Figure 13 Flip© Camera Equipment

Twelve centres used the original Canon© equipment and two centres trialled the FLIP© equipment.

3.5.2. Video Conversion Software

In planning the study, we anticipated from prior pilot work that video clips would be 5-10 minutes in length and around 7-15 MB in size in MPEG format. More modern camera equipment was purchased for the CARS study which recorded video footage with higher resolution in MTS format. These MTS format video clips ranged from 150-250 MB for a similar length of mRS assessment. Allowing for NHS and University network speeds these files were too large to be uploaded over the internet in an acceptable time frame and therefore a video conversion step was included with the initial Canon camera. We used AVS Video Converter© Software, Online Media Technologies Ltd. UK. This allowed conversion of MTS files to WMV files which were 5-15 MB in size and were easily viewed on any Windows personal computer (PC) (Microsoft©, USA) with Windows Media Video© software.

Later in the study, with the addition of the FLIP camera technology the file format was updated. We began to receive AVI files which were of greatly increased size (100-200MB). In response to this we undertook re-programming of the web portal with the Robertson Centre for Biostatistics and all files were able to be transferred in a non-converted format. This allowed direct upload of the camera files to the co-ordinating centre with much greater ease and without the requirement for conversion software.

3.6. The CARS web portal

The CARS web portal was developed with technical assistance from the Robertson Centre for Biostatistics (RCB), University of Glasgow. The functionality of the web portal was planned and designed by the study team and then constructed by the web design team at the RCB. The RCB were responsible for the administration of the many functions of the portal under instruction from the study team. All study documentation and links to training materials were accessible to investigators. After enrolment, the portal was used to generate an electronic case report form (eCRF) for each study participant. This facilitated remote entry of participant details, NIHSS assessment scores, standard mRS assessment scores and upload of mRS video assessments for review by the co-ordinating centre. Secure access to the portal was granted through the use of username and password, provided to authorised investigators by the RCB, University of Glasgow (Figure 14). On first use, users were prompted to change their password. Smart passwords were required.

3.6.1. Investigator access to CARS web portal

Different functions of the web portal were available to study investigators allocated by their role within the trial. This was designed to ensure that data were kept secure, accessible by the minimal number of investigators and to allow full blinding of outcomes assessment. The various components of the portal and degree of access available to each study role are detailed in figure 15.

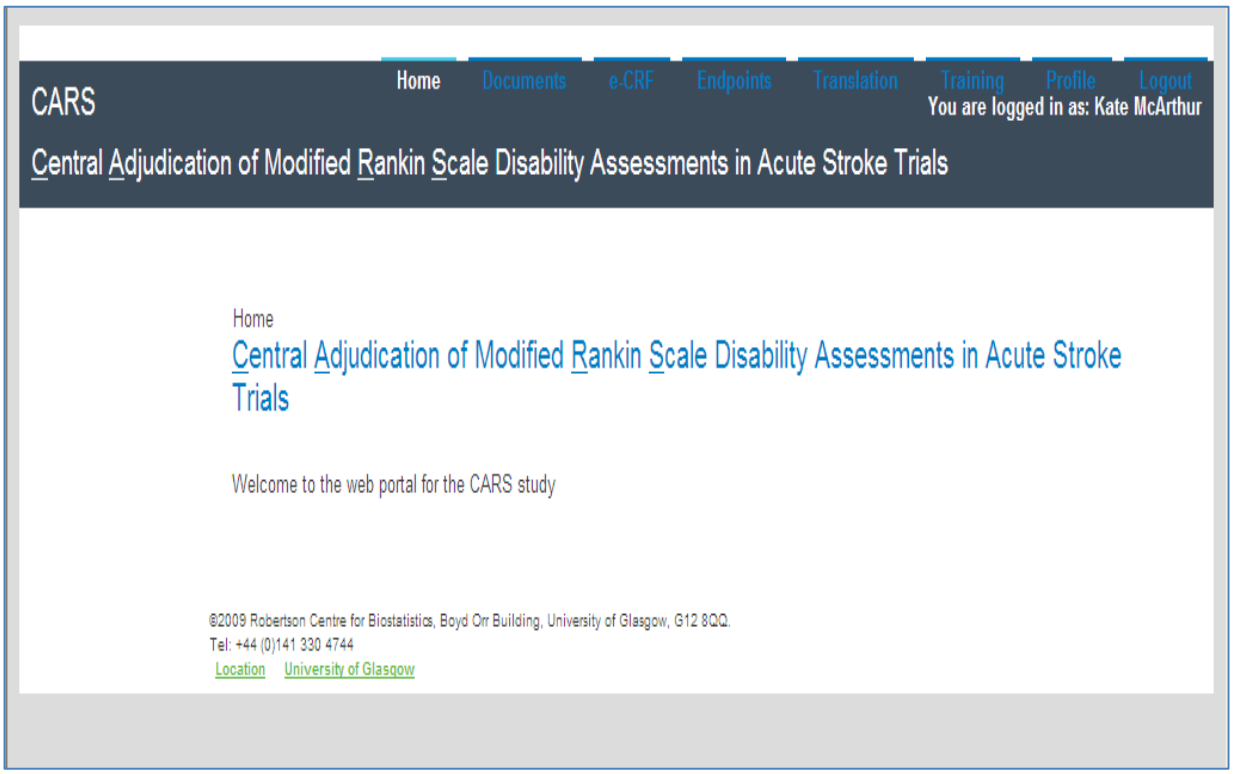


Figure 14 - The CARS Web Portal

Portal function	Access available to:
Study documents Study Protocol Participant and Relative information sheets Consent documents Paper case report forms	• All users
Electronic Case Report forms (eCRF) Electronic entry of participant data	• Outcomes manager - able to view all sites • Local investigators - able to view local site
Endpoints Assessment of mRS video clips Diary to prospectively record periods when unavailable	• Outcomes manager - able to see all clips • Endpoint committee members - able to see clips for individual review and group review if misclassified
Training Links to internet based training and certification modules in NIHSS and mRS	• All users
Profile Link to change password	• All users

Figure 15 CARS web portal. Functions and User Access.

3.6.2. Electronic Case Report form (eCRF)

Standard paper case report forms were used for each study participant, stored locally and completed at the time of the study visit by the investigator. These data were also entered into the eCRF, available to all local investigators through the CARS web portal. This allowed automatic entry of source data to our database, held by the RCB. As far as possible all data were entered using drop down menus or tick boxes to minimise possible responses, limit human error and aid data analysis. Where free text was more practical or necessary this was available. Any change / correction made to the data after initial entry was archived and the investigator making the change was required to state a reason for this in line with standard data monitoring practice.

3.6.3. Endpoint Assessment

Endpoint assessment was performed by a committee of academics with a clinical and academic interest in stroke. All had extensive experience in administration of the mRS assessment and many had been involved in prior mRS reliability research.

One member of the endpoint committee had a dual role as outcome manager. After initial review of each clip for technical adequacy this member distributed the clips to a minimum of four endpoint committee members. This was done in an unsystematic manner. Where endpoint committee members were unavailable to review clips for any period of time they could note this in the web portal to ensure that they did not appear on the list of potential reviewers at the time of distribution. The aim of this was to ensure timely review of all clips, our target was to ensure all clips were adjudicated within 7 days of upload to allow for repeat assessment or further information where clips were contentious.

After allocation of an mRS clip, endpoint committee members were alerted by automated email. They were prompted to log in to the web portal to view and score the clip. Once all four reviews were in place the portal was programmed to recognise any discrepancy in scoring among the four endpoint committee opinions. If all scores agreed then the clip and corresponding scores were archived securely in the database. If there was any disagreement,

the clip was deemed “misclassified” and this prompted a further automated email to all seven members of the endpoint committee to alert them all to log in and review the clip.

Regular endpoint committee meetings were arranged by the outcomes manager as a forum to discuss the misclassified clips and reach a consensus score where possible. At the endpoint committee meetings each member attended with a note of their scores and comments for discussion. Access to the CARS web portal was available for entry of scores and comments. The misclassified clips were available via the portal for review if the committee members wished to review the clip collectively in order to facilitate discussion and scoring.

As the study progressed we expanded the facility to store endpoint committee scores in order to facilitate endpoint committee discussion. Initially the portal had the facility to accept only the four original committee scores and a consensus score. We expanded this to allow all seven members to enter a score at the stage of full committee review and also provided a comments box for each committee member to use if there were specific details about the clip that were contentious or guided their scoring decision. In the later stages of the study all of these details were available through the portal for endpoint committee meetings. These alterations were devised to help facilitate remote discussion in cases where endpoint committee members were not available to attend.

Where a score was allocated after committee discussion it was noted if this score was allocated with unanimous or non-unanimous committee agreement in order to highlight the most controversial clips. Where no score could be allocated due to inadequate technical quality or inadequate interview consideration was given to request further information from the local site or a repeat assessment. There was a link next to each assessment in the web portal to email the investigator directly in the case of any queries.

3.7. Conclusions

The CARS study was designed to emulate a typical phase III RCT in acute stroke, similar to recent large studies investigating reperfusion or neuroprotectant agents. We have designed a facility to incorporate central adjudication of mRS assessments within an acute stroke trial

and a process for providing adjudicated endpoint blinded to the original / local mRS assessment. We aim to examine the benefits of such an outcome strategy. The results of the study programme are presented in the following chapters. We aimed to assess the benefits of central adjudication of mRS outcomes in terms of feasibility of collecting this data (Chapter 4) together with an assessment of the reliability (Chapter 5) and validity (Chapter 6) of the outcomes recorded.

Chapter 4

Feasibility: is a central adjudication model feasible?

4.1. Introduction

As a novel method of collecting modified Rankin outcome data, we aimed to establish whether central adjudication of locally recorded mRS assessments can be performed in a multicentre trial setting. In this chapter we will discuss the practical details of conducting the CARS study. It will detail the recruitment, data collection and central adjudication processes to examine the feasibility and acceptability of using this model in future stroke trials.

4.2. Results

4.2.1. Study Sample

We set up a "virtual" acute stroke trial across UK hospitals. Fourteen study centres were included in the CARS study; six in Scotland (ethical approval granted by Scotland A Research Ethics Committee on 12th November 2008) and eight in England / Wales (ethical approval granted by Essex 2 research ethics committee on 28th January 2009).

4.2.1.1. Recruitment

Three hundred and seventy three participants were recruited to the study. Recruitment began at the co-ordinating centre and extended to all fourteen sites as each centre received training and achieved local research and development approval. Recruitment commenced on 17th December 2008 and completed on 22nd September 2010. Table 8 details the date each

centre commenced the study, the majority joined in a six month period between May and October 2009. The final day 90 follow up visits were completed in January 2011. The number of participants recruited by month as the trial progressed is detailed in Figure 16. Total recruitment at each site is detailed in Figure 17.

Of 373 participants recruited, 3 were excluded from analysis. One was later found to have a diagnosis other than stroke (demyelinating disease) and two withdrew their consent and asked to have their data removed from analysis.

Table 8 Date for first participant recruited at each site

Site	Date first participant recruited
SCOTLAND	
Aberdeen Royal Infirmary	15.07.09
Glasgow Royal Infirmary	10.02.09
Ninewells Hospital, Dundee	19.01.10
Southern General Hospital, Glasgow	22.12.09
Stobhill Hospital, Glasgow	08.05.09
Western Infirmary, Glasgow	17.12.08
ENGLAND AND WALES	
Countess of Chester Hospital, Chester	28.06.09
Cumberland Infirmary, Carlisle	19.03.10
Harrogate District General Hospital	10.09.09
Leeds General Infirmary	21.08.09
Royal Devon and Exeter Hospital	28.07.09
Royal Glamorgan Hosptal	16.10.09
Wakefield (Pinderfields) Hospital	01.10.09
York Hospital	19.08.09

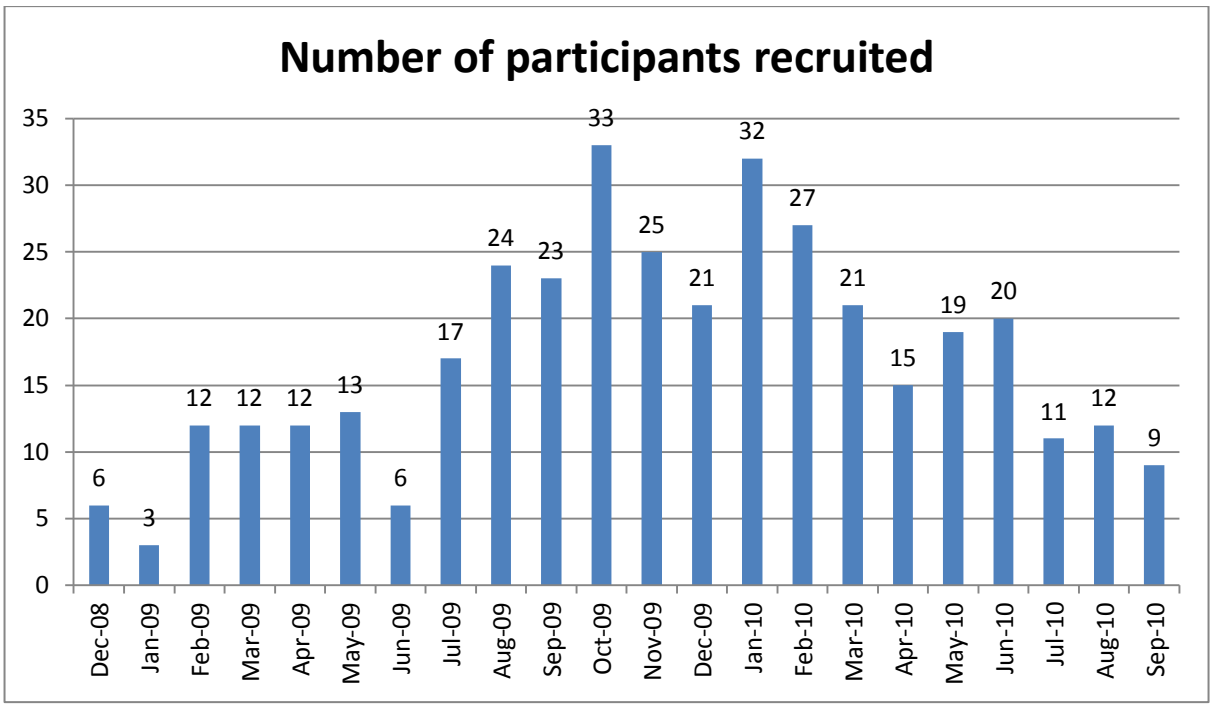


Figure 16 Number of participants recruited by month

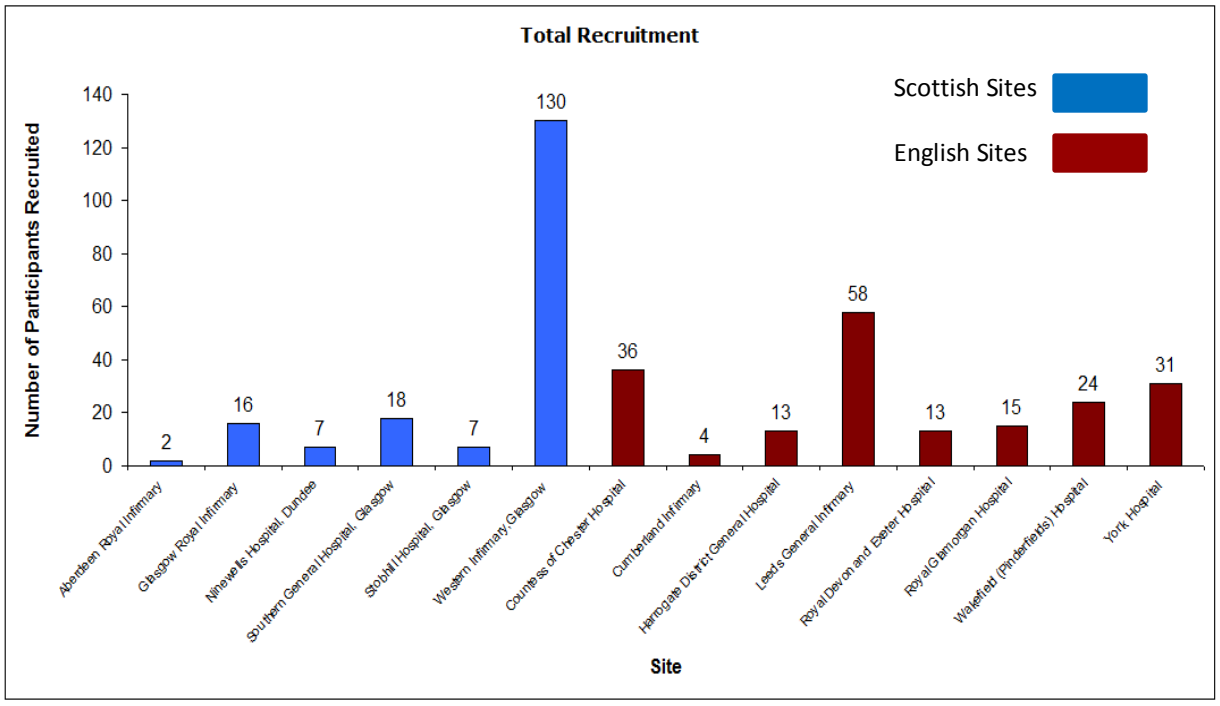


Figure 17 Total Recruitment by site

4.2.1.2. Consent

All participants provided written informed consent or where a participant was unable to do so, proxy consent from their nearest relative was used. 46 participants were included with proxy consent. Participants included with proxy consent were slightly older than those who gave their own consent (mean 70.5 years versus 67.4 years). Stroke was more severe in the group who were included with proxy consent (mean baseline NIHSS (bNIHSS) 4.8 versus 11.8), indicating communication difficulties or other incapacitating symptoms at the time of the index event.

Of the 46 included with proxy consent, 30 completed the study. Seven were withdrawn due to a serious adverse event. In this group we see the most severe strokes (mean bNIHSS 18.3). The nine participants who chose to withdraw from the study after initial proxy consent were younger than those completing (66.1 years versus 70.6 years). One of the participants included with proxy consent actively withdrew consent for use of data during the study. Age and stroke severity of the participants included with proxy consent are detailed in Table 9.

Table 9 Age and Stroke Severity of participants included with own consent / proxy consent

	N	Range	Mean (SD)	Median (IQR)
Age (years)				
All	373	22 – 99	67.8 (13.0)	69 (60-77)
Own consent	327	22 – 99	67.4 (13.2)	69 (59-77)
Proxy Consent (All)	46	37 – 88	70.5 (10.6)	71 (67-76)
Proxy Consent (Completed)	30	45 – 88	70.6 (9.6)	72 (67-76)
Proxy Consent (Withdrew)	9	37 – 86	66.1 (15.2)	68 (66-71)
Proxy Consent (SAE)	7	67 – 83	75.6 (6.4)	77 (72-80)
Baseline NIHSS				
All	373	0 - 23	5.6 (4.8)	4 (2-7)
Own consent	327	0 – 21	4.8 (3.8)	3 (2-6)
Proxy Consent (All)	46	2 – 23	11.8 (6.4)	11 (7-17)
Proxy Consent (Completed)	30	2 – 22	10.8 (6.2)	10 (6-17)
Proxy Consent (Withdrew)	9	3 – 22	10.3 (6.0)	9 (6-13)
Proxy Consent (SAE)	7	15 – 23	18.3 (3.1)	18 (16-20)

4.2.2. Demographics and Baseline Characteristics

61.4% (227/370) of participants were male with a mean (SD) age of 67.8 (13.2) years. The participants ranged from 22 years to 99 years old.

Most participants had an excellent functional status prior to involvement in the study. The majority had a premorbid mRS of 0 (75.5%, 269/370). There were ten participant recruited with a pre-morbid mRS of 3 and two participants recruited with a pre-morbid mRS of 4 in violation of the protocol.

Stroke severity in the study population was mild. Baseline NIHSS ranged from 1 to 23 with a mean (SD) of 5.6 (4.8). Median baseline NIHSS was 4, IQR 2-7. The distribution of stroke severity by the Oxford Classification was as follows: Total Anterior Circulation Stroke (TACS) 12.7%, Partial Anterior Circulation Stroke (PACS) 41.6%, Posterior Circulation Stroke (POCS) 14.9% and Lacunar Stroke (LACS) 30.8%. Stroke side was equally distributed, left hemisphere affected in 50.2%, right hemisphere affected in 49% and both sides affected in 0.8%.

Stroke risk factors were widespread. 14.9% (55/370) had a history of previous stroke. 20.3% had a pre-existing history of atrial fibrillation or were found to be in atrial fibrillation at the time of presentation. A history of smoking (65.1%, 240/370), hypertension (55.6%, 205/370), hyperlipidaemia (39.2%, 145/370) ischaemic heart disease (21.6%, 80/370) or diabetes (14.1%, 52/370) was highly prevalent in the study population. 21.6% (80/370) of participants had a family history of stroke.

The baseline blood pressure ranged from 96/45 mmHg to 232/140 mmHg with an average reading of 152/81mmHg. There were 29 missing values for admission blood glucose. The mean (SD) glucose level was 6.7 (1.4) mmol/l, ranging from 3.1 – 20.7. Demographic details are summarised in Table 10.

Table 10 Baseline Demographic Characteristics of CARS study participants

Baseline Demographics of CARS study participants (n=370)		
Age (years)	Mean (SD)	67.8 (13.2)
	Median (IQR)	69 (60-77)
	Range	22-99
Sex	Male	227 (61.4%)
	Female	143 (38.6%)
Oxford Classification	TACS	47 (12.7%)
	PACS	154 (41.6%)
	POCS	55 (14.9%)
	LACS	114 (30.8%)
Side of Stroke	Right	187 (50.2%)
	Left	180 (49%)
	Both	3 (0.8%)
Pre Morbid mRS	Mean (SD)	0.9 (1.25)
	Median (IQR)	0 (0 – 0.25)
	Range	0 – 4
Baseline NIHSS	Mean (SD)	6.2 (4.8)
	Median (IQR)	4 (2 – 7)
	Range	1-23
Blood Pressure (mmHg)	Mean (SD)	152/81 (33/18)
	Median (IQR)	150/80 (132/70-170/92)
	Range	96/45 – 232/140
Blood Glucose (mmol/l)	Mean (SD)	6.7 (1.4)
	Median (IQR)	6.0 (5.2-7.1)
	Range	3.1 – 20.7
Risk Factors (n= (%))	Smoking Current	124 (33.5%)
	Former	117 (31.6%)
	Never	129 (34.9%)
	Alcohol Excess	26 (7.0%)
	Hypertension	207 (55.9%)
	Hyperlipidaemia	145 (39.2%)
	Previous Stroke	55 (14.9%)
	Peripheral Vascular Disease	20 (5.4%)
	Atrial Fibrillation	75 (20.3%)
	Family History	80 (21.6%)
	Diabetes	52 (14.1%)
	Ischaemic Heart Disease	80 (21.6%)

4.2.3. Trial Termination

A considerable proportion of the study population did not complete all follow up visits. Of the 370 participants included in analysis, 267 completed 90 days of follow up. 65 participants did not return for 30 day follow up. In 10 cases this was due to a serious adverse event (SAE). 38 further participants withdrew after 30 day follow up, six due to an SAE. Figure 18 and Table 11 describe the flow of participants through the study and detail the reasons for trial termination.

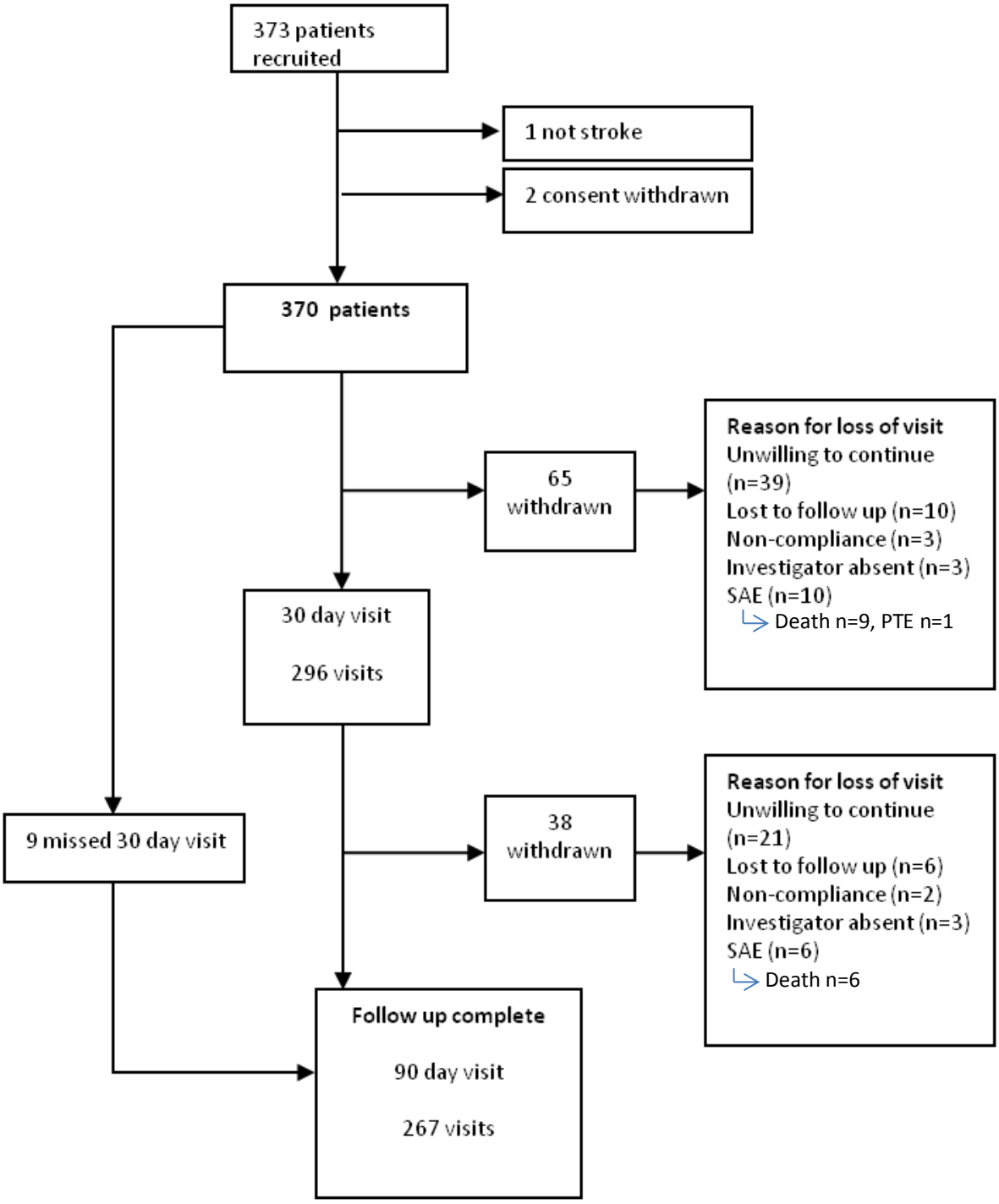


Figure 18 - Flow diagram of participant follow up

Table 11 Trial Termination Details for participants who attended at day 30, day 90 or did not attend either visit

Trial Termination	Day 30 visits (n= 296)	Day 90 visits (n=267)	No visits (n=68)	Total
Study Completion	258	267	0	267
Subject unwilling to continue	21	0	39	60
Lost to follow up	6	0	10	16
SAE	6	0	10	16
Non Compliance	2	0	3	5
Subject violated protocol	0	0	1	1
Investigator terminated participation	0	0	0	0
Subject withdrawn consent for use of data	0	0	2	2
Investigator absence	3	0	3	6
Total	296	267	68	373

Six participants were withdrawn from the study due to investigator absence. A single investigator at one site was unavailable for several months due to ill health leading to a number of follow up visits being omitted.

Stroke severity and baseline demographic characteristics were similar in participants who completed versus participant chosen withdrawals. Tables 12 and 13.

Table 12 Stroke severity of participants at baseline, 30 days and 90 days in each follow up group

		Follow up Complete (n=267)	No visits (n=52)	30 day visit only (n=29)
Baseline mRS	Mean (SD)	0.37 (0.76)	0.5 (0.92)	0.79 (1.11)
	Median (IQR)	0 (0-0)	0 (0-1)	0 (0-1)
	Range	0-4	0-3	0-4
Baseline NIHSS	Mean (SD)	5.39(4.53)	4.63 (3.36)	5.83 (5.43)
	Median (IQR)	4 (2-7)	4 (2-6)	4 (2-6)
	Range	0-22	1-17	0-22
30 day mRS	Mean (SD)	2.24 (1.21)		2.21 (1.26)
	Median (IQR)	2 (1-3)		2 (1-3)
	Range	0 – 5		0-4
30 day NIHSS	Mean (SD)	2.83 (3.64)		1.5 (1-4.25)
	Median (IQR)	2 (1-3)		2.50 (2.41)
	Range	0 – 22		0-8
90 day mRS	Mean (SD)	1.91 (1.20)		
	Median (IQR)	2 (1-3)		
	Range	0 – 5		
90 day NIHSS	Mean (SD)	2.14 (2.83)		
	Median (IQR)	1 (0–3)		
	Range	0 – 19		

Table 13 Baseline Demographic Characteristics of participants in each follow up group

		Both Visits (n=267)	No visits (n=52)	30 day visit only (n=29)
Age (years)	Mean (SD)	67.7 (12.7)	69 (12.4)	65 (15)
	Median (IQR)	69 (60-77)	68 (60.8-70.5)	63 (56-79)
	Range	22-99	43-92	37-90
Sex	Male	163 (61.0%)	28 (53.8%)	21 (72.4%)
	Female	104 (39.0%)	24 (46.2%)	8 (27.6%)
Oxford Classification	TACS	31 (11.6%)	7 (13.5%)	2 (6.9%)
	PACS	106 (39.7%)	21 (40.4%)	16 (55.2%)
	POCS	45 (16.9%)	8 (15.4%)	3 (10.3%)
	LACS	86 (32.2%)	16 (30.8%)	8 (27.6%)
Side of Stroke	Right	134 (50.2%)	24 (46.2%)	22 (75.9%)
	Left	131 (49.1%)	28 (53.8%)	6 (20.7%)
	Both	2 (0.7%)	0 (0%)	1 (3.4%)
Blood Pressure (mmHg)	Mean (SD)	153/81 (27/16)	150/83 (27/17)	148/81 (28/17)
	Median (IQR)	158/80 (133/70 - 170-90)	143/80 (130/68 – 167-95)	143/80 (129/71 – 161/92)
	Range	98/45 – 232/140	96/47 – 226/118	97/49 – 230/119
Blood Glucose (mmol/l)	Mean (SD)	6.7 (2.4)	6.6 (2.9)	6.4 (2.5)
	Median (IQR)	6.7 (5.2-7.1)	5.6 (5.2-6.6)	5.8 (5.3 – 6.5)
	Range	3.1 – 19.9	3.8 – 20.7	4 – 15.3
Risk Factors	Smoking Current	86 (32.2%)	22 (42.3%)	15 (51.7%)
	Former	81 (30.3%	13 (25%)	16 (55.2%)
	Never	100 (37.5%)	17 (32.7%)	8 (27.6%)
	Alcohol Excess	15 (5.6%)	4 (7.7%)	5 (17.2%)
	Hypertension	147 (55.1%)	31 (59.6%)	23 (79.3%)
	Hyperlipidaemia	108 (40.4%)	21 (40.4%)	12 (41.4%)
	Previous Stroke	43 (16.1%)	9 (17.3%)	3 (10.3%)
	Peripheral Vascular Disease	13 (4.9%)	5 (9.7%)	3 (10.3%)
	Atrial Fibrillation	54 (20.2%)	12 (23.1%)	7 (24.1%)
	Family History	60 (22.5%)	11 (21.2%)	6 (20.7%)
	Diabetes	45 (16.9%)	4 (7.7%)	3 (10.3%)
	Ischaemic Heart Disease	58 (21.8%)	12 (23.1%)	7 (24.1%)

4.2.3.1. Serious Adverse Events (SAEs)

As stroke severity was mild in the study population SAEs were infrequent. There were fifteen deaths. Sixteen participants were withdrawn due to an SAE. Thirteen patients (3.5%) had a recurrent cerebrovascular event during the study period. The recorded SAEs are detailed in Table 14.

Table 14 Serious Adverse Events

Serious Adverse Events	
Death	15
DVT/PTE	5
Elective Procedure	2
Endarterectomy	5
Fall	6
Unrelated Illness requiring admission	18
MI/CCF/AF	7
Pneumonia	9
TIA/Stroke/New Neurology	13
Pt's withdrawn due to SAE	16
Total	**80**

4.2.4. Imaging

All patients had brain imaging. 96.8% (358/370) had initial CT scanning and 13.2% (49/370) had MRI. The majority of participants had ischaemic stroke (94.9%, 351/370). 27.3% (101/370) of participants had normal imaging on CT and in 10 cases this was confirmed with a normal MRI scan.

Carotid Doppler studies were performed in 73.2% of participants (271 / 370). In some cases an alternative modality of carotid imaging was performed (CT or MR angiography) and in the remainder no carotid imaging was done. We did not collect data regarding carotid imaging other than Doppler ultrasound. Severe carotid stenosis was found in 26 participants, in 14 cases these were symptomatic stenoses. Five carotid endarterectomies were recorded during the study period. Details of imaging are described in Table 15.

Table 15 Imaging – Frequency of CT, MRI and Carotid Doppler Ultrasound studies and Results

CT Brain (n=358 [96.8%])	Normal	101 (28.2%)
	Primary Intracerebral Haemorrhage	18 (5.0%)
	Subcortical Infarction	95 (26.5%)
	Cortical Infarction	95 (26.5%)
	Posterior Circulation Infarction	48 (13.4%)
	Other	32 (8.9%)
MRI Brain (n=49 [13.2%])	Normal	10 (20.4%)
	Primary Intracerebral Haemorrhage	1 (2.0%)
	Subcortical Infarction	11 (22.4%)
	Cortical Infarction	14 (28.6%)
	Posterior Circulation Infarction	11 (22.4%)
	Other	11 (22.4%)
Carotid Doppler USS Right (n=271 [73.2%])	Normal	196 (72.3%)
	Mild Stenosis	33 (12.2%)
	Moderate Stenosis	18 (6.6%)
	Severe Stenosis	10 (3.7%) (Symptomatic Side n= 3)
	Occlusion	14 (5.2%)
Left	Normal	200 (73.8%)
	Mild Stenosis	35 (12.9%)
	Moderate Stenosis	14 (5.2%)
	Severe Stenosis	16 (5.9%) (Symptomatic Side n=11)
[Grading of severity of carotid stenosis was at the discretion of each investigating team based upon local protocols]	Occlusion	6 (2.2%)

4.2.5. Medications

Consistent with the high prevalence of stroke risk factors in the study population, a substantial proportion of participants were prescribed primary preventative medications at baseline. At the time of recruitment 41.4% (153/370), 55.4% (205/370) and 44.3% (164/370) of the study population were taking aspirin, antihypertensive treatment and statin treatment respectively. At follow up there were increased rates of prescription of all secondary preventative treatment. The rate of antiplatelet prescription with aspirin and dypiridamole dual therapy was greatest at 30 days. At 90 days the proportion of participants taking monotherapy with clopidogrel was greater. The rate of warfarin prescription increased from 7.3% (27/370) at baseline to 27% (72/267) at 90 days. We did not collect data regarding rates of new atrial fibrillation at follow up. The rate of antihypertensive prescription and statin

prescription increased at each follow up visit. A summary of secondary preventative medication is detailed in table 16.

Table 16 Frequency of secondary preventative medication prescription at each study visit

	Baseline (n=370)	30 days (n=289)	90 days (n=267)
Aspirin	153 (41.4%)	188 (65.1%)	155 (58.1%)
Dypiridamole	20 (5.4%)	108 (37.4%)	89 (33.3%)
Clopidogrel	19 (5.1%)	45 (15.6%)	46 (17.2%)
Warfarin	27 (7.3%)	63 (21.8%)	72 (27.0%)
LMWH	1 (0.3%)	8 (2.8%)	2(0.8%)
Thrombin Inhibitor	0 (0%)	1 (0.3%)	1 (0.4%)
Antihypertensives	205 (55.4%)	194 (67.1%)	190 (71.2%)
Statins	164 (44.3%)	241 (83.4%)	229 (85.8%)

4.2.6. Home Time

Home time was recorded at 30 and 90 days. Home times were clustered at high and low levels, a reflection of the mild stroke population being discharged very early with a small proportion having a more prolonged length of stay in hospital for rehabilitation after more severe stroke.

The average home time at 30 and 90 days was 16.5 and 65.8 days respectively. In each case there was a full range of possible home time (0 -30 days or 0 – 90 days) and the median (IQR) was 22 (0 – 26) and 82 (59 – 86). Cumulative Frequency distribution of 90 day home time is displayed in Figure 19.

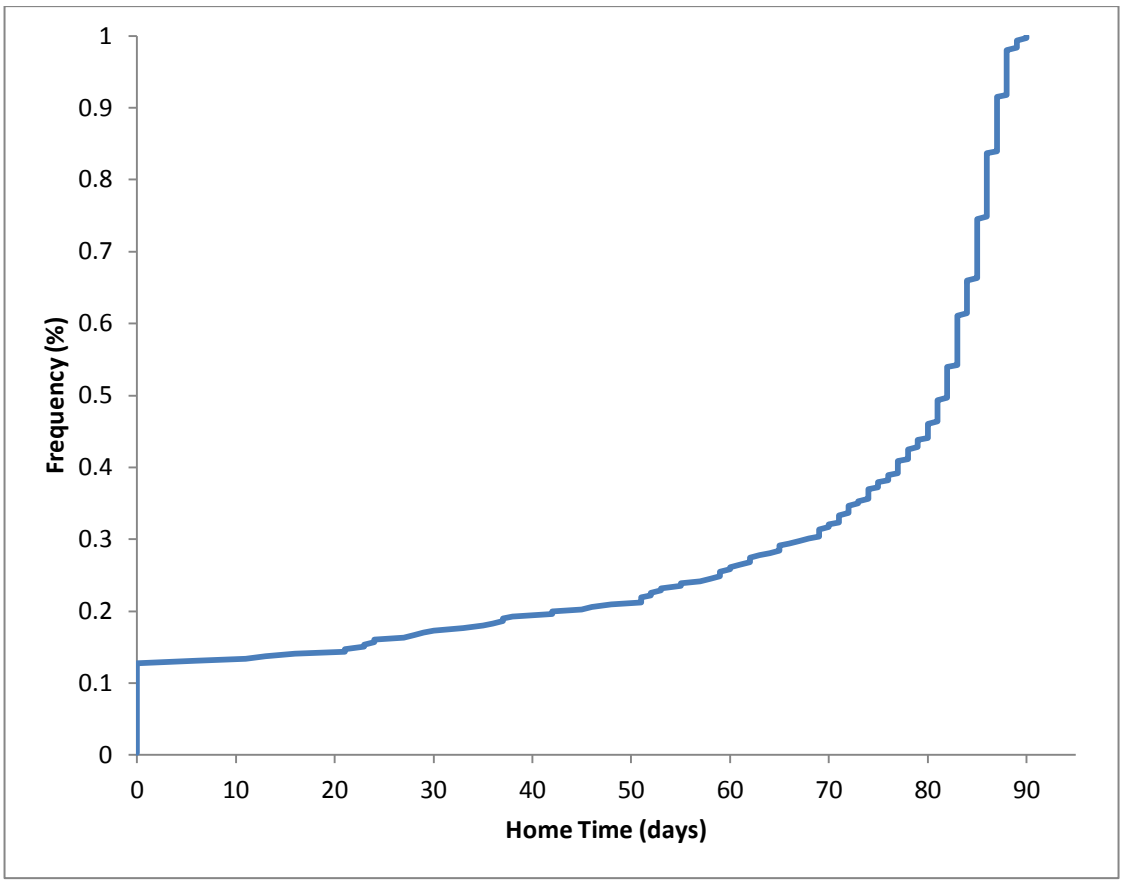


Figure 19 Cumulative Frequency Distribution of Home Time (90 days)

4.2.7. mRS assessment Videos

Five hundred and sixty three mRS video assessments were uploaded for review. Twelve centres used the original Canon HF100© camera and two centres used the Flip Mino© camera. Interview duration ranged from under one minute to 24 minutes; mean (SD) 5mins 32 seconds (3mins 20seconds), median (IQR) 4mins 47seconds (3mins 9 seconds – 7mins 6 seconds). Figure 20.

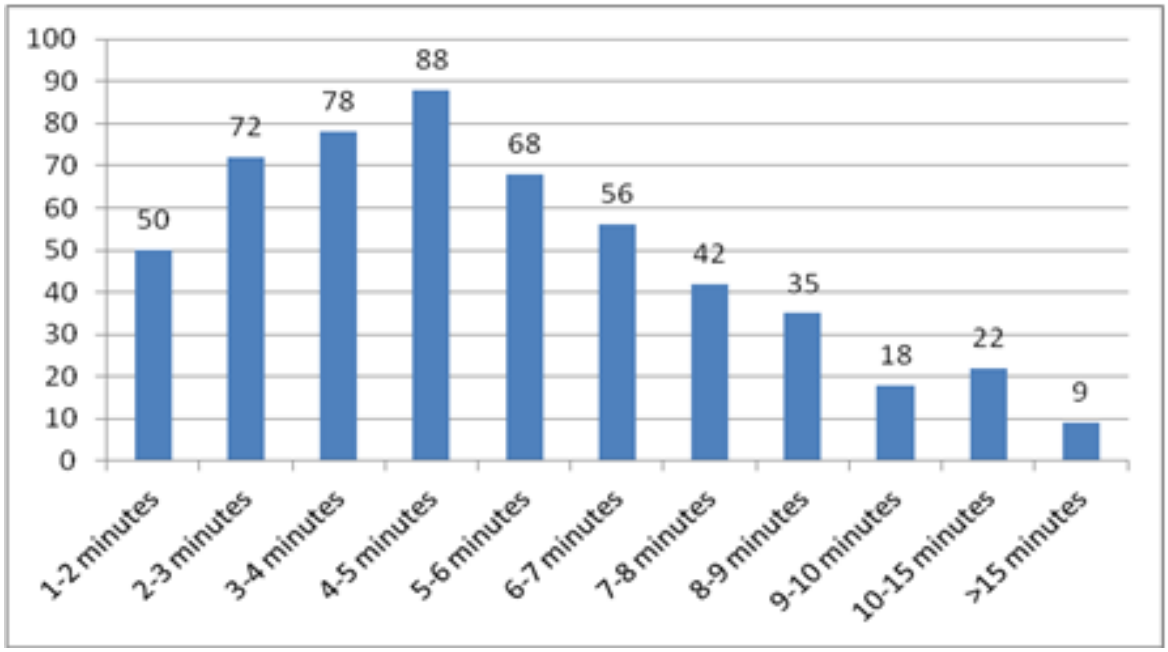


Figure 20 - mRS video length (minutes)

File sizes ranged from 1920KB to 116095KB (113MB); mean (SD) 18230KB (16453KB). (Table 17) Smaller files were received where the original Canon camera was used; these files were converted to a compressed file type prior to upload. Five hundred and thirty eight clips were scored by the adjudicating committee. The reasons for missing adjudicated scores will be discussed in section 4.2.8.1.

Table 17 File size and duration of video mRS assessments

File Size (KB)	Mean (SD)	18230 (16453)
	Median (IQR)	13152 (8510 – 20635)
	Range	1920 – 116095
Clip Duration (mins:secs)	Mean (SD)	05:29 (0:14)
	Median (IQR)	04:45 (03:09 – 07:03)
	Range	00:52 – 23:51

Editing for anonymity was required in thirty nine (7.2%) mRS assessments due to mention of participants forename, surname or date of birth. A proxy was used in the assessment in 106 (19.3%) assessments. In the majority of these clips both the participant and a proxy were used to provide the interview for clarity and corroboration of details. The full interview was provided by a proxy in only 22 (4.1%) cases. In 14 of these the mRS interview was conducted

with a member of the clinical team involved in the participant's daily care (nurse / physiotherapist / occupational therapist / doctor) and in 8 the mRS interview was conducted with the participant's relative or informal carer. There was language disorder noted in 145 (20.7%) assessment videos. The majority of assessments were filmed in an outpatient clinic setting (389 clips, 72.3%). Eighty six assessments (16.0%) were filmed in the participants home and sixty three (11.7%) were filmed in a ward setting prior to discharge. To assess the level of pre-morbid disability in line with the pre-morbid mRS score it was noted how many clips involved a discussion of prior disability. There was no mention of previous functional limitations in 415 clips (77.1%). In 14 clips there was mention of previous stroke (43 participants who attended both follow up visits had a documented history of prior stroke). In 109 clips there was mention of other co-morbidity affecting pre-morbid function (arthritis, respiratory or other cardiovascular illness). Details of video clips scored are shown in Table 18.

Table 18 Details of video mRS assessments

Proxy	No proxy Participant provides full interview	432 (80.3%)
	Proxy provides full interview Clinical staff	14 (2.6%)
	Proxy provides full interview Relative	8 (1.5%)
	Participant and proxy Both provide interview together	84 (15.6%)
Language Disorder	Yes	145 (27.0%)
	No	393 (73.0%)
Discussion of prior disability	None	415 (77.1%)
	Yes – previous stroke	14 (2.6%)
	Yes – other co-morbidity	109 (20.3%)
Location	Hospital Study Visit / Clinic	389 (72.3%)
	Inpatient / Care facility	63 (11.7%)
	Patient's home	86 (16.0%)

4.2.8. Adjudication of mRS videos by Endpoint Committee

96% (538/563) of study visits resulted in an adjudicated mRS score.

The majority of mRS assessments were performed using a traditional, unstructured approach. 45 structured mRS scores were recorded at 30 days and 28 structured mRS scores were recorded at 90 days. Structured mRS scores were limited to six sites, in four of these all study visits recorded an mRS score at 30 and 90 days, in two sites there were structured mRS scores recorded in some study visits only. One site recorded structured mRS at 30 days only in 8.2% (4/49) of participants. One site recorded structured mRS at 30 and 90 days in 30.4% (7/23) and 26.1% (6/23) of participants at each time point. An adjudicated mRS score was available in 88.9% (40/45) and 92.9% (26/28) of mRS assessments where a structured mRS was recorded at 30 and 90 days respectively.

4.2.8.1. Missing adjudicated scores

Technical failures were responsible for the majority of missing adjudicated scores. The clips where technical failure precluded scoring were recorded early in the study; median (IQR) 159 (111-221) days into study [Total study days Median (IQR) 510 (458-558)].

Poor audio was encountered in nineteen clips (3.4%), rendering scoring impractical in seven (1.2%). A repeat assessment was requested by the adjudicating centre in 15 cases (2.7%). In six cases a duplicate clip was uploaded in error. In two cases an accessory file (.exe) was uploaded instead of the video file. At one centre there were a number of video clips stored on the camera which were not uploaded to the co-ordinating centre immediately. These assessments were lost when the camera malfunctioned. In two cases the camera or microphone battery died during the assessment interview, resulting in an incomplete assessment. The reasons for missing adjudicated scores are detailed in Table 19.

Table 19 Reason mRS video unable to be scored by endpoint committee

Reason (n)	Day 30 (Total n = 296))	Day 90 (Total n= 267)
Poor Audio	5	2
Duplicate Clip uploaded in error	1	5
Camera Malfunction – data lost	5	-
Incorrect (assessory) File type uploaded in error	2	-
Incomplete Assessment (battery failure)	1	1
Committee Unable to reach consensus	2	1
Total	16	9

4.2.8.2. Time to adjudicated mRS score

The time taken to complete the adjudication process varied substantially during the study. We were unable to meet our target of timely mRS adjudication within 7 days. Each committee member (C1-C7) entered scores at their convenience and in many cases committee members waited for several clips to be ready for scoring before assigning mRS via the web portal. Committee members C1, C2 and C3 chose to score clips as they were allocated. Committee members C6 and C7 saved large numbers for scoring in batches, resulting in a considerable delay in scoring some clips. The median number of days to scoring for each committee member is displayed in Figure 21. Once misclassified, the adjudication committee assigned a final score within a median (IQR) of 21 (13-37) days.

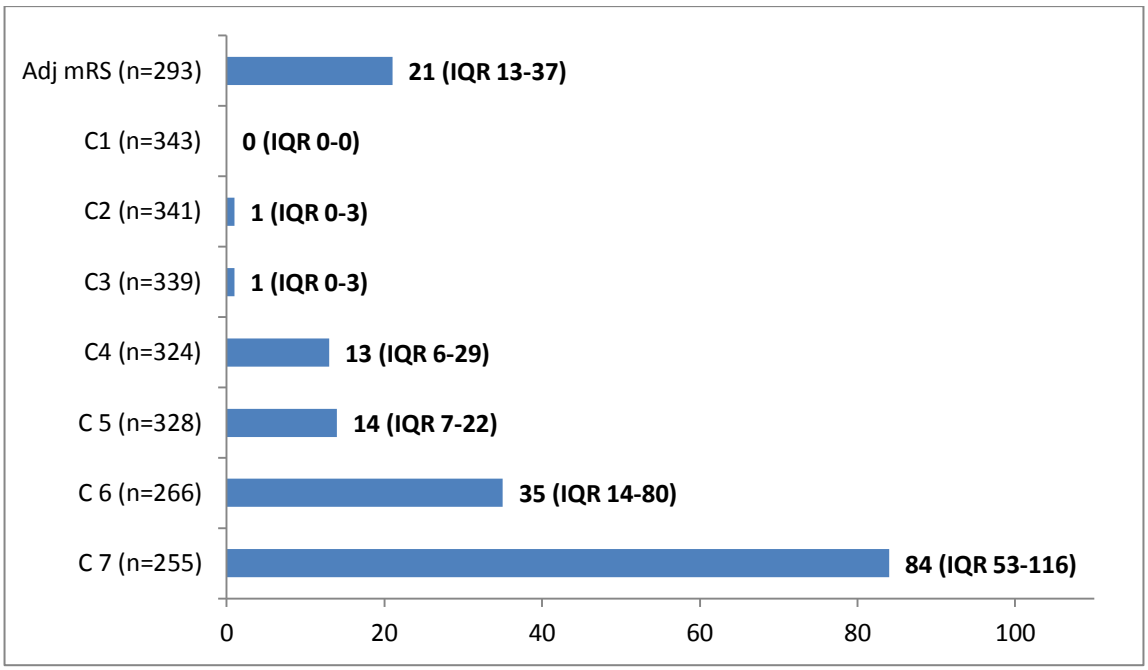


Figure 21 Number of days (median and IQR) to mRS score entry (Committee members C1 to C 7 and final adjudicated mRS score)

4.3. Discussion

4.3.1. The "CARS" web portal

The central adjudication model relies upon the appropriate IT infrastructure to support transfer and timely scoring of local mRS assessments. As an exploratory trial designed in part to assess the feasibility of the central adjudication model, we were prepared to make changes as challenges arose. During the course of the trial we modified aspects of the trial protocol and data collection processes in response to initial experience.

4.3.2. The "CARS" web portal – experience of investigators

Each centre received face to face training in the use of the video equipment and web portal. Written material (See Appendix A) was also provided together with remote advice and assistance from the trial outcomes manager by email and telephone. We encountered very few problems and in general received very positive feedback from investigators.

4.3.2.1. Data entry changes

During the course of the trial the outcomes manager had several meetings with the web design team at RCB in order to discuss experience of the portal design and make changes to improve data collection. Given the large amount of data collected in the CARS portal we encountered very few issues and made only minor changes to the collection of laboratory, imaging and home time data.

4.3.2.2. Laboratory results

In the initial design we entered limits for each laboratory value which were considered physiologically plausible as a safeguard against data entry error. As an observational study we did not specifically request that certain blood tests were done in the protocol. We expected that the laboratory parameters that were included in the CARS study would be tested ubiquitously in all centres, however this was not the case and in some centres there were no values available for certain fields in the web page (most commonly blood glucose). We also found that our physiological limits were too narrow for the blood glucose field, where a patient had a blood glucose result at the extremes of physiological parameters these data could not be entered.

The web page had originally been designed so that the page could not be saved until all fields were complete, again to safeguard against error and missing data. However, we quickly found that this feature was in fact reducing the completeness of data collection as some investigators were unable to save and progress though the form where there were missing data. We made changes to widen the data entry limits for blood glucose and to allow any data entered to be saved. We also included an option “not done” so that the empty fields could be acknowledged as complete despite missing data.

Each form of the eCRF was colour coded in the summary view for each centre. Once a form was “complete” the link to that form changed from green to blue. This colour change served as a visual prompt to investigators to ensure that data entry was complete. The colour of the box which was selected for each form on the eCRF did not change to “complete” until all fields were filled, either with a value or “not done”.

4.3.2.3. Imaging results

In the initial design of the web portal we had not considered the heterogeneous design of radiology services across various sites. Based upon our local experience we anticipated that most imaging results would be available within a fairly short time frame. Delays in accessing imaging services and reporting of examinations varied considerably among sites. Imaging results were included in the baseline eCRF, usually completed within the first 24-72 hours of enrolment. Again, as an observational study we did not stipulate that reports must be available for enrolment. Participants could be enrolled on the basis of a clinical diagnosis without imaging. We made changes to allow entry of imaging data retrospectively at future follow up visits and again to allow the form to be saved at an early stage with data awaited or the "not done" field selected if no results would be expected. This was most common in the carotid Doppler section due to varied practices in frequency and modality of carotid imaging.

4.3.2.4. Home Time

Home time is the number of nights that a participant spends back in their own home between enrolment and the follow up visit; in CARS at day 30 or day 90[181]. In many cases this was simply the number of nights between discharge and follow up, however in some cases it was more complex and collection of these data in a standardised fashion within the eCRF was challenging.

The following is an example of how home time is calculated for illustrative purposes. A participant is enrolled on day 0 and followed up on day 30 and day 90. If they are discharged from hospital on day x and remain at home then home time would be (30 – x) or (90 – x) at each follow up visit. However, if that participant is re-admitted to hospital for any reason or spends time in any other institutional care (eg. Nursing home for respite), then the number of nights (y) that they spent in this care environment must be subtracted. Home time is expressed as (30 – (x + y)) or (90 – (x + y)). We felt that leaving this calculation to individual centres was open to error and would not provide enough certainty that the data was entered accurately.

Initially the web portal was designed to calculate the difference between the discharge date and the date of the follow up visit. We soon discovered that this method made two assumptions which jeopardised the home time data. The first was that participants were not readmitted. The second was that the follow up visit always occurred at exactly 30 or 90 days after enrolment. This was not practical given the necessary flexibility that was required to account for weekends, holidays and availability of participants and investigators.

We made changes to the web portal to allow entry of each portion of the above equation and to standardise the limit of the calculation as 30 or 90 days after enrolment. Investigators were prompted to answer the following questions detailed in Figure 22.

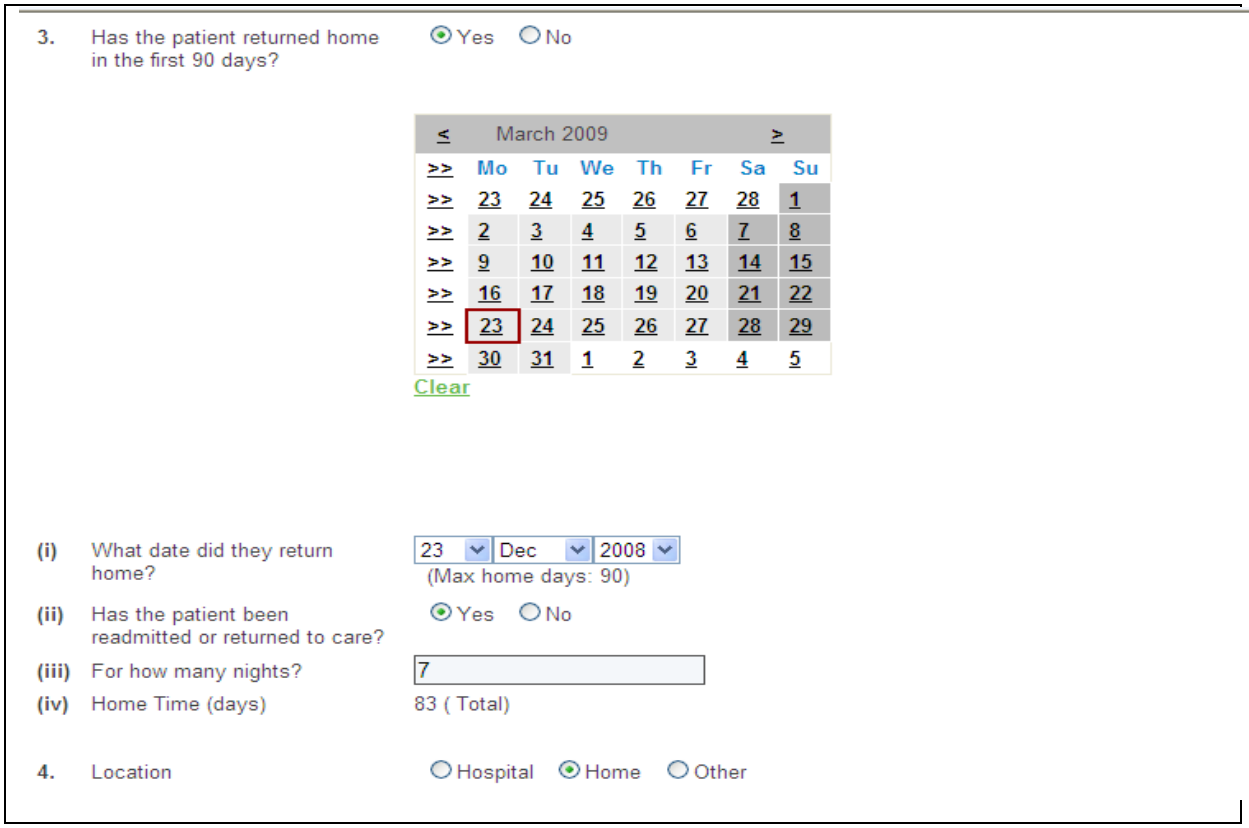


Figure 22 - Data collection screen for home time data

4.3.2.5. NIHSS assessment

Initially we planned to include a video recording of the NIHSS assessment at 90 days for additional adjudication. Our initial experience, having completed 90 days of follow up in the first 10 participants demonstrated that there were considerable technical problems with

adequate recording of NIHSS. The NIHSS scale requires a view of the participant from several angles and with varying degrees of proximity to the camera – i.e. in order to view large movements from a distance (such as limb power) and fine movements at close range (such as eye movements). In order to do this successfully appropriate accommodation and a dedicated camera operator was necessary. The video footage required substantial editing to produce a single clip for upload and scoring. Investigator numbers were not sufficient at local sites to facilitate this process and to avoid this affecting future recruitment we chose to remove this video assessment from the protocol. Ethical approval for this amendment was granted in June 2009 in light of the fact that this did not impact upon the overall aims of the study.

4.3.2.6. mRS assessment and video upload

In most cases local investigators were able to deliver mRS assessments of high quality and negotiate their upload through the web portal with little difficulty. Written guidance and remote assistance via telephone or email was available from the trial outcomes manager throughout.

In seven clips the sound quality was too poor to enable scoring. The use of an external boundary microphone with the Canon camera added multiple opportunities for error; connecting the microphone, switching the microphone on and ensuring adequate battery power. At the end of each interview the microphone had to be switched off to ensure that the battery did not run flat between interviews.

The Canon camera required an external USB cable to upload clips. There were four steps to access the correct folder once the camera was recognised by the computer. Once the correct clip was identified it then had to be converted using the AVS converter software and re-saved to the investigators "CARS" folder with the appropriate file name for that visit. Again, these complexities increased susceptibility for error. In six cases a duplicate file was uploaded incorrectly, presumably due to error in re-naming the video files. In one centre five clips were lost when their Canon camera had a fault. Investigators were encouraged to upload each clip on the day of filming, however, due to the time taken to connect, convert, save, rename and

upload each clip some investigators found it time efficient to save the clips up for upload in batches.

In the planning stages of the trial there were concerns regarding the capabilities of the existing NHS IT infrastructure to handle video files. For data protection reasons our ethical approval was granted on the understanding that secure NHS networks were used for file manipulation and delivery. We found that local IT departments were very helpful in setting up the required video conversion software. We were also pleased to find that the NHS network speed coped well with video upload in all centres. These were not barriers to using this technology in clinical practice.

Each of these steps rendered the process more susceptible to error. Despite these complexities, it is important to emphasise that over 95% of clips were uploaded successfully and scored by the adjudicating committee. The newer Flip camera system together with technical advances in the two years that the study progressed simplified and substantially improved the upload process. Large files could be uploaded, mitigating the need for file conversion with either camera system. The direct access to upload with the Flip camera's integrated USB stick encouraged prompt upload.

In a small minority of cases (n=23, 4.2%) there were significant technical problems which precluded scoring of the mRS clips (See table 19, section 4.2.8.1). The majority of these occurred early in the study and were related to the use of the original Canon camera system. A learning effect was seen in investigators as the study progressed in both mRS interviews and technical skills. It is important to note that had this been a real intervention trial there would have been a backup outcome measure in these cases in the form of the local mRS score.

4.3.3. The "CARS" web portal – experience of the outcome manager and endpoint committee members

The trial outcome manager had access to all areas of the portal to enable informal study monitoring and to keep track of enrolment, data entry and video upload. This was achieved

through several tabs at the top of the web page to enable the outcome manager to switch between rolls easily (See figure 15, Section 3.6).

Automated emails were generated to the outcome manager to inform them of new video upload. This enabled timely review of clips for anonymity, technical adequacy and completeness of interview.

4.3.3.1. Anonymity

Editing for anonymity was required in a small number of clips (n=39, 7.2%). Clips were edited if there was any mention of patient identifying information such as forename, surname or date of birth. Editing was minimised to avoid disruption of the interview and did not affect the content of the mRS assessment except where the investigator had mentioned their score at the end of the clip. In these cases the score was removed to ensure blinding.

Anonymity was an important component in the ethical considerations in this study and details of this were incorporated in the protocol and patient information sheets to reassure participants of our commitment to data protection.

4.3.3.2. Interview content and quality

In general, the interview quality in all video clips was very high. Very few centres performed a structured or semi-structured interview (structured mRS score recorded in 75/563 assessments; 13.0%). In four centres a structured mRS score was recorded as standard. In two centres there was a structured mRS score recorded in some participants only. There was no clear pattern in which participants had a structured mRS score recorded suggesting that this was not part of standard practice.

As this was an exploratory study of feasibility we were liberal in accepting most interviews unless there were stark omissions in the discussion regarding mRS grade. No clips were considered inadequate on interview content alone. In two cases there was a failure of microphone battery midway through the interview which precluded scoring. In three cases endpoint committee adjudication requested further details from the local investigator to

clarify the adjudicated score. In three cases the committee were unable to reach consensus but scoring was delayed to a point where a request for further information was not practical.

There was heterogeneity in the length and content of each interview. The majority of clips were under 10 minutes long, most were between 4 and 6 minutes long (mean duration 5minutes 32seconds (SD 3minutes 20seconds). This heterogeneity is a function of all traditional mRS interviews (face to face or video) and reflects the complexities of some stroke survivors disability, particularly where there are co-morbidities or non-stroke related limitations. These complexities are present in all traditional mRS scores; including non-video mRS interviews and are not a function of the central adjudication model per se. We encouraged local investigators to complete their interview as per their usual practice with no expectation that the video clip should be deliberately succinct.

As our intent in the CARS study was to assess the variation in scores between local investigators and video based scoring; we deliberately designed the trial such that we were blinded to the local score at the time of consensus meeting. In practice, cases where there were omissions in the interview or disagreements at a committee level a reasonable approach would be to contact the local site evaluator for further information. Such contact has both scientific and educational value unless the local rater may be prejudiced through knowledge of treatment assignment.

4.3.3.3. Time to adjudicated mRS scores

The different approaches to scoring misclassified clips were a limitation in the conduct of this study. As each endpoint committee member was scoring clips in their own time there was considerable delay in discussion of each clip at an endpoint committee meeting for a consensus score to be recorded. This did reduce the availability of further information from local centres if there were queries regarding the content of the interview. Adjudicated scores were recorded a number of months after the original mRS interview was uploaded in extreme cases. In two cases the endpoint committee could not reach consensus but were unable to ask for further information due to the significant delay in committee discussion. There were also a number of clips where an adjudicated score was reached but without

unanimous committee agreement in scoring (See chapter 5), in these cases the ability to ask further details of the local investigator may have clarified points causing controversy.

4.3.3.4. Co-ordination of endpoint committee activity

Endpoint committee members were based in the same city (Glasgow) but worked in various centres across the city. The co-ordination of endpoint committee activity was achieved predominantly using email. As clips were allocated to endpoint committee members for scoring an automated email would alert them that there was a clip for review. Once a clip was scored by the original four endpoint committee members an automated email would be generated to the outcome manager in the case of any disagreement (misclassified clips) and this clip would become available for all committee members to view prior to committee meeting discussion. End point committee meetings were arranged via online meeting scheduling software, Doodle©.

Our target was that each clip would be viewed and adjudicated within 7 days of upload, to allow further information or repeat assessment to be requested from local investigators where required. Unfortunately this was not possible due to the time taken by some endpoint committee members to score their allocated clips (see section 4.3.3.3). Most committee members did not have dedicated time for the study and were scoring clips at their convenience around other clinical and academic commitments.

As the study progressed it became clear that with each contentious clip there were often similar reasons for a committee member to choose a particular score or swing to one or other side of a controversial boundary. These thought processes are important in assigning the initial score and in subsequent discussion of each clip. We felt that allowing committee members to document comments pertaining to these scoring decisions would help to facilitate endpoint committee meetings. We arranged to have the web portal changed and allow each committee member to input comments together with their score if they wished. These comments and scores were then available at each committee discussion.

This alteration had multiple benefits. In the CARS study endpoint committee meetings were held face to face. Due to competing commitments it was not always possible for all

committee members to be present. The availability of comments from each member allowed their thoughts to be considered when assigning an adjudicated score. An extrapolation of this might allow committee discussion to occur remotely, via email or video/teleconferencing; facilitating more timely adjudication, improving access to further information from local investigators and allowing involvement of endpoint committee members from geographically disparate areas.

4.3.4. Study Completion

The majority of participants completed the study, attending for three study visits with two mRS video assessments. We experienced a greater proportion of withdrawals than might be expected in an active treatment trial (17% visits missed). The completeness of video recordings was excellent; in 96% of cases where a local mRS was recorded an adjudicated mRS was reached.

A small number of participants were withdrawn following a serious adverse event; 15 deaths and one following pulmonary embolism. The majority of withdrawals were recorded as "subject unwilling to continue" or "lost to follow up". Only two participants actively withdrew consent for the study, one of whom had originally been included with proxy consent.

Our study population had a large proportion of participants with mild stroke. Investigators favour participants able to give their own consent in observational research therefore emulating a true acute stroke interventional trial is challenging.

4.3.4.1. Withdrawals

In any clinical trial there are opposing goals; it is important to collect sufficient data but also to minimise participant burden as far as possible by minimising frequency and intensity of intervention. At the point of consent it is important that the participant understands what is involved in the study to the point of completion. Attrition of participants is a threat to the quality of all data collected and losses of 20% or greater are considered a considerable threat to the validity of trial data[183]. One of the cornerstones of ethical research is that participants can withdraw at any time, for any reason and without any detrimental effect on their care.

Some argue that this may jeopardise clinical research and that a non-exploitative and autonomous "contract" might be reasonable prior to enrolment to ensure that participants understand the wider effects should they choose to withdraw[184, 185].

Withdrawals are a common phenomenon in clinical research and the rates of non-completion in trials varies considerably depending upon the nature of the medical condition in question, the study population, the intervention and the study procedures. In studies involving psychiatric disorder[186-188] or dementia[189-191] there are high rates of attrition. It is well recognised that the elderly are more likely to dropout from research than other populations[191-193]. Randomisation is frequent in clinical research and best practice would blind treatment allocation and outcome to both the participant and investigator. The act of randomisation can affect consent to participate and remain in clinical research[194]. In unblinded trials it is recognised that participants allocated to a control group have increased rates of withdrawal[195]. There is a reasonable perception that participation in clinical research is less attractive in a placebo group or where there has been no intervention, such as in our observational study.

Several techniques are used to optimise participant retention despite these issues. In many areas of the world where access to healthcare is limited and expensive; involvement in research is an incentive in itself as it ensures provision of treatment and follow up. In areas where there is publicly funded universal health care this is less important. Remuneration for participants directly is ethically unsound, however payment of expenses for travel and time from work are often utilised. Remuneration of investigators is more widely accepted, not only for enrolment but for follow up visits as well. Academic studies are more susceptible to participant withdrawal; modest funding budgets are less likely to extend to generous site remuneration for follow up and close data monitoring.

The larger than expected proportion of withdrawals in our study may be attributable to the observational nature of the study and the mild clinical deficits. The mild nature of stroke events allowed the majority of participants to return to their usual active lives within the study period, reducing their motivation for participation in research. In a study of risk factor management in stroke patients, recruiting at a similar time point to the CARS study found

that 48% of patients did not attend for follow up at one year, often due to lack of interest[196]. As there was no intervention in our study there may also have been a perception amongst participants that they would not be impacting on valid research by not attending.

Withdrawals might be considered an important endpoint in themselves[197], particularly if they are due to poor tolerability of a study drug or unacceptability of study procedures. The majority of withdrawals in our study occurred before the first follow up assessment and therefore before any active intervention. There were fewer withdrawals between the 30 and 90 day visits which could arguably be attributable to study procedures. The video adjudication process was acceptable to the substantial majority of investigators and participants.

4.4. Conclusions

In stroke trials, mRS outcome data can be collected using a central adjudication model. The use of patient interview videos to assign mRS grades is used in mRS training and certification[151] and so is familiar and acceptable to most investigators. Across 14 sites, we initiated and trained investigators, in many cases with limited stroke research experience. Recruitment was fast and exceeded our target of n=300.

The use of video technology (see section 3.5) and the "CARS" adjudication web portal (see section 3.6) were the main focus in the investigation of feasibility. Both video systems and the web adjudication portal were very successful. An adjudicated score was available for 96% of study visits. We have demonstrated a high rate of technical success for assessment upload with the majority of failures occurring early in the study. A learning effect was seen in investigators as the study progressed and it is important to note that had this been a real intervention trial there would have been a backup outcome measure in these cases in the form of the local mRS score.

We were able to utilise technological advances as the study progressed and incorporate these into our central adjudication model. It is important to emphasise that video and computer technology advance very quickly; with the advent of tablet computers with video

and internet capabilities this process could be simplified further without greater expense. The cost of the original Canon camera system was £700 in this study; current tablet computers with internet access and integrated video cameras are substantially cheaper than this. Recording a brief video clip and uploading this to the internet is familiar to any user of social media. We have demonstrated that using analogous technology in clinical trials is possible with the necessary data protection security.

We have developed and assessed a system for central adjudication of mRS endpoints that is feasible; data collection is simple, inexpensive and acceptable to participants. Technical handling of video recording and uploads and committee review has been successful. Further technical advances and the use of the portal review comments may help facilitate faster remote adjudication, increasing the usefulness of local investigating teams to answer queries and provide clarification.

Chapter 5

Reliability: is a central adjudication model reliable?

5.1. Introduction

Previous studies assessing the inter-rater reliability of the mRS have predominantly been conducted in small, single centre studies with highly motivated individuals. Inter-observer reliability in a large scale clinical trial with the associated challenges is likely to be poorer. No prior study has assessed measures of mRS reliability in a large multicentre and multinational study with observers from varied professional backgrounds and with varied levels of experience; a design which parallels a contemporary stroke RCT in practice.

In our 14 centre study we collected multiple mRS scores for each study visit, as detailed in the methods section (Chapter 3). Each mRS interview was initially scored by a local investigator and was subsequently scored by four adjudication panel members. A clip was designated as "misclassified" where there was disagreement amongst panel members, these clips were then forwarded for full adjudication committee review to reach a consensus score. At the close of this process each clip had a final adjudicated score, either taken from an agreed score at initial review (classified clips) or based upon the consensus decision taken at committee review (misclassified clips).

5.2. Method - Statistical Analysis

5.2.1. Inter-Observer reliability

Agreement between assessors was measured using kappa statistics (κ / κ_w Fleiss-Cohen Weights $[1-[(i-j)/(1-\kappa)]^2$), intraclass correlation coefficient (ICC) and Bland Altman plot. Agreement was assessed for the following comparisons: local mRS versus final adjudicated panel mRS, local mRS versus mean panel mRS, local mRS versus a random panel mRS (generated using R statistical software). We also assessed agreement between the mean of two paired panel mRS scores, between all individual panel mRS scores and between the individual panel mRS scores and the local score together.

We assessed for any differences in reliability as the study progressed to quantify any "learning" effect in the adjudication committee by measuring reliability in clips scored early or late in the study. Each participant had video clips assessed at various times during the study (local mRs assignment immediately, four endpoint committee mRS scores in the following weeks and finally the adjudicated mRS score assigned later still). We grouped the early and late assessments by splitting the participants into those assessed early and late around the median of the mean assessment dates. Agreement in clips scored early and late (κ and κ_w ± 95% confidence interval) was calculated for the difference between groups (based on 10000 bootstrap samples).

5.2.2. Measurement Error among observers

It is important to quantify if there is any error (systematic bias or inconsistency) in the scores of individual raters, both to ensure that individual scores do not skew the panel process but also as a measure of quality control of individual assessors. We cannot quantify the "true" level of disability on the mRS; neither the local or adjudicated mRS scores are an accurate gold standard. For this reason we used statistical modelling techniques to quantify measurement error among observers. We used only day 90 mRS clips to avoid including repeated measures in the model.

We examined the error among individual raters using a mixed model designed to estimate the error between observers, measured as variance in mRS scores. This included a measure of each adjudicators bias (or accuracy) and consistency (or precision). Bias was quantified in comparison to a surrogate gold standard mRS score, the mean of all mRS scores for each clip. Consistency was modelled as the individual error of each score entered by an individual rater.

ICC was estimated using linear mixed effects models with normally distributed random effects fitted by restricted maximum likelihood (REML). Variance in mRS was decomposed into components due to: variation between patients (V_P) and between study centres (V_C), which together comprise the real variation in disability that we want to measure; consistent differences (bias) between observers relative to the mean of all observers (V_O); and residual error (V_E), which quantifies the measurement error, or inconsistency, of an individual observer. The square root of V_E can be interpreted as an observer's "inconsistency standard deviation (SD)" measured in mRS units.

ICC is the proportion of variance that is due to differences between patients, that is:

$$ICC = \frac{V_{patient} + V_{centre}}{(V_{patient} + V_{centre}) + (V_{0[bias]} + V_{e[Error\ inconsistency]})}$$

The real variation in disability that we aim to measure

The real variation in disability that we aim to measure + unobserved noise term (bias and error)

Specific hypotheses were tested using likelihood ratio tests of REML models refitted using maximum likelihood. We tested for bias among observers, which is a test of the null hypothesis $V_O = 0$, and for heterogeneity of inconsistency SD between observers and groups of patients. We compared reliability between local and panel-derived mRS scores by

estimating the inconsistency SD separately for local and panel mRS scores and testing for a difference. We did not include observer bias in this comparison of reliability because, in the absence of a gold standard mRS score, absolute bias cannot be estimated.

5.2.3. Predicted Reliability with multiple mRS scores

To estimate the predicted reliability that would be delivered by combining multiple ratings we used the Spearman Brown prediction formula[198].

$$\text{Predicted Reliability} = \frac{np}{1 + (n-1))p}$$

n = Number of raters

p = Reliability of a single rater

We used the observed reliability of a single panel member (ICC) at day 90 to predict the likely improvement in reliability (ICC) with groups of up to ten raters. Using our data we compared our observed figure for reliability in two raters to the predicted figure. The observed reliability for two raters was calculated by selecting at random (using R statistical software) two pairs of panel mRS scores from each subject without replacement, taking the means of each pair and estimating the ICC between the two mean mRS scores.

A summary of the mRS reliability analysis and the sample of mRS clips used for each component of the analysis is shown in figure 23

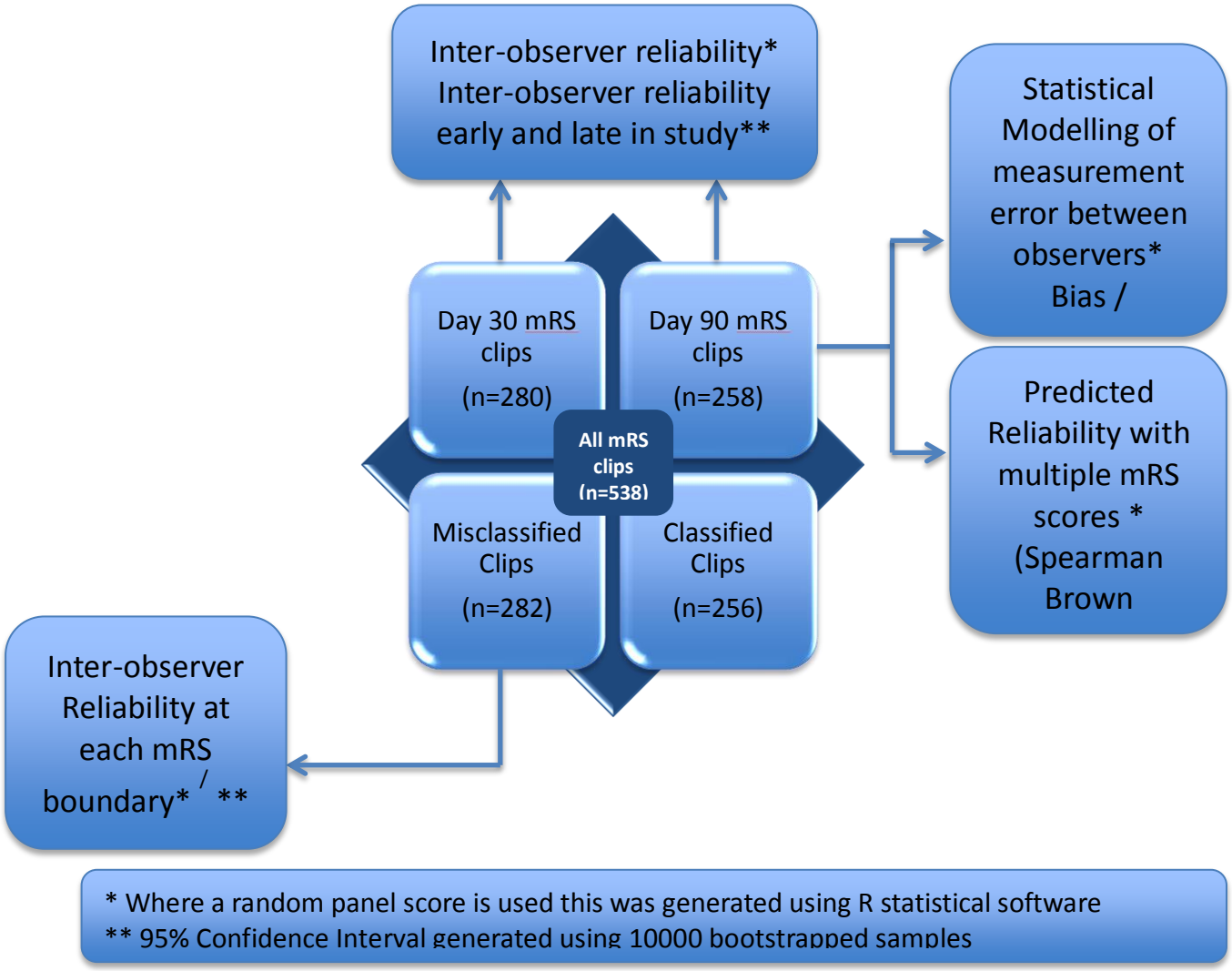


Figure 23 - Summary of reliability analysis and sample of mRS video clips used for each component of analysis.

5.3. Results

5.3.1. Misclassified mRS assessments

Clips were misclassified if there was disagreement among the four scoring members of the adjudication committee. At 30 and 90 days respectively, 57.5% (161/280) and 50.8% (131/258) of clips were misclassified. See figure 24

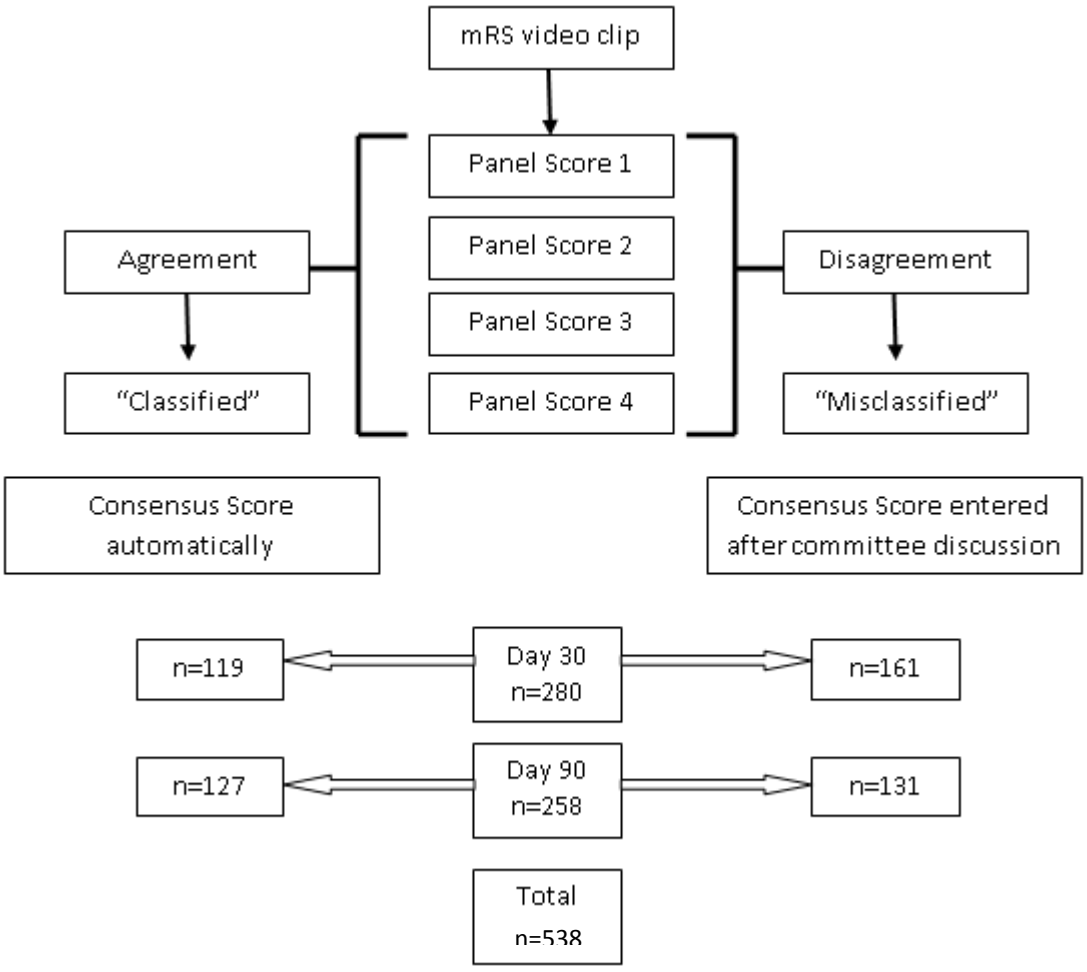


Figure 24 - mRS video clip adjudication process: classified / misclassified clips at Day 30 ad Day 90

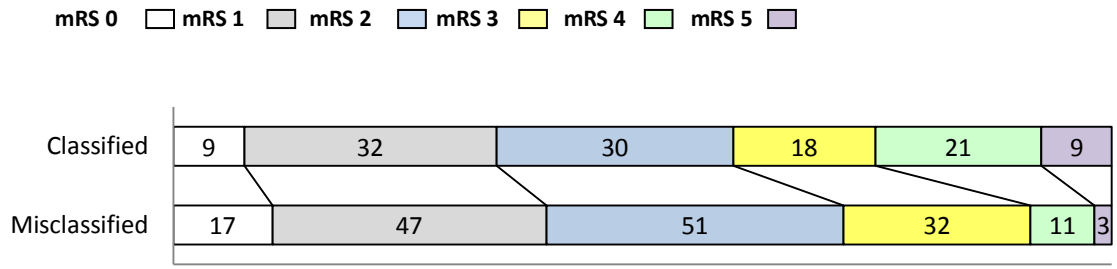
Misclassified clips were forwarded to the entire endpoint committee for review and discussion. The endpoint committee met on 12 occasions during the study to discuss misclassified clips. Details of the endpoint committee meetings are described in section 3.6.3.

Unanimous committee agreement was reached after panel review in 89.4% (261/292) of mRS assessments. A non-unanimous committee decision was reached in 9.2% (27/292). In only three cases were the committee unable to reach consensus (0.5%). The distributions of local and adjudicated mRS at 30 and 90 days are shown in figure 25.

Of the clips that resulted in complete agreement in end point committee scores ("classified") there was disagreement with the local score in a proportion of cases. At 30 days, 31/119 (26.1%) classified clips disagreed with the local score. At 90 days, 40/127 (31.5%) clips disagreed with the local score. Again there was no clear pattern or systematic bias in these disagreements. The local versus "classified" adjudicated mRS scores are cross tabulated in Table 20.

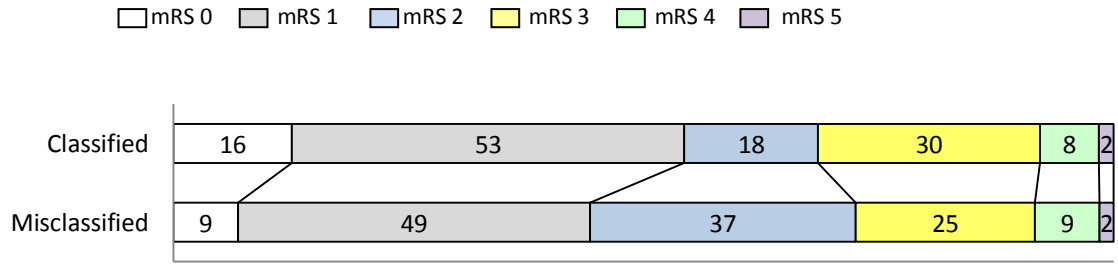
5.3.2. Inter-Observer Reliability in mRS assessments

Agreement between the adjudicated panel score and local mRS score at 90 days was good; κ 0.48 (95% CI 0.40-0.55), κ_w 0.80 ((95% CI 0.75-0.84) and ICC 0.8. The use of the mean panel mRS score improved reliability at 90 days; κ 0.50 (95% CI 0.42-58), κ_w 0.83 ((95% CI 0.78-0.87) and ICC 0.84. The use of a single random panel score did not result in a benefit in reliability; κ 0.43 (95% CI 0.34-0.50), κ_w 0.78 ((95% CI 0.78-0.83) and ICC 0.79. Agreement amongst panel members was very good at 90 days; κ 0.59 (95% CI 0.53-0.63), κ_w 0.86 ((95% CI 0.82-0.88) and ICC 0.87. The addition of the local mRs score did not reduce reliability; κ 0.55 (95% CI 0.51-0.60), κ_w 0.84 ((95% CI 0.80-0.87) and ICC 0.84. Agreement at 30 days was similar to that seen at day 90. Table 21.



Common Adjudicated Score	119
Adequate assessment with unanimous committee agreement	138
Adequate assessment with non-unanimous committee decision	20
Inadequate assessment - scored with additional info from centre	3
Clip unable to be scored	16
Total	**296**

A Day 30



Common Adjudicated Score	127
Adequate assessment with unanimous committee agreement	123
Adequate assessment with non-unanimous committee decision	7
Inadequate assessment - scored with additional info from centre	1
Clip unable to be scored	9
Total	**267**

B Day 90

Figure 25 - Distribution of mRS Scores and committee outcomes at (A) Day 30 and (B) Day 90.

Table 20 - Cross tabulation of mRS scores where "classified" (agreement among committee members) scores disagree with local score

Day 30 mRS scores								
		Local mRS Score						
Adjudicated mRS Score (Classified)		**0**	**1**	**2**	**3**	**4**	**5**	**Total**
	0		2					2
	1	2		2				4
	2	2	9		1			12
	3		1	2		1		4
	4			1	4		1	6
	5					3		3
	Total	4	12	5	5	4	1	**31**
Day 90 mRS scores								
		Local mRS Score						
Adjudicated mRS Score (Classified)		**0**	**1**	**2**	**3**	**4**	**5**	**Total**
	0		9					9
	1	3		3	1			7
	2	2	9		3			14
	3			3		2		5
	4				4		1	5
	5							0
	Total	5	18	6	8	2	1	**40**

Table 21 - Inter observer reliability in mRS scores at Day 30 and Day 90. [Agreement between local score and various methods of generating adjudicated score; agreement amongst panel members and agreement amongst all available scores]

	Number of scores compared (n)	Kappa (κ) (95% CI)	Weighted Kappa (κ_w) (95% CI)	Intraclass Correlation Coefficient (ICC)
Day 30				
Adjudicated Panel vs Local	2	0.53 (0.45-0.60)	0.84 (0.80-0.88)	0.85
Mean Panel vs Local	2	0.51 (0.42-0.58)	0.84 (0.79-0.88)	0.87
Mean of Paired Panel Scores	2	0.63 (0.56-0.70)	0.92 (0.89-0.94)	0.92
Random Panel vs Local	2	0.46 (0.38-0.53)	0.82 (0.77-0.86)	0.82
Individual Panel Scores	4 – 7	0.55 (0.50-0.60)	0.85 (0.81-0.88)	0.86
Individual Panel and Local Scores	5 – 8	0.53 (0.48-0.58)	0.84 (0.81-0.87)	0.85
Day 90				
Adjudicated Panel vs Local	2	0.48 (0.40-0.55)	0.80 (0.75-0.84)	0.8
Mean Panel vs Local	2	0.50 (0.42-0.58)	0.83 (0.78-0.87)	0.84
Mean of Paired Panel Scores	2	0.65 (0.57-0.71)	0.92 (0.89-0.94)	0.92
Random Panel vs Local	2	0.43 (0.34-0.50)	0.78 (0.78-0.83)	0.79
Individual Panel Scores	4 – 7	0.59 (0.53-0.63)	0.86 (0.82-0.88)	0.87
Individual Panel and Local Score	5 – 8	0.55 (0.51-0.60)	0.84 (0.80-0.87)	0.84

Bland Altman plots demonstrating the agreement between local and adjudicated mRS clips is demonstrated in Figure 26.

Inter-rater reliability varied across different mRS boundaries. The proportion of clips with disagreement at each level of mRS score was as follows (mRS 0: 51%, mRS 1: 53%, mRS 2: 65%, mRS 3: 54%, mRS 4: 41%, mRS 5: 31%). Greatest agreement was seen across the mRS 2-3 boundary (κ 0.81 95% CI 0.74-0.87) with poorer agreement seen at 0-1 (κ=0.66 95% CI 0.53-0.75), 1-2 (κ=0.70 95% CI 0.64-0.76) and 4-5 (κ=0.75 95% CI 0.60-0.84). Reliability around the 4-5 boundary was not assessed because too few participants were assigned a score of 5 to allow κ to be estimated. Table 22

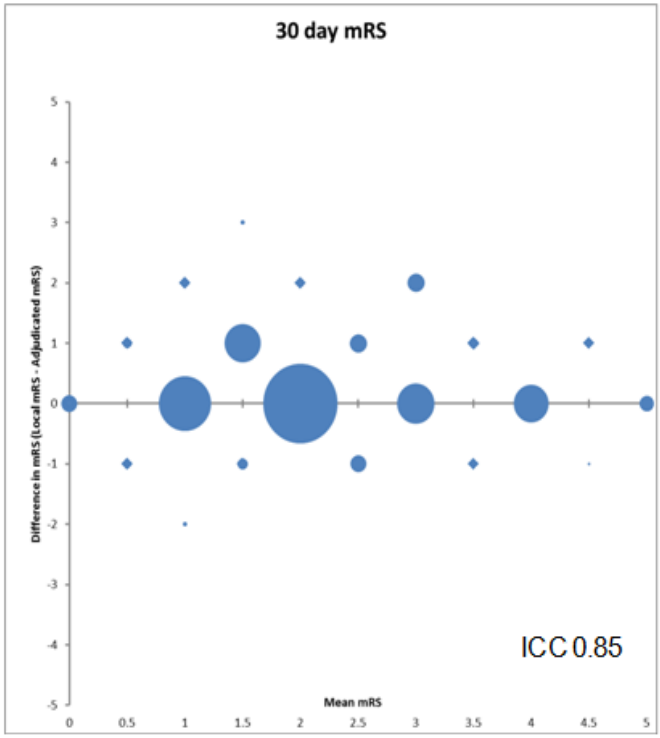
Table 22 - Reliability of dichotomised mRS scores at each mRS boundary at day 90. Inter-rater reliability (κ) with 95% CI derived from 10 000 bootstrapped samples.

mRS boundary	0-1	1-2	2-3	3-4
κ (95% CI)	0.66 (0.53 – 0.75)	0.70 (0.64 – 0.76)	0.81 (0.74 – 0.87)	0.75 (0.60 – 0.84)

There was no significant difference in agreement in day 90 clips that were scored early (κ_w 0.876) versus late (κ_w 0.824) in the course of the study (p=0.146). This applied also for 30 day assessments (κ_w 0.876 early versus κ_w 0.824 late p=0.146). Table 23

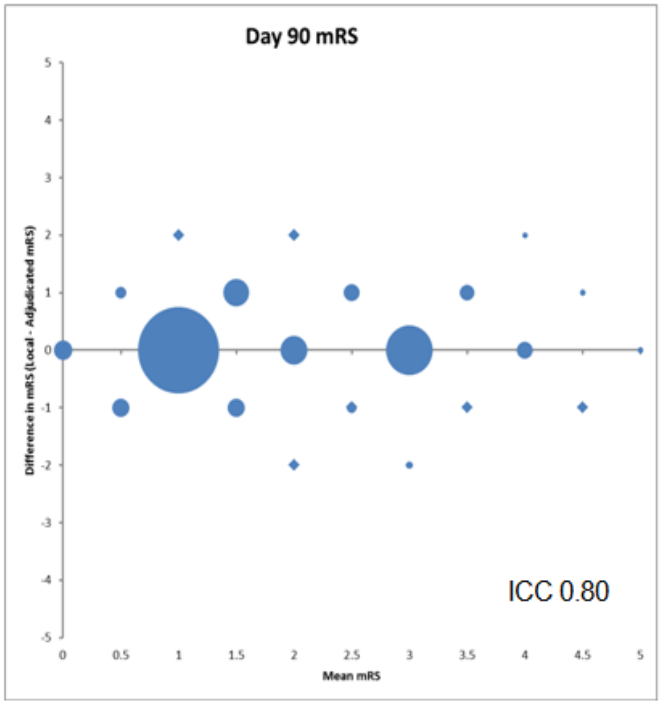
Table 23 – Inter-Observer agreement of clips scored early and late in study. Participants divided by those assessed early and late around the median of the mean assessment dates (* 95% CI for difference derived from 10 000 bootstrap samples)

Visit	Early (κ_w)	Late (κ_w)	95% CI for difference *	P
Day 30	0.856	0.841	-0.081 – 0.057	0.344
Day 90	0.876	0.824	-0.134 – 0.013	0.146



Local Score

Adjudicated Score	0	1	2	3	4	5	
0	12	8	5	1			26
1	5	40	29	3			77
2	1	9	58	13	1		82
3			12	29	9		50
4				2	28	3	33
5					1	11	12
	18	57	104	48	39	14	**280**



Local Score

Adjudicated Score	0	1	2	3	4	5	
0	15	7	4				26
1	13	63	20	5			102
2		11	31	12			54
3		6	7	36	11	1	57
4			1	3	10	1	15
5					3	1	4
	28	83	63	67	24	3	**258**

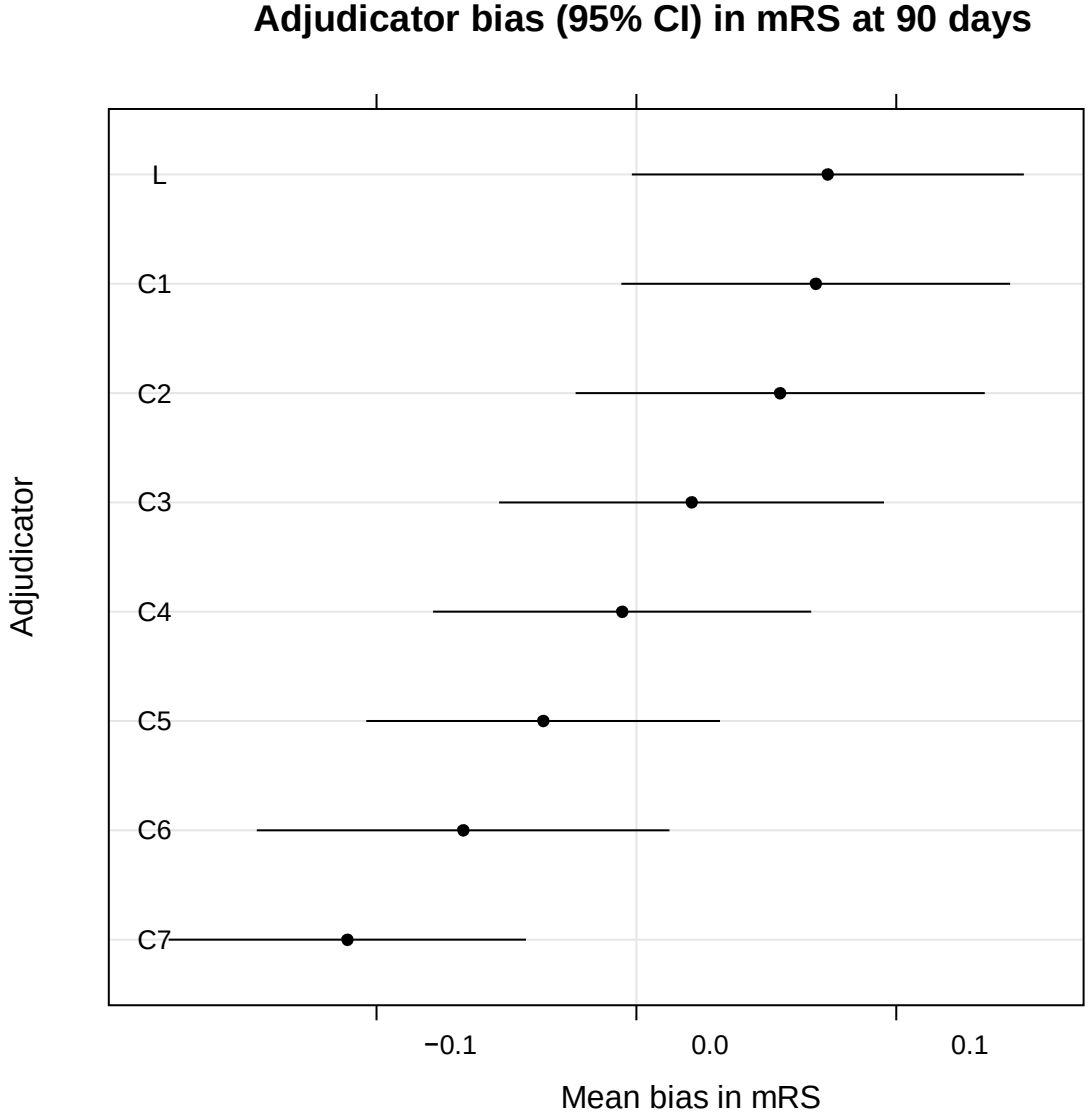
Figure 26 -Bland Altman Plot and cross tabulation of day 30 and 90 Local and Adjudicated mRS scores. Bland Altman Plot: [Difference in mRS (local – adjudicated) with mean mRS]

5.3.3. Measurement error among observers

The magnitude of disagreement among raters is demonstrated in Figure 27. Very small levels of variability on the mRS scale are seen; typically less than a tenth of an mRS grade. When comparing measurement error / consistency of panel members against performance of local raters there was no significant difference when using a single panel score (adjudicated score or random panel score). Where a mean panel score was used the panel score was more consistent than the local mRS (p=0.025). Table 24.

Table 24 - Inconsistency standard deviation (SD) estimates for panel and local mRS scores at day 90. P-values are presented from tests of the null hypothesis of homogeneity of inconsistency SD across adjudicators.

Scores compared	Panel Inconsistency SD	Local Inconsistency SD	Heteroscedasticity P-Value
Random panel score and local score	0.53	0.57	0.696
Consensus panel score and local score	0.45	0.58	0.194
Mean panel score and local score	0.33	0.56	0.025



Bias (± 95% CI) in the seven individual panel scores and local score at visit 3 estimated as the adjudicator-level residuals in a mixed model where the outcome is mRS (n = 1322) and the random effects are study site (n = 14), patient within study site (n = 258), and adjudicator crossed with patient (n = 8).

Figure 27 - Magnitude of disagreement among mRS scores

(L=local, C1-7=seven adjudication committee members)

5.3.4. Predicted Reliability with multiple mRS scores

Using the estimated reliability of a single panel member (ICC) at day 90 of 0.87 (Table 22) and the Spearman Brown prediction formula we can estimate the reliability of the mRS with multiple raters. The observed reliability with two raters (ICC 0.92) was similar to the predicted figure (ICC 0.93). Increasing the number of raters to four is predicted to increase reliability to ICC 0.96. Table 25

Table 25 - Spearman-Brown predicted mRS reliability at day 90 based on single panel rater reliability (ICC) 0.87.

N raters	Reliability (ICC)	
	Spearman-Brown prediction for ICC = 0.870	Observed
2	0.930	0.923
3	0.952	-
4	0.964	-
5	0.971	-
6	0.976	-
7	0.979	-
8	0.982	-
9	0.984	-
10	0.985	-

5.3.5. Reliability where a structured mRS approach was recorded

A small number of assessments were accompanied by a structured mRS score. Each site was able to enter a mRS score based upon a structured interview where this was available. There was no formal protocol within the study to advise upon the use of a structured approach and therefore the scores are based upon various structured mRS instruments with substantial local variation. 45 structured mRS scores were recorded at 30 days and 28 structured mRS scores were recorded at 90 days. These assessments were limited to certain sites (see section 4.2.8). Structured mRS scores were accompanied by longer clips; mean (SD) 7mins 28secs (3mins 49 secs), median (IQR) 6mins 30 secs (5mins 4secs to 8mins 52secs). An adjudicated mRS score was available in 88.9% (40/45) and 92.9% (26/28) of mRS assessments where a

structured mRS was recorded at 30 and 90 days respectively. There were more misclassified clips amongst the group that had a structured mRS score recorded; 65.0% (26/40) of structured mRS assessments at 30 days and 61.5% (16/26) of structured mRS assessments at 90 days. The local (structured) mRS scores and adjudicated mRS scores are cross tabulated in Table 26. There was more disagreement between structured mRS scores and local adjudicated scores at the higher end of the mRS score.

Table 26 - Cross tabulation of structured mRS interviews: local (structured) mRS scores and adjudicated mRS scores

Day 30 mRS scores								
		Local (Structured) mRS Score						
Adjudicated mRS Score		**0**	**1**	**2**	**3**	**4**	**5**	**Total**
	0	1						1
	1	1	5	2				8
	2	1	5	9				15
	3		1		2			3
	4				1	10		11
	5						2	2
	Total	3	11	11	3	10	2	**40**
Day 90 mRS scores								
		Local (Structured) mRS Score						
Adjudicated mRS Score		**0**	**1**	**2**	**3**	**4**	**5**	**Total**
	0	3	1					4
	1		2	1				3
	2	2		3	3			8
	3			1	2			3
	4				1	6	1	8
	5							
	Total	5	3	5	6	6	1	**26**

Agreement between structured mRS score (based upon varied structured mRS approaches) and adjudicated score was comparable to the results seen in the whole CARS sample. At 30 days κ / κ_w (95% confidence interval) was 0.64 (0.48 – 0.80) / 0.88 (0.58 – 1.19). At 90 days κ / κ_w (95% confidence interval) was 0.52 (0.33 – 0.70) / 0.86 (0.47 – 1.23). ICC at 30 and 90 days was 0.88 and 0.86 respectively.

5.4. Conclusions

Intra-rater reliability has been previously reported[99, 199]; we considered only inter-rater reliability in this study with a "virtual trial" design. In the context of a multicentre clinical trial where endpoints will be assessed only once, inter-observer variability is most relevant. Comparing local and remote mRS scores gives some measure of the inherent variation in scoring across a multicentre study.

A considerable proportion of mRS assessments were misclassified, meaning that there was disagreement among committee members. A further proportion of mRS assessments demonstrated disagreement between the local score and the agreed panel score. Inter-observer variation was considerable (κ=0.48), albeit not as pronounced as in a previous smaller study (κ=0.25)[100].The endpoint committee scores performed favourably in comparison to the local mRS score and inter-observer variability within the endpoint committee was excellent (κ 0.59, κ_w 0.86). Panel scores demonstrated a trend to suggest less disability overall than those allocated by the local investigator, but this "bias" represents a very small change in the mRS score (typically one tenth of an mRS rank). It is important to recognise that in terms of interpretation, the mean difference in mRS scores created by a therapy such as rtPA may well be less than one mRS grade and when dealing with late treatment, this could be a very small difference. One tenth of an mRS grade may represent a meaningful proportion of the anticipated treatment effect.

Our figures are comparable to those seen in systematic review of mRS reliability (κ 0.46, κ_w 0.90)[99]. In their three centre study, Wilson et al reported inter-observer variability of κ 0.25, κ_w 0.71[100]. Thus, with the addition of centrally adjudicated mRS scores we delivered in a 14 site study an outcome substantially more reliable than the nearest prior estimate of mRS reliability in a multicentre study.

There is a perception that there is less disagreement at the extremes of the scale (0-1 and 4-5 boundaries); possibly because there can only be a disagreement in one direction, because no disability or severe disability are less complex to identify or because there are fewer

participants at these points in the scale. In fact our work has found substantial disagreement at all levels of the scale.

The agreement between a single random panel score and the local score is less favourable than that seen between the adjudicated panel score, a mean panel score or the mean of paired panel scores. This suggests that there is an advantage with multiple scores. The advantage of multiple scores is supported further by the Spearman Brown prediction formula. Multiple mRS scores may be desirable but without added inconvenience to participants are unlikely to be achievable without the use of a remote scoring model such as that used in a central adjudication model.

There was no significant difference in scores generated early or late in the study. This emphasises the complexity of scoring mRS assessments and the need to evaluate each participant individually. As the study progressed, the members of the end point committee had frequent opportunity to discuss controversies and difficulties in scoring clips; including discussion of several issues that recurred. Examples of these issues include how to score a patient who has chosen not to return to an activity, who receives help with activities that may not be necessary or who had prior disability that complicated the assessment of post stroke function. Despite reaching consensus on each of these issues for individual clips, these discussions did not improve reliability as the study progressed.

Structured mRS assessment tools, including the recent Rankin Focussed Assessment[133], have been proposed to improve mRS reliability. The comprehensive mRS structured inverview was originally developed by Wilson et al[105] and has been adapted to the Rankin Focussed Assessment (RFA) by Saver et al[133]. This structured assessment form is designed to be completed in conjunction with interview of the participant, relative or caregiver and review of participant medical records. The RFA reports excellent inter observer reliability (κ_w 0.99) in a sample of 50 paired mRS assessments and has been prospectively validated as part of the ongoing FAST-MAG trial. A simplified modified Rankin questionnaire (smRSq) has found good reliability (κ_w 0.82) with a very fast administration time (average 1.67 minutes) in a sample of 50 paired mRS assessments[132] and has been validated in comparison to stroke severity[135], stroke size[136] and quality of life measures[137]. The validity of the smRS has been demonstrated

to that of the standard mRS and NIHSS[138]. It has also been validated for use via telephone consultation[137].

Compared to our group adjudication system, it could be argued that these structured tools offer an economical and simplified method of improving use of the mRS in stroke trials. It is worth noting that in the small sample of mRS assessments with structured mRS scores in our study we found increased frequency of misclassification at committee review, suggesting more variability in scoring. However, each of these structured assessments demonstrated heterogeneity in the approach taken with local variation between instruments. For this reason we must exercise caution in drawing conclusions from this small sample of structured mRS assessments. The agreement between a structured mRS score and the adjudicated score was comparable. Few of the tools have been independently validated or assessed in a contemporary randomised controlled trial context. In meta-analysis, subgroup analysis comparing structured and unstructured approaches did not affect reliability[99] and previous data suggest that questionnaire based mRS assessments confer no advantage when used in a "real world" setting[199]. The global and unstructured nature of the traditional mRs is a great advantage; without relying on individual activities of daily living there are no floor or ceiling effects in its application which are common to structured instruments.

Further study of each approach to mRS assessment is required in large multicentre trials and is necessary to determine if the limitations placed upon mRS outcomes by either a structured interview approach or central adjudication model may diminish its value as an outcome measure.

Chapter 6

Validity: is a central adjudication model valid?

6.1. Introduction

Functional outcome after stroke is highly variable and reflects the heterogeneity in the location and extent of the neurological insult. However, there are several factors that are known to predict functional outcome regardless of the nature and classification of the stroke event. In assessing a novel outcome assessment method it is important to ensure that it is both feasible (chapter 3), reliable (chapter 4) and valid in comparison to the current accepted method of outcome assessment.

As discussed in section 1.4.2.1, the mRS has demonstrated good convergent validity (with NIHSS and Barthel Index) and construct validity (with infarct volume / recanalisation rates). The aim of this chapter is to investigate the validity of a centrally adjudicated mRS by two means. First, we assessed criterion validity in comparison to the current "gold standard"; the local mRS. Second, we assessed construct validity in comparison to independent factors known to predict functional outcome after stroke.

As discussed in Chapter 4, adjudicated mRS scores show a trend to suggest less disability than their local counterparts, we sought to quantify this to ensure that there is no significant or systematic bias in adjudicated mRS scores.

6.2. Methods

6.2.1. Criterion Validity

Criterion validity describes the performance of a test in comparison to a criteria already held to be valid. We sought to compare the current accepted method of collecting mRS outcome data (local investigator interview) with our novel method of collecting outcome data (centrally adjudicated video interview). We analysed the local and committee adjudicated scores for any systematic differences using the Kruskal Wallis test of distributions for non parametric data (StatsDirect software). We included all mRS assessments with both a local and adjucated mRS score (n=538)

6.2.2. Construct Validity

Construct validity refers to a tests relationship to other accepted indicators of measuring the desired attribute. This can be used in developing a novel outcome assessment tool by correlating the test result with other known prognostic indicators. We sought to assess the performance of adjudicated mRS outcomes in comparison to local mRS scores by comparing each method of scoring with factors known to affect functional outcome after stroke. We collected data regarding several factors predictive of functional outcome after stroke, as described in section 3.4.5.1.

The predictors used were baseline NIHSS (bNIHSS), Systolic Blood Pressure (SBP), blood glucose and home time.

Spearman Rank correlation coefficients were generated for each variable with each method of assigning mRS outcome to demonstrate a simple test of association. The Spearman Rank correlation coefficient is a non parametric measure and takes into account the ordinal nature of the mRS scale.

To further assess the nature of the relationship we performed (adjusted and unadjusted) proportional odds ordinal logistic regression. The software used was SAS (Version 9.3). The significance of the association in ordinal logistic regression models using the mRS can be

tested using the the non-parametric Cochran mantel Haenszel (CMH) test which accounts for confounders in analysis[200].

6.2.2.1. Ordinal Logistic Regression and the Cochrane-Mantel-Haenszel Test

Each of these analyses was performed using ordinal logistic regression with the proportional odds model. Originally described by McCullagh in 1980[201], this model provides a useful extension of binary logistic regression (with a yes:no response variable) to situations such as the mRS with a response variable using ordered categorical variables[202].

Ordinal logistic regression allows an analysis of how a predictive value is associated with a response variable; for example: for every point increase in bNIHSS or systolic blood pressure there is a tendency to worse mRS outcome with a quoted odds ratio. This enables an assessment of how a local or adjudicated mRS outcome score is related to other factors known to affect stroke outcome (baseline NIHSS, systolic blood pressure, blood glucose and home time) and to determine if either method of assigning mRS is more closely related to these factors. Ordinal logistic regression can also be performed with adjustment for the other covariates.

Intrinsic to the proportional odds model is an assumption of ordinality. In simple terms, this assumes that each level of mRS is affected equally by the predictor, that is the odds ratio for a better or worse outcome is identical at each level of the scale[203]. This assumption is often not met and therefore a conservative measure of significance testing is required. The Cochrane-Mantel-Haenszel Test can be used to test the significance of the association between the predictor and outcome variable (or in clinical trials the treatment and outcome variables[200]).

6.3. Results

6.3.1. Criterion Validity

We analysed for any systematic difference between local and committee scores. The mRS distributions of local and adjudicated scores were similar. There was a trend for local scores

to be higher than committee scores (i.e describing more disability) but this was not statistically significant (p=0.160). Figures 28 and 29 display the distribution of mRS scores from local investigators and the adjudication committee (mean and median of initial n=4 committee scores and the final adjudicated score).

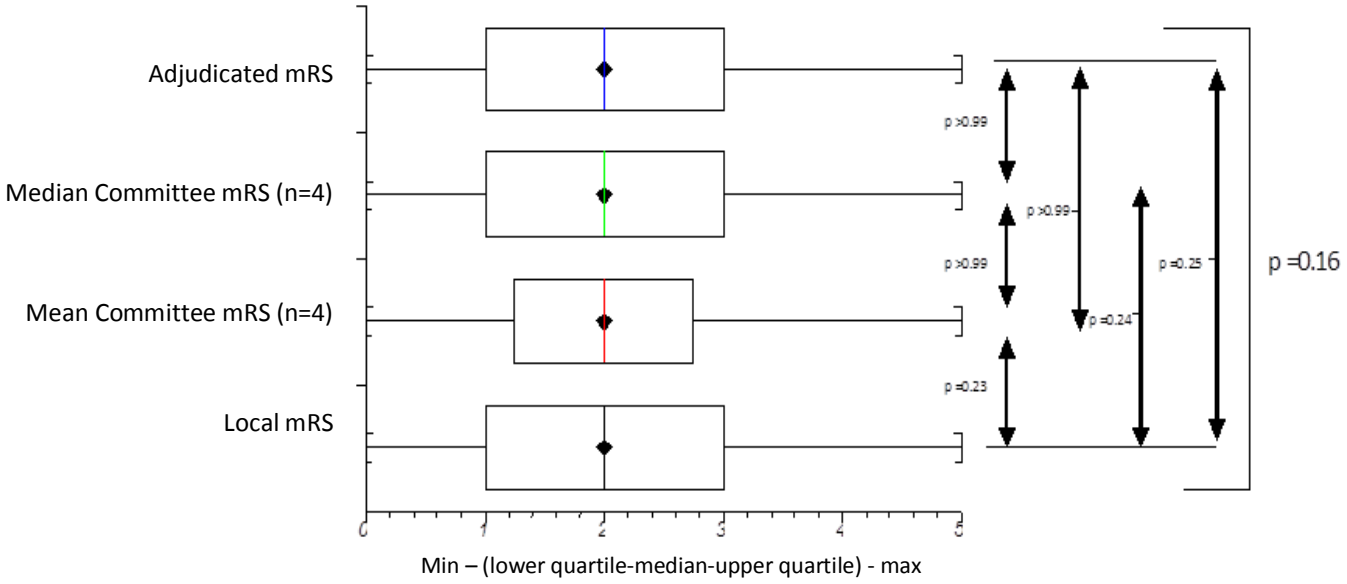


Figure 28 -mRS Distribution of Local and Committee Scores (Median / IQR). p-values represent the Kruskal Wallis test of difference between distributions. n=538

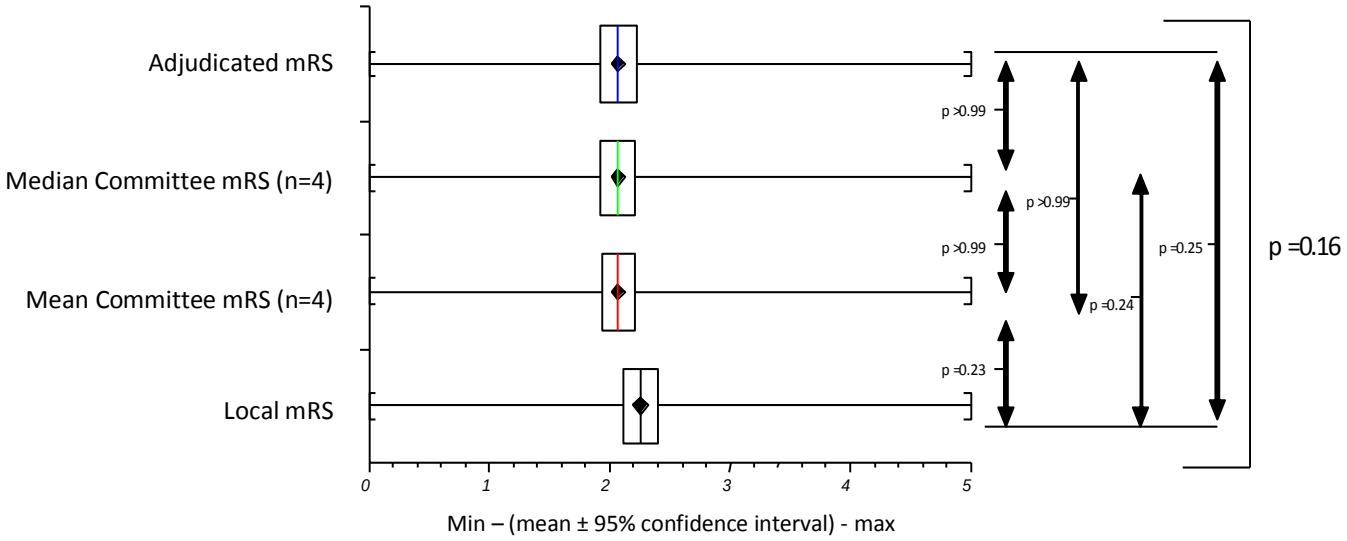


Figure 29 - mRS Distribution of Local and Committee Scores (Mean / 95% CI) p-values represent the Kruskal Wallis test of difference between distributions. n=538

6.3.2. Construct Validity

6.3.2.1. Spearman Rank Correlation

Initially we performed simple Spearman Rank correlation coefficients to determine any simple relationship among variables known to be associated with stroke outcome (bNIHSS, Blood Pressure, Blood Glucose and Home Time). There were similar results with local and adjudicated mRS outcomes with any variable. All were found to be significantly associated with mRs outcome; except SBP with 90 day adjudicated mRS which did not reach statistical significance (p=0.068). Table 27.

Table 27 - Spearman Rank correlation coefficients (p value) for bNIHSS, SBP and Glucose with each mRS outcome

mRS method	n	bNIHSS	SBP	Glucose
Day 30 Local mRS	280	0.508 (<0.0001)	0.140 (0.014)	0.205 (0.005)
Day 30 Adjudicated mRS	280	0.509 (<0.0001)	0.132 (0.027)	0.188 (0.003)
Day 90 Local mRS	258	0.508 (<0.0001)	0.140 (0.014)	0.205 (0.001)
Day 90 Adjudicated mRS	258	0.417 (<0.0001)	0.113 (0.068)	0.176 (0.006)

6.3.2.2. Unadjusted Proportional Odds Ordinal Logistic Regression

Unadjusted proportional odds logistic regression was then performed for each of the above predictors with each method of mRS outcome. Again bNIHSS was a strongly significant predictor of outcome with either method of assigning mRS. SBP and Blood Glucose were not found to be consistently significantly associated with outcome. There was no significant difference between local and adjudicated mRS scores. Table 28 and figure 30.

Table 28 - Unadjusted proportional odds logistic regression of relationship between bNIHSS, SBP and blood glucose with each method of mRS assessment.

Outcome	n	Odds Ratio (OR)	95% CI for OR	P (CMH test)
Baseline NIHSS				
Day 30 Local mRS	280	1.315	1.248 – 1.385	<0.0001
Day 30 Adjudicated mRS	280	1.343	1.267 - 1.423	<0.0001
Day 90 Local mRS	258	1.260	1.198 – 1.325	<0.0001
Day 90 Adjudicated mRS	258	1.275	1.204 – 1.350	<0.0001
Systolic Blood Pressure				
Day 30 Local mRS	280	1.009	1.002 – 1.016	0.670
Day 30 Adjudicated mRS	280	1.008	1.000 – 1.016	0.303
Day 90 Local mRS	258	1.008	1.000 – 1.016	0.558
Day 90 Adjudicated mRS	258	1.007	0.999 – 1.015	0.581
Blood Glucose				
Day 30 Local mRS	280	1.178	1.074 – 1.293	0.071
Day 30 Adjudicated mRS	280	1.169	1.059 – 1.289	0.005
Day 90 Local mRS	258	1.178	1.075 – 1.291	0.099
Day 90 Adjudicated mRS	258	1.105	1.005 – 1.216	0.143

6.3.2.3. Adjusted Proportional Odds Ordinal Logistic Regression

Repeat proportional odds logistic regression was then performed, adjusted for each baseline predictor. Again, only bNIHSS was consistently significantly associated with outcome and there was no significant difference between local or adjudicated mRS scores. The CMH test was not possible in the adjusted analysis because of the nature of the variables, there were too many values for this be calculated. Table 29 and Figure 31.

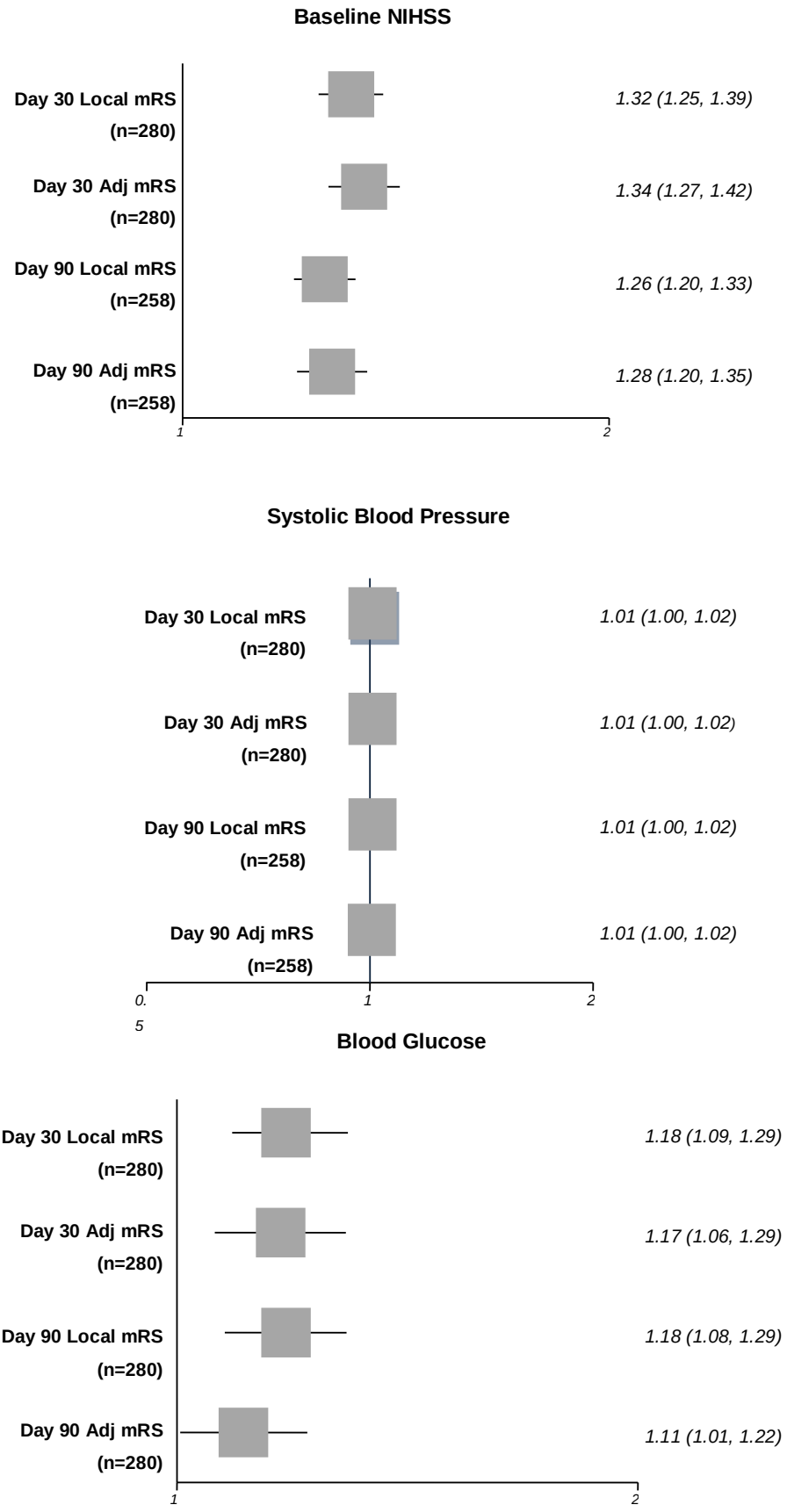


Figure 30 - Unadjusted proportional odds logistic regression of relationship between bNIHSS, SBP and Blood glucose with each method of mRS outcome (Odds Ratio and 95% CI)

Table 29 - Adjusted proportional odds logistic regression of relationship between bNIHSS, BP and blood glucose with each method of mRS assessment.

Outcome	Odds Ratio (OR)	95% CI for OR	P (CMH test)
Day 30 Local mRS (n=280)			
Baseline NIHSS	1.302	1.233 – 1.375	Not calculated
Systolic Blood Pressure	1.014	1.005 – 1.023	Not calculated
Blood Glucose	1.097	0.997 – 1.208	Not calculated
Day 30 Adjudicated mRS (n=280)			
Baseline NIHSS	1.336	1.257 – 1.420	Not calculated
Systolic Blood Pressure	1.014	1.005 – 1.024	Not calculated
Blood Glucose	1.125	1.017 – 1.245	Not calculated
Day 90 Local mRS (n=258)			
Baseline NIHSS	1.256	1.191 – 1.325	Not calculated
Systolic Blood Pressure	1.015	1.005 – 1.024	Not calculated
Blood Glucose	1.115	1.015 – 1.224	Not calculated
Day 90 Adjudicated mRS (n=258)			
Baseline NIHSS	1.276	1.201 – 1.355	Not calculated
Systolic Blood Pressure	1.015	1.005 – 1.026	Not calculated
Blood Glucose	1.058	0.959 – 1.167	Not calculated

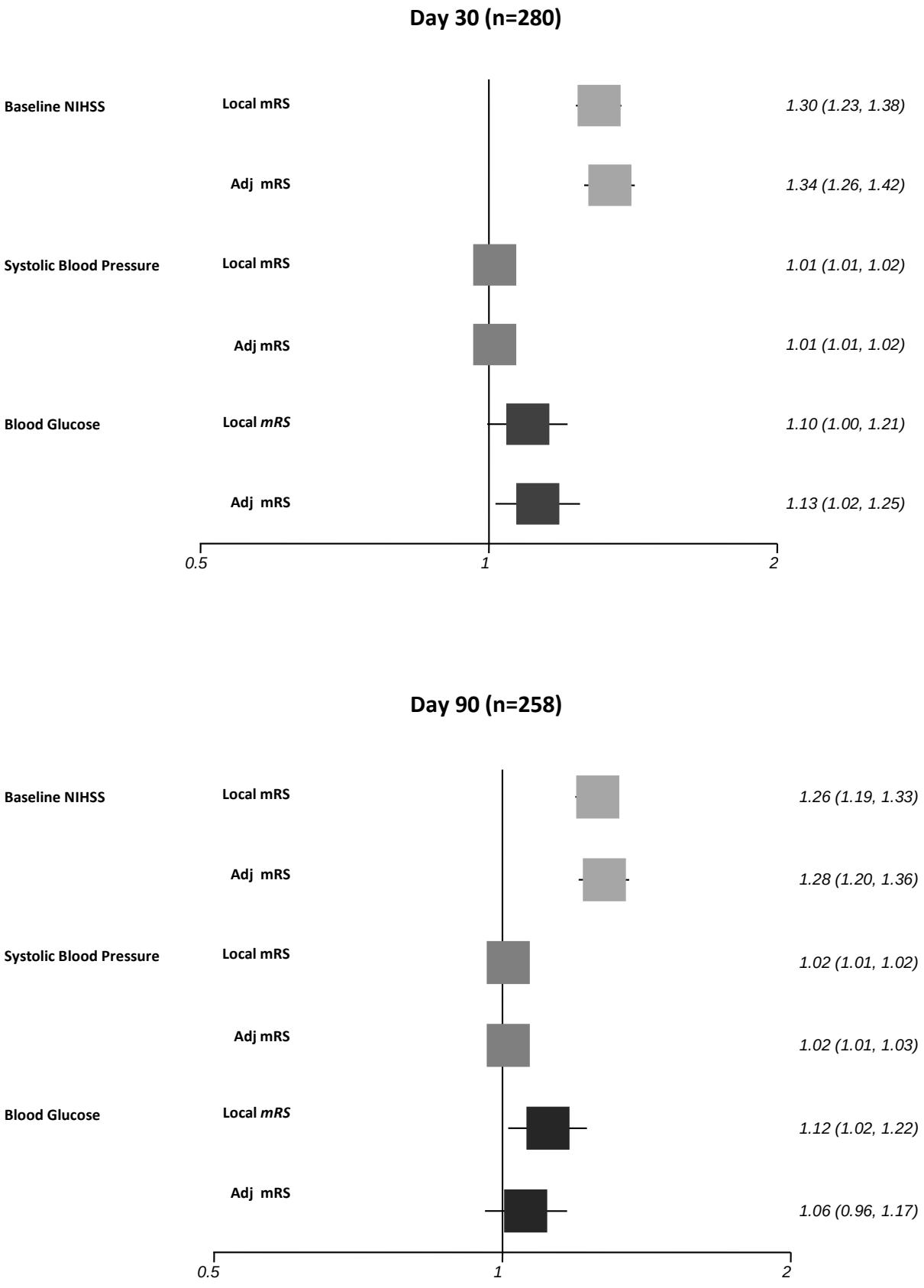


Figure 31 - Adjusted proportional odds logistic regression of the relationship between bNIHSS / SBP / blood glucose with each method of mRS assessment. Odds Ratio (95% CI)

6.3.2.4. Home Time

Home time is not a baseline predictor of outcome but is validated as a useful measure of functional outcome[181, 182]. We repeated the previous analysis using home time as an independent surrogate marker of functional outcome and then as a further predictor in the adjusted proportional odds model.

Initial Spearman Rank correlation coefficients confirmed the significant relationship between home time and functional outcome as measured using the mRS scale. There were no differences between local and adjudicated mRS scores. Table 30

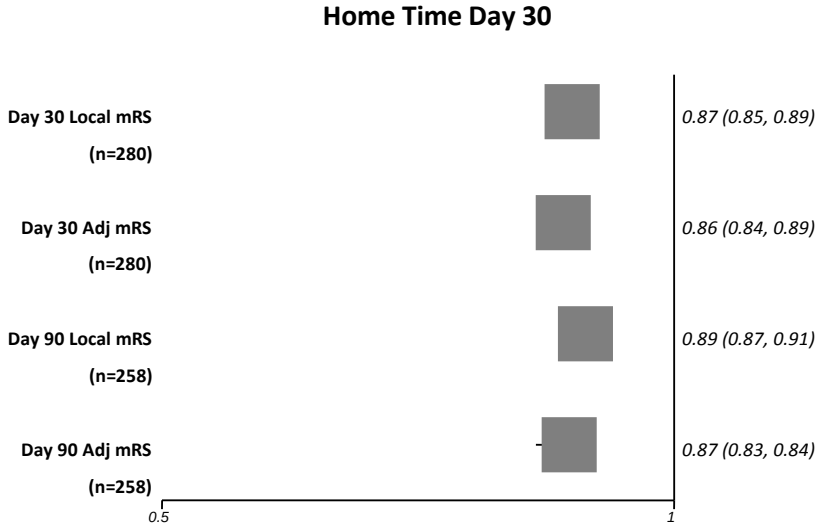
Table 30 - Spearman Rank correlation coefficients (p value) for Home Time with each mRS outcome

mRS method	n	Home Time (Day 30)	Home Time (Day 90)
Day 30 Local mRS	280	-0.585 (<0.0001)	-0.619 (<0.0001)
Day 30 Adjudicated mRS	280	-0.605 (<0.0001)	-0.642 (<0.0001)
Day 90 Local mRS	258	-0.547 (<0.0001)	-0.599 (<0.0001)
Day 90 Adjudicated mRS	258	-0.578 (<0.0001)	-0.607 (<0.0001)

Unadjusted proportional odds logistic regression was performed with home time at day 30 or day 90 as a predictor for mRS outcome. Again, home time was found to be a consistently significant predictor of mRS outcome and there was no significant difference between local and adjudicated mRS scores. Table 31 and figure 32

Table 31 - Unadjusted proportional odds logistic regression of relationship between Home Time and each method of mRS assessment.

Outcome	Odds Ratio (OR)	95% CI for OR	P (CMH test)
Home Time Day 30 (n=280)			
Day 30 Local mRS	0.871	0.850 – 0.893	<0.0001
Day 30 Adjudicated mRS	0.861	0.839 – 0.885	<0.0001
Day 90 Local mRS	0.887	0.866 – 0.909	<0.0001
Day 90 Adjudicated mRS	0.868	0.844 – 0.892	<0.0001
Home Time Day 90 (n=258)			
Day 30 Local mRS	0.933	0.922 – 0.944	<0.0001
Day 30 Adjudicated mRS	0.929	0.916 – 0.941	<0.0001
Day 90 Local mRS	0.947	0.938 – 0.957	<0.0001
Day 90 Adjudicated mRS	0.934	0.921 – 0.947	<0.0001



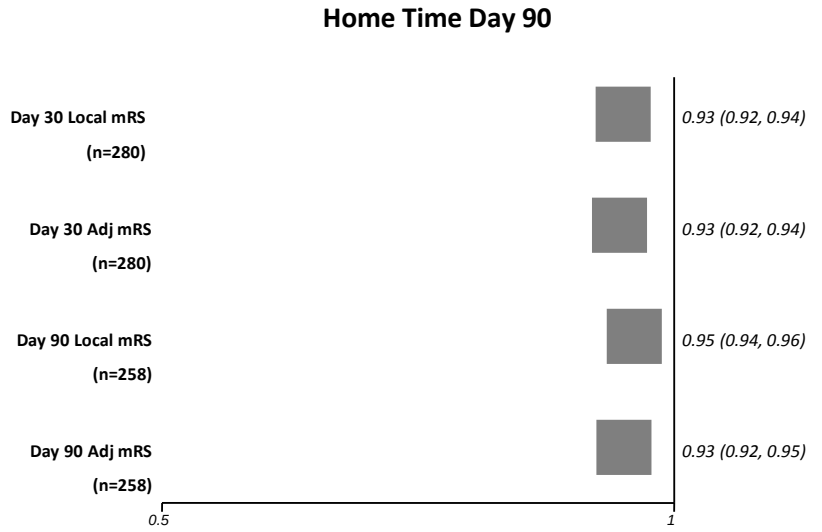


Figure 32 - Unadjusted proportional odds logistic regression of relationship between Home Time and each method of mRS assessment. Day 30 (n=280) and Day 90 (n=258). Odds Ratio (95% CI)

When included in the adjusted model with bNIHSS, SBP and Blood Glucose, only home time and bNIHSS remained significantly associated with mRS outcome. There was no change in the relationship using either local or adjudicated mRS scores. Day 90 home time: Table 32 and Figure 33. Results for Day 30 are shown in Appendix C (Table 41 and Figure 51)

Table 32 - Adjusted proportional odds logistic regression of relationship between bNIHSS, SBP, blood glucose and home time at 90 days with each method of mRS assessment. Day 30 and 90 mRS

	Odds Ratio (OR)	95% CI for OR	p
Day 30 Local mRS (n=280)			
Baseline NIHSS	1.151	1.085 – 1.222	<0.0001
Systolic Blood Pressure	1.007	0.999 – 1.015	0.096
Blood Glucose	1.059	0.959 – 1.169	0.257
Home Time Day 90	0.945	0.934 – 0.957	<0.0001
Day 30 Adjudicated mRS (n=280)			
Baseline NIHSS	1.174	1.101 – 1.252	<0.0001
Systolic Blood Pressure	1.006	0.997 – 1.014	0.207
Blood Glucose	1.084	0.977 – 1.204	0.130
Home Time Day 90	0.941	0.928 – 0.954	<0.0001
Day 90 Local mRS (n=258)			
Baseline NIHSS	1.104	1.040 – 1.172	0.001
Systolic Blood Pressure	1.008	1.000 – 1.017	0.052
Blood Glucose	1.094	0.995 – 1.203	0.062
Home Time Day 90	0.956	0.945 – 0.967	<0.0001
Day 90 Adjudicated mRS (n=258)			
Baseline NIHSS	1.113	1.042 – 1.190	0.002
Systolic Blood Pressure	1.007	0.998 – 1.016	0.123
Blood Glucose	1.044	0.943 – 1.155	0.408
Home Time Day 90	0.945	0.932 – 0.959	<0.0001

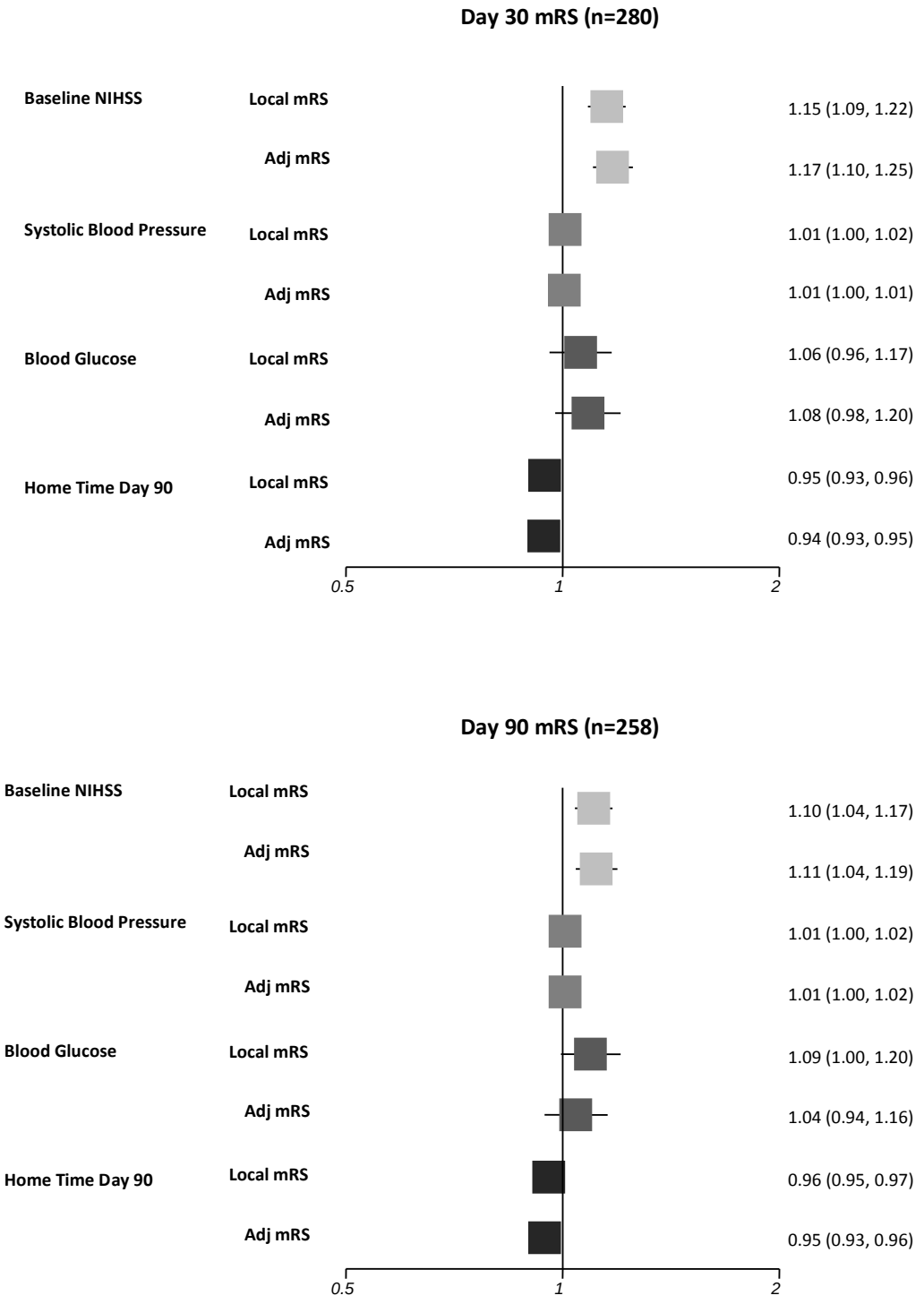


Figure 33 -Adjusted proportional odds logistic regression of relationship between bNIHSS, SBP, blood glucose and home time at 90 days with each method of mRS assessment. Day 30 (n=280) and 90 mRS (n=258)

6.4. Conclusions

We report sufficient agreement between local and adjudicated mRS scores to demonstrate validity of an adjudicated mRS outcome, without any significant or systematic bias. This validity has been demonstrated in relation to the local mRS assessment and has been demonstrated to correlate well with other independent factors known to affect stroke outcome.

There are limitations to this analysis. As there is no method of quantifying the "true" disability of a participant we are reliant on a scale to document this. Through use of endpoint committee review we hope to score the "true" mRS of the study participant. In the absence of a gold standard disability assessment we are unable to assess this directly. We are encouraged that there was a degree of disagreement between local and adjudicated mRS (suggesting that the adjudication process adds something to standard assessment); but this variability was not too large (which may suggest that adjudicated mRS is systematically different to traditional mRS).

Chapter 7

Factors associated with variability in mRS scoring

7.1. Introduction

Inter-observer variability in mRS has been apparent and documented since inception of the scale. As we have discussed, even with dedicated training and increasing familiarity with the scale this variability remains. The reasons for the high variability in clinical mRS scoring are likely to be multifactorial. Arguably vague definitions used to define each mRS grade allows for a degree of subjectivity in assessment. The inherently unstructured nature of the traditional mRS allows the investigator the freedom to explore the functional aspects that are relevant to each individual patient, without the need to concentrate on mobility or continence where these factors may not be relevant. Despite the perception that the application of mRS is more consistent at extremes of disability; because high and low mRS grades are better defined or as a reflection that variability can be in one direction only; we have found disagreement in assigning mRS grades at all levels of the scale.

There are potential factors associated with variability a) specific to the interviewer; b) specific to the interview subject and c) specific to the interview situation. If these sources of variability were better understood, one might be able to predict “problem” mRS cases and target interventions to improve reliability. We aimed to identify factors noted in our cohort of mRS assessment videos that might be associated with variability in mRS scores.

7.2. Methods

7.2.1. mRS scores and variability grading

All mRS videos scored by the endpoint committee in the CARS study were included. Those clips that did not have an adjudicated score were excluded. Table 19 in Section 4.2.8.1 describes the reasons for missing adjudicated scores. The majority were due to duplicate clips, incorrect file type, poor audio that precluded scoring or incomplete assessment.

Each clip was scored by a minimum of six independent assessors (local mRS, four adjudicated committee scores and a further independent mRS score from an assessor that did not participate in the original trial). All assessors were trained and certified in mRS. The final adjudicated mRS score was not included in the analysis as it was generated by the four committee scores and was therefore not an independent score.

Variability in mRS scores for each clip was graded by determining the number of mRS scores that agreed from the six mRS scores. (Range 2-6)

7.2.2. Identification of factors predictive of variability

For each of the mRS interviews studied, we collated descriptive data on the participant and the quality and content of the interview. The variables chosen for study were those domains thought to be factors potentially associated with mRS variability based upon the results of previous qualitative analysis[204]. Clips were reviewed by two assessors, noting variables felt to impact upon scoring; if either rater felt that the variables contributed to difficulty in scoring then they were included in analysis.

Patient specific variables included participant age, pre stroke mRS, baseline stroke severity as graded by baseline NIHSS (bNIHSS) and presence of language disorder. Interview specific variables included length of interview, poor sound quality, location of the interview, use of a proxy or discussion of prior disability.

Variables included were continuous and categorical. For continuous variables (such as interview length or participant age) all clips were included as a potential factor associated

with variability. For each categorical variable a standard was identified as the most common factor in the cohort of videos. From this standard, analysis was based upon whether the less frequent variable was associated with variability in scoring. For location of interview, the majority of clips were recorded in an outpatient clinic setting; there were a minority of clips recorded in either the participant's home or in an inpatient hospital setting (most commonly a rehabilitation facility). The majority of clips did not involve a proxy. In describing the use of a proxy a minority of clips used a proxy to provide the entire interview (e.g. a nurse or other caregiver) and some included both the participant and a proxy in the interview (e.g. the participant together with their relative or caregiver). Most clips did not include discussion of prior disability; in those that did this was noted as due to either prior stroke or to other comorbidity (e.g. arthritis / lung disease).

7.2.3. Statistical Analysis

Each of the variables were input as factors potentially associated with variability in mRS scoring using ordinal logistic regression with the proportional odds model (see section 6.2.2.1). The dependant variable was mRS variability grade. Odds ratios with corresponding 95% confidence intervals and p values generated by the Cochrane-Mantel-Haenszel (CMH) test are reported. Day 30 and day 90 scores were analysed independently as they represent repeated measures on each subject.

7.3. Results

538 video clips were included in the analysis (Day 30: 281 and Day 90: 257). Variability grading for video clips ranged from 2 (2 scores of 6 in agreement) to 6 (all scores in agreement). 21% of clips at each study visit noted full agreement in mRS scores (59/281, 21% at 30 days and 56/257, 21.8% at 90 days). The distribution of variability grading was similar at each visit. Table 33

Table 33 - Variability rating for videos at 30 and 90 days

Variability Rating	Day 30 N=280 (%)	Day 90 N=258
2	3 (1.0%)	6 (2.3%)
3	68 (24.2%)	45 (17.5%)
4	80 (28.5%)	76 (29.6%)
5	71 (25.3%)	74 (28.8%)
6	59 (21.0%)	56 (21.8%)

Continuous variables included in analysis were similar at both 30 and 90 days. There was some variation in the population of participants included at 30 and 90 days due to missed visits and missing videos. There were no statistically significant differences between the two groups for any variable. Table 34. Categorical variables were similarly represented in the day 30 and day 90 groups. In some cases there were very small group numbers for categorical variables (e.g. poor sound quality, full proxy interview and discussion of prior disability due to stroke). Table 35.

Table 34 - Continuous variables as predictors of scoring variability in proportional odds logistic regression model

Continuous Variables	Day 30 N=280			Day 90 N=258			
	Range	Median (IQR)	Mean (SD)	Range	Median (IQR)	Mean (SD)	p
Age	22-99	69 (59-77)	67.3 (13.2)	22-99	69 (60-76)	67.5 (12.4)	0.691
Pre-stroke mRS	0-4	0 (0-1)	0.4 (0.8)	0-3	0 (0-0)	0.3 (0.7)	0.128
Baseline NIHSS	0-23	4 (2-7)	5.6 (4.8)	0-22	4 (2-7)	5.4 (4.4)	0.560
Interview Length (min:sec)	00:52-23:51	5:08 (3:19-7:16)	5:45 (3:20)	1:00–22:39	4:30 (3:18-6:54)	5:15 (3:19)	0.066

Table 35 - Categorical variables as predictors of scoring variability in proportional odds logistic regression model

Categorical Variables		Day 30 N=280	Day 90 N=258
		N (%)	N (%)
Location of Interview (Standard: Outpatient Clinic)	Participants home	40 (14.2%)	48 (18.7%)
	Inpatient / Rehab	48 (17.1%)	15 (5.8%)
Poor Sound Quality (Standard: No)	Yes	11 (3.9%)	9 (3.5%)
Presence of Language disorder (Standard: No)	Yes	89 (31.7%)	56 (21.8%)
Use of Proxy (Standard: No Proxy)	Yes – full interview	12 (4.3%)	8 (3.1%)
	Yes – participant & proxy	45 (16.0%)	40 (15.5%)
Pre-stroke disability (Standard: None)	Yes – previous stroke	8 (2.8%)	6 (2.3%)
	Yes – other co-morbidity	62 (22.1%)	47 (18.3%)

Prior disability was poorly reflected in the pre-stroke mRS score. 22.9% (123/538) videos discussed prior disability; only 1.6% (9/538) of videos had a pre-stroke mRS of ≥ 3.

At both 30 and 90 days only “interview length” was significantly associated with agreement in mRS scoring. Tables 36 and 37. Baseline NIHSS at 30 days was inversely related to variability in mRS scoring, however this did not reach statistical significance using the CMH significance test. The small numbers in some groups limited analysis in some categorical variables and the regression model provided very wide confidence intervals. In plotting the data groups with n ≤ 15 have been removed. Figures 34 and 35.

Table 36 - Factors associated with variability in mRS scoring at 30 days. Frequency of variable and odds ratio (95% CI) from proportional odds logistic regression. P value generated from CMH test.

Day 30 (n=280)	N	Odds Ratio (95% CI)	CMH p value
Categorical Variables			
Prior disability – stroke	8	1.44 (0.41-5.12)	0.448
Prior disability – other comorbidity	62	1.69 (1.01-2.82)	0.448
Proxy – full interview	12	0.16 (0.05-0.52)	0.119
Proxy – participant and proxy	45	2.11 (1.18-3.79)	0.119
Language Disorder	89	0.62 (0.40-0.98)	0.139
Poor sound quality	11	1.59 (0.54-4.69)	0.149
Location – inpatient/rehab	48	0.35 (0.20-0.63)	0.238
Location – patient's home	40	1.37 (0.74-2.56)	0.238
Continuous Variables			
Interview length	281	1.12 (1.05-1.19)	0.023
Age	281	1.01 (0.99-1.02)	0.913
Baseline NIHSS	281	0.95 (0.91-0.99)	0.263
Pre Stroke mRS	281	0.90 (0.68-1.12)	0.380

Table 37 – Factors associated with variability in mRS scoring at 90 days. Frequency of variable and odds ratio (95% CI) from proportional odds logistic regression. P value generated from CMH test.

Day 90 (n=258)	N	Odds Ratio (95% CI)	CMH p value
Categorical Variables			
Prior disability – stroke	6	3.30 (0.76-14.47)	0.146
Prior disability – other comorbidity	47	1.78 (0.99-3.16)	0.146
Proxy – full interview	8	1.93 (0.50-7.49)	0.072
Proxy – participant and proxy	40	1.72 (0.93-3.18)	0.072
Language Disorder	56	0.94 (0.55-1.61)	0.360
Poor sound quality	9	0.32 (0.09-1.06)	0.002
Location – inpatient/rehab	15	1.14 (0.44-2.91)	0.876
Location – patient's home	48	1.17 (0.66-2.07)	0.876
Continuous Variables			
Interview length	257	1.13 (1.06-1.21)	0.001
Age	257	1.01 (0.98-1.02)	0.511
Baseline NIHSS	257	1.01 (0.96-1.06)	0.683
Pre Stroke mRS	257	1.15 (0.84-1.56)	0.189

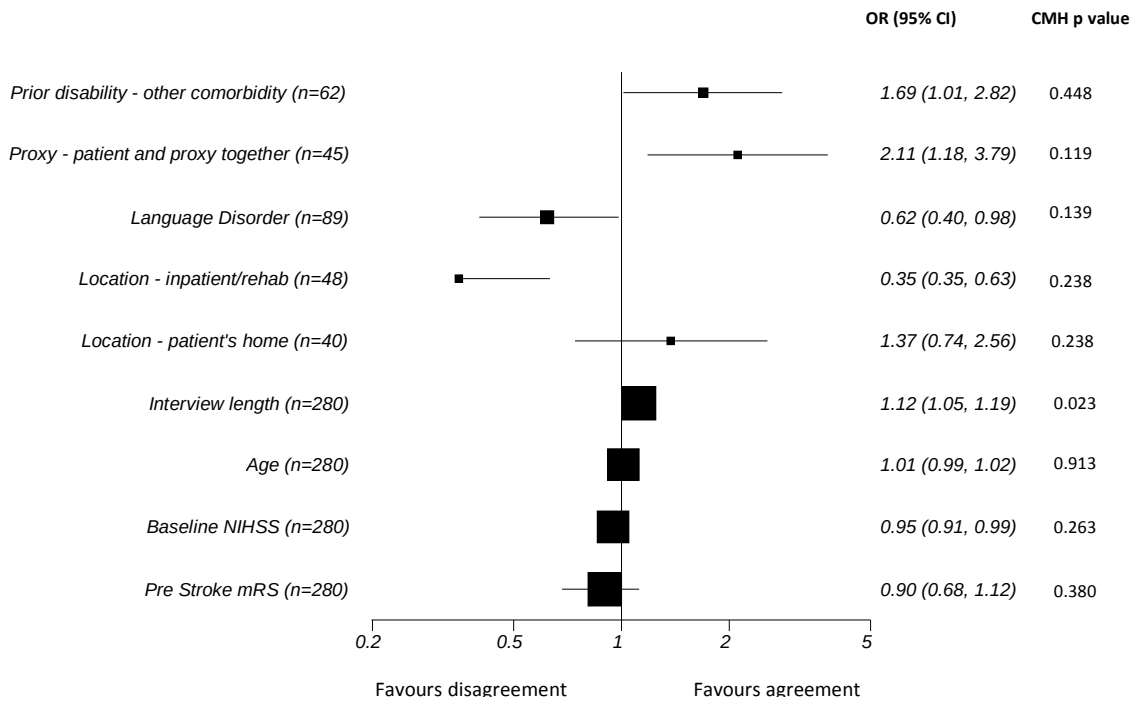


Figure 34- Forest Plot: Factors associated with variability in mRS scoring at 30 days. Odds ratio (95% CI) and CMH p value.

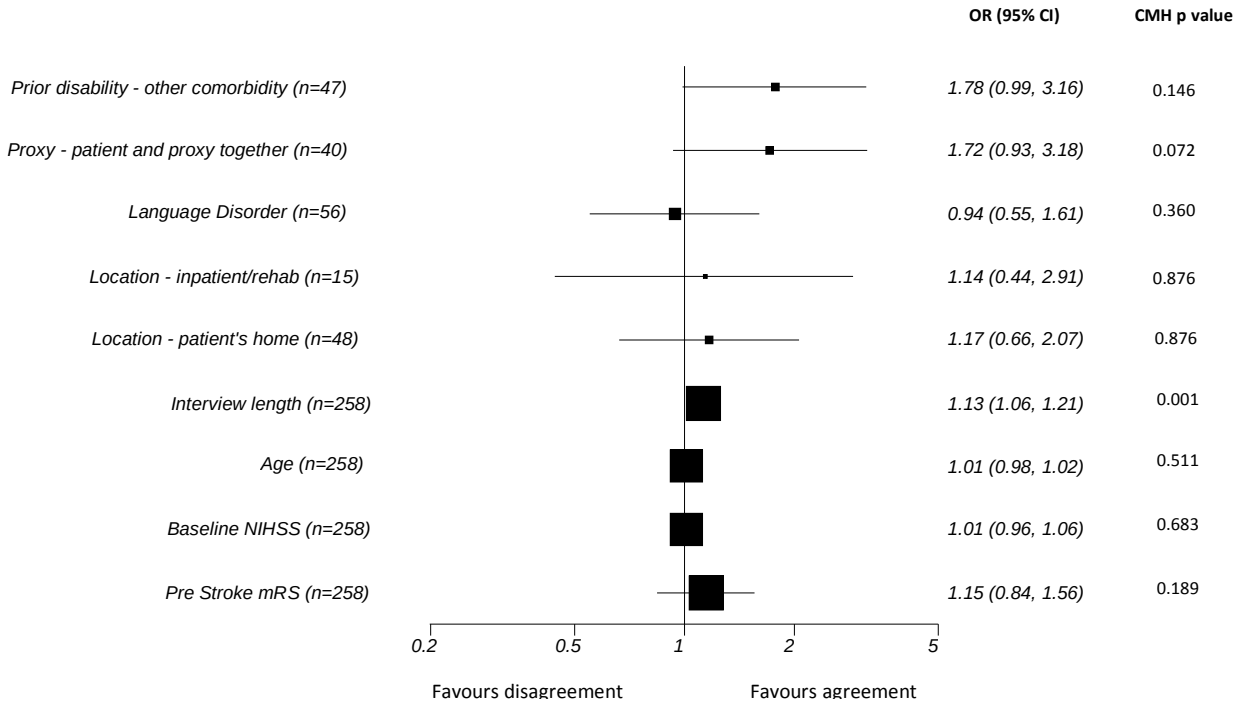


Figure 35 - Forest Plot: Factors associated with variability in mRS scoring at 90 days. Odds Ratio (95% CI) and CMH p value

7.4. Conclusions

As we have previously described, inter-observer variability limits the utility of mRS scoring. There are several proposed strategies to minimise this but tools to predict potential "problem" mRS cases might allow targeted interventions to minimise misclassification. We aimed to assess which factors in the mRS video interview affect variability in scoring to identify the features of an optimal adjudication video.

In this study, the selection of factors potentially associated with variability was arbitrary by necessity due to a paucity of published background literature on predictors of mRS variability. Our selection of variables was guided by prior qualitative study[204]. Other factors that might have been relevant such as employment status, socio-economic status, marital status etc. may be pertinent but the corresponding data were not captured at the time of initial interview.

The majority of factors identified as possible factors associated with variability were not demonstrated as such in this study. This is in part attributable to the small numbers in some groups. The use of a proxy, presence of language disorder or location of the interview does not appear to impact upon scoring variability. There is a perception that at the extremes of disability there is less variability in mRS scoring; however measures of initial stroke severity were not associated with improved reliability in this study.

We were only able to identify one factor associated with agreement in mRS scoring. Increasing length of interview was associated with less variability. This was unexpected in light of the experience of end point committee discussions. In several instances adjudication committee members commented that longer recorded interviews were difficult to follow and contained unnecessary information. However, on the basis of this data we might surmise that extra information, which may be interpreted as superfluous, could be useful in forming an mRS judgement.

The clinical implications of these data are uncertain, although duration of mRS interview may be associated with variability the effect seems small and by its nature duration of interview

cannot be predetermined or assessed until the interview is complete. Perhaps we should conclude that interviewers should not deliberately prolong the interview but should take as much time as is needed to cover all the salient points of a complete mRS assessment.

7.4.1. Scoring controversies in CARS study

During the course of the CARS study the endpoint committee met to discuss all mRS videos that had resulted in disagreement. This permitted focussed discussion to highlight potential areas that may contribute to mRs uncertainty. There were common themes discussed during these meetings that cause difficulty in mRS grading. In several instances the controversies arose simply because of a value judgement and were individual to each interview. Factors that frequently generated discussion were with regards to possible non stroke symptoms or incomplete information. We did not formally study the details of these conversations in a qualitative manner; however, it is worth brief discussion of these controversies as potential topics in the content of a clip that might impact upon mRS variability in. Across each mRS boundary there were recurring themes.

Post stroke symptoms (mRS 1 boundary): A wide array of post stroke symptoms and syndromes are recognised; those which are easily recognisable such as paralysis or dysphasia and others which are less objective such as post stroke pain, fatigue or mood disorder. There is a clear distinction between interpretation of physical and psychological symptoms after stroke, largely related to the objective / subjective nature of their impact. The interpretation of more subjective symptoms (such as fatigue or mood disorder) may vary more among individuals and the link between them and functional ability may be less clearly related to the initial stroke insult or any response to treatment.

Should all of these symptoms be regarded as equal in terms of the limitations that they place on the stroke survivor? This is not clear in mRS scoring guidance; an inability to perform ADLs due to hemiplegia is noticeably different to an inability to perform ADLs due to fatigue. In some societies it must be acknowledged that the extent and impact of symptoms can have consequences in terms of physical and financial support or welfare. These influences are difficult to extricate in accurate functional assessment for clinical trial purposes, particularly

where a participant knows that a recording of their description of disability is going to be stored. It is important that the nature and effect of all symptoms is fully explored, acknowledging the limits and possible motivations in these discussions.

Return to previous activities (mRS 2 boundary): For the purposes of the mRS scale, a regular activity is defined as one which was carried out at least once monthly in the time period immediately prior to stroke. Controversy may arise where a participant may be able to perform an activity, but they feel that they perform less well than prior to the stroke. An example might be a participant who had a complex hobby prior to their stroke (playing an instrument / painting); if their music or art is perceived to be of poorer quality after stroke should this be deemed a loss of activity? Where an individual is able to perform physical tasks to a degree that allows resumption of normal activities, then perhaps caution should be exercised before giving a score of mRS 2.

In some situations participants may not have had the requirement or opportunity to attempt such activities. Such a situation arises when participants are scored during inpatient rehabilitation. Our prejudice is likely to be that during an inpatient stay, where assistance with personal care, meals and domestic tasks is provided; a participant must mandate a score of mRS 3. Length of stay in rehabilitation facilities and the activities that are expected of patients during their stay vary widely across geographical areas in relation to healthcare systems, facilities and funding. On this basis it is impossible to offer international guidance on how to score this population except to highlight that consideration should given to ensure that the assistance received is absolutely required.

Societal barriers to functional recovery: The primary issue here arose in discussions regarding a participant's return to driving. Where participants have not returned to an activity due to a legal constraint it is unclear if this should be scored as a disability. Similar issues arise with return to work, caring roles and other societal roles. As the CARS study was performed across sites subject to similar laws and regulations this was fairly easy to deal with. Agreement was reached that where a legal constraint regarding driving was the only limitation with an otherwise good recovery, this must not be scored as a disability. There was an acknowledgement that where a central adjudication model might be adopted on an

international basis there are likely to be other "societal barriers" to full, functional recovery and these must each be considered on their own merits.

Definition of Requiring Assistance (mRs 3 Boundary): Activities of daily living (ADL) can be grouped into two categories of basic and instrumental ADL's. Basic ADL's include self care tasks such as personal hygiene, dressing, eating, transfers, elimination and mobility. If help with these is required then a score of at least 3 should be assigned. Instrumental ADL's are not necessary for fundamental functioning, but allow an individual live independently in a community. These are again of equal importance but given that they are more complex, the range of activity is great. Examples of such activities are doing light housework, preparing meals, taking medications, shopping for groceries or clothes, using the telephone, managing money. If a patient is able to perform each of these activities in a basic form then they would not warrant a score of 3. An example might be a patient who is able to visit the local shop daily for a small selection of groceries but couldn't manage a large supermarket shop. This should not in isolation warrant a score of 3. The ability to complete the task in question at a simple level is the crucial point. In this circumstance, if minor modification within their physical capabilities could obviate need for such help (e.g. more frequent smaller shopping trips) this may not warrant a score of 3.

What constitutes independent mobility? (mRS 4 boundary): The mRS scoring guidance is clear that where a participant can mobilise without the help of another person they are considered independent. The use of walking aids is permissible but it must be clear that the participant can transfer independently and use the aid without the assistance of another person. However, in our cohort we discussed a few examples where this description was not clear cut. The use of a wheelchair is a complex issue. Many would consider this to be an exclusion to "mobility", but where a participant is able to move freely and independently using this aid (similar to the use of a cane or walking frame) they might be considered mobile. This is a particular issue where a participant used a wheelchair prior to stroke; often for reasons not attributable to stroke disease.

What constitutes "bedbound" or "constant nursing care"? (mRS 5 boundary): The mRS scoring guidance for mRS 5 requires a participant to be bedbound, incontinent and requiring

constant nursing care. If any one of these criteria is met it would be reasonable to consider a score of mRS 5. However, we encountered several situations where these descriptions became less clear. If a patient has been hoisted from bed to sit in a chair with support, does this remove them from the category of bedbound? If a participant is physically able but has cognitive impairment following stroke that requires constant supervision could this scored as mRS 4 or 5 despite mobility?

These factors raise questions about the mRS scoring guidance that require thought and discussion amongst the stroke community. Although there are clear uncertainties, these are less important where there is consistency in scoring. By describing common themes relating to difficulties in grading mRS, other mRS assessors might be able to incorporate this into their future assessments. These topics might contribute in future to extended guidance on mRS scoring in clinical trials; for example paying particular attention to local regulations concerning driving, return to work; or taking account of pre-morbid disability during interview.

In conclusion we have not convincingly demonstrated factors associated withmRS variability. Difficulties in mRS grading appear to be specific to the patient, assessor and interview. Difficult mRS cases cannot be prospectively identified on the basis of this work indicating that strategies to improve reliability should be applied universally. However, we acknowledge the preliminary nature of this exercise and further prospective studies may better inform the debate.

Chapter 8

Translation of mRS assessments: Validity, reliability and feasibility of incorporation in the central adjudication model.

8.1. Introduction

Stroke is a leading cause of death and disability worldwide. For acute stroke trials to be generalisable to an international population it is important that participants are enrolled from geographically, culturally and genetically diverse backgrounds. Thus, contemporary randomised controlled trials in stroke are international, multicultural and multilingual.

Including patient reported outcomes is a challenge in international trials. Participant responses must be documented in a way that is standardised, repeatable and reliable. Researchers must consider in their study design how they will address language barriers and consider the use of translators and interpreters in the collection of data.

The use of translated materials in international medical research is accepted and the methodological difficulties posed are acknowledged. There is a body of literature to offer

guidance on the translation of standardised tools used in research of patient reported outcomes. However, this guidance pertains to the use of structured assessment tools with a fixed query : response model, such as patient questionnaires or quality of life measures. There is little guidance in the use of translation with qualitative or semi-structured tools such as patient interviews as an outcome measure. Translation with the aim of achieving direct lexical equivalence and that with the aim of achieving cultural and conceptual equivalence may be quite different, further complicated by the communication difficulties that are common after stroke.

Social, cultural and linguistic factors may affect perception of disability after stroke which in turn may influence mRS scoring. For central adjudication of mRS outcome assessments to be successful in this context it is important to consider the challenges that may be posed in achieving culturally sensitive translation; how to address these challenges in the central adjudication model described whilst ensuring that validity and reliability are maintained.

The aims of our translation study were twofold. Firstly we aimed to determine the validity and reliability of translated mRS assessments in collaboration with a team in Beijing, China. We then aimed to assess the feasibility of incorporating a translation step into the central adjudication model and evaluate the reliability of these translated assessments.

8.2. Methods

8.2.1. mRS translation pilot study

Our pilot study was conducted in collaboration with a team of stroke researchers from the Department of Neurology, Peking University First Hospital, Beijing, China. Ten assessors (5 Glasgow, UK and 5 Beijing, China), trained and certified in the use of mRS, scored digitally recorded mRS assessments of consenting patients from each site. UK Ethical approval was granted by Scotland A Research Ethics Committee. mRS videos were scored in English and Mandarin, with each assessor working in his or her native language and after translation. Translations were provided by written transcript which accompanied the mRS video file. Both native language and translated versions were scored by the respective teams.

A convenience sample of mRS clips in English was selected from the CARS cohort of video mRS assessments. A further convenience sample of mRS clips in Mandarin was filmed in China specifically for this study. The translation pilot study was undertaken in two parts. In an initial sample, to assess the impact of language disorder, two versions of the translated transcript were prepared (dual translation); one with the input of an mRS certified clinician and one by a linguist with no medical background. The translated mRS interviews were scored twice, using each transcript in turn at least 2 months apart. In the second phase a larger sample of mandarin clips was translated and subsequently scored to assess inter-observer variability with a larger sample size.

8.2.2. CARS translation sub study

Using a sample of the original CARS mRS video clips we assessed the feasibility and reliability of using translated mRS assessments in the central adjudication model. Ethical approval for this study extension was granted by Scotland A Research Ethics Committee and Essex 2 Research Ethics committee. Trained and certified mRS assessors from the original CARS investigator team and end point committee were involved in both translation and scoring roles. Training in the translation procedures was provided face to face or by telephone depending upon the distance involved and was supplemented by written information (see appendix B)

We used study sites in Scotland and in England/Wales to represent distinct geographical areas. There was no true language barrier in scoring clips from each area; however they were used to signify different countries in the CARS web portal model. All clips were in the English language but there were local variations in accent and colloquialisms evident. A sample of clips to meet certain criteria was selected from the cohort of CARS videos using R statistical software. The criteria used to limit the CARS sample were as follows: 90 day assessments only (to ensure no duplication of participant videos), equal number from Scotland : England/Wales and equal number of classified : misclassified clips, maximum clip length 10 minutes. Using these criteria clips were randomly selected from the CARS video cohort using R statistical software.

The sample of translation clips was allocated to the trial outcome manager through the web portal to appear as a newly uploaded video file. The origin of the file was identified in the mRS upload list. The file was checked for quality by the outcomes manager before being allocated to an investigator in a "translator" role who was based in the alternative geographical area (i.e. Scottish clips were "translated" by English investigators and vice versa). The investigator allocated as "translator" was notified by automated email that a clip had been allocated for translation; from here they were able to log in to the web portal to view the clip. Translations were provided using a digital dictation device (Phillips Digital Pocket Memo© and SpeechExec© software) in mp3 format. Translators were advised to provide their translations in plain, clear English without the use of colloquialisms where possible. They were also requested to provide clear and unambiguous statements of what was said in the clip by both the assessor and the participant. The dictated mp3 file was uploaded to the web portal by USB connection and on receipt it was automatically merged with the original .wmv video file. This provided a new file with the original video but replaced audio component (a video dubbed with the translation file). The translated file was again checked by the trial outcome manager for quality and verified as ready for committee review. In the translation sub study the end point committee was broadened beyond that used in the main CARS trial to include several investigators from other sites. Four committee members from both Scotland and England/Wales were then selected to review and score the translated clip. Each endpoint committee member was notified by automated email that there was a translated clip ready for review. Endpoint committee members based at the centre from which each clip originated were excluded from final review.

Translated clips were scored by four endpoint committee members not less than one year after first review of the original clip. The original four endpoint committee scores were then compared to the four translated endpoint committee scores to assess validity and reliability.

8.2.3. Statistical Analysis

Validity of translated mRS assessments was assessed by analysis for any systematic differences in the distribution of native language and translated mRS scores using the Kruskal Wallis test of distributions for non parametric data. Reliability of translated mRs assessments was assessed using kappa (κ), weighted kappa (κ_w) and ICC statistics.

All analyses were undertaken using Statsdirect statistical software.

8.3. Results

8.3.1. mRS translation pilot study

Sixty nine mRS clips were scored (9 English and 60 mandarin). Twenty mRS clips underwent dual translation (9 English and 11 Mandarin). . Median mRS score was 3 (IQR 2-4). There was no significant or systematic bias in native or translated mRS scores. The distribution of native language and translated scores (mean (95% confidence interval) and median (IQR)) are displayed in figures 36 to 41. There was no significant or systematic difference between all clips scored by native language scorers and all translated assessments (p=0.896) or in the groups that underwent dual translation (medical translation p=0.999; linguist translation p=0.999).

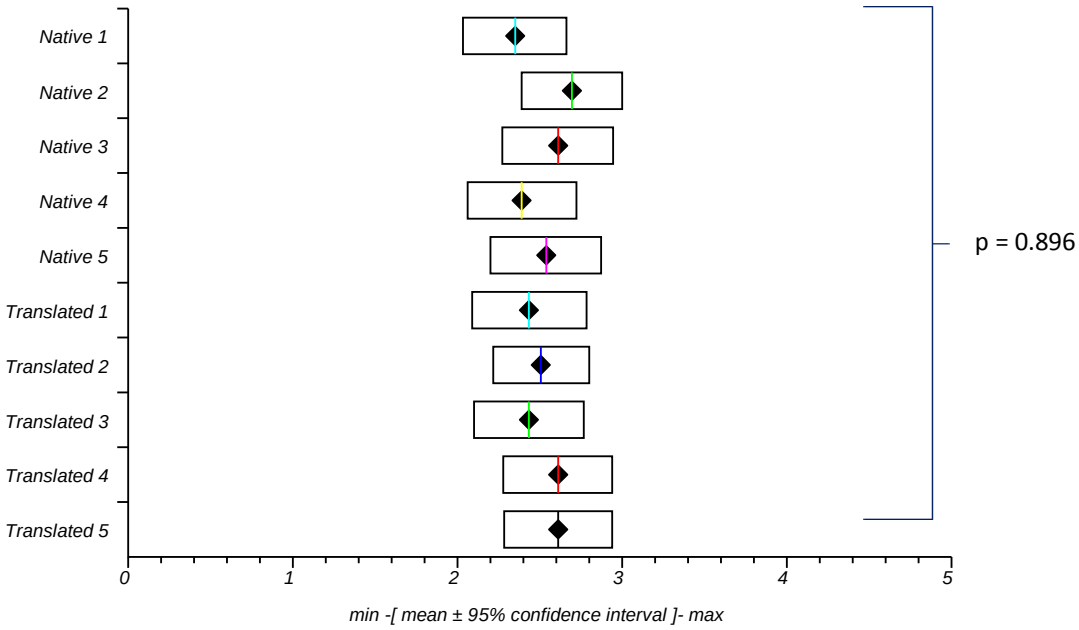


Figure 36 - Distribution of mRS scores in native language and all translated clips (n=69). Mean ± 95% Confidence Interval. p value represents the Kruskal Wallis test of difference between distributions.

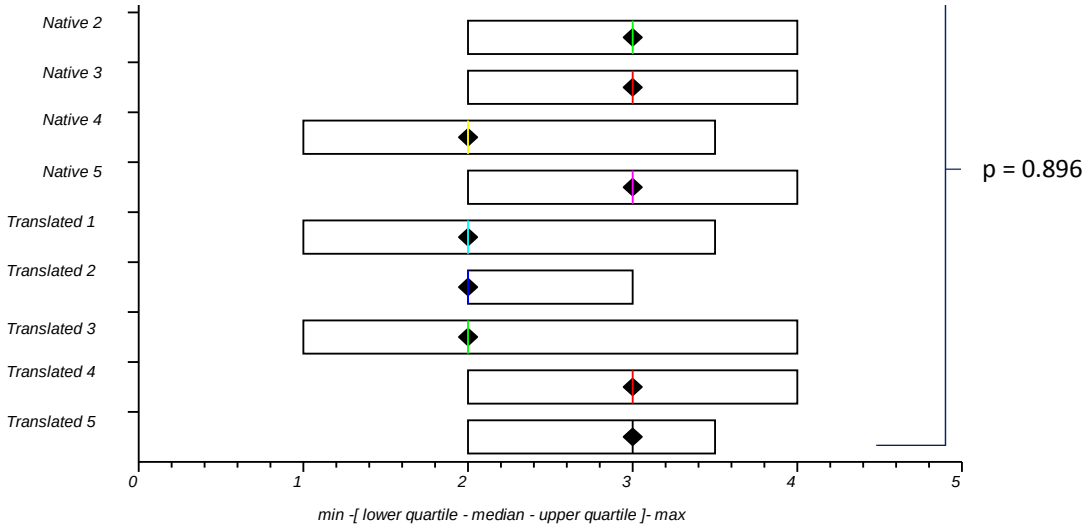


Figure 37 - Distribution of mRS scores in native language and all translated clips (n=69). Median ± IQR, Range. P value represents the Kruskal Wallis test of difference between distributions.

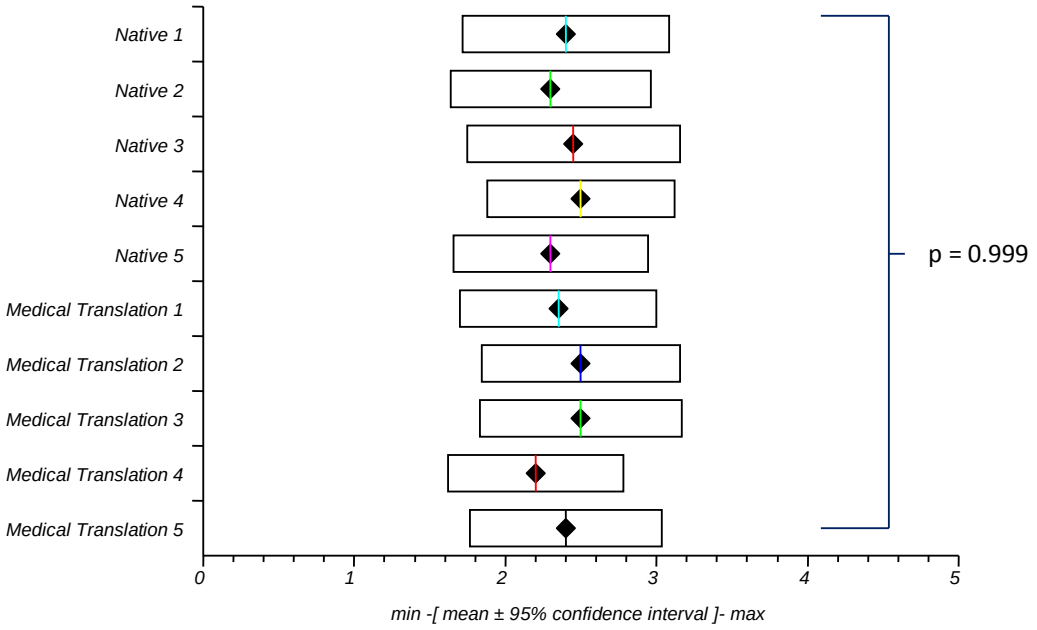


Figure 38 - Distribution of mRS scores in native language and medical translated clips (n=20). Mean ± 95% Confidence Interval. P value represents the Kruskal Wallis test of difference between distributions.

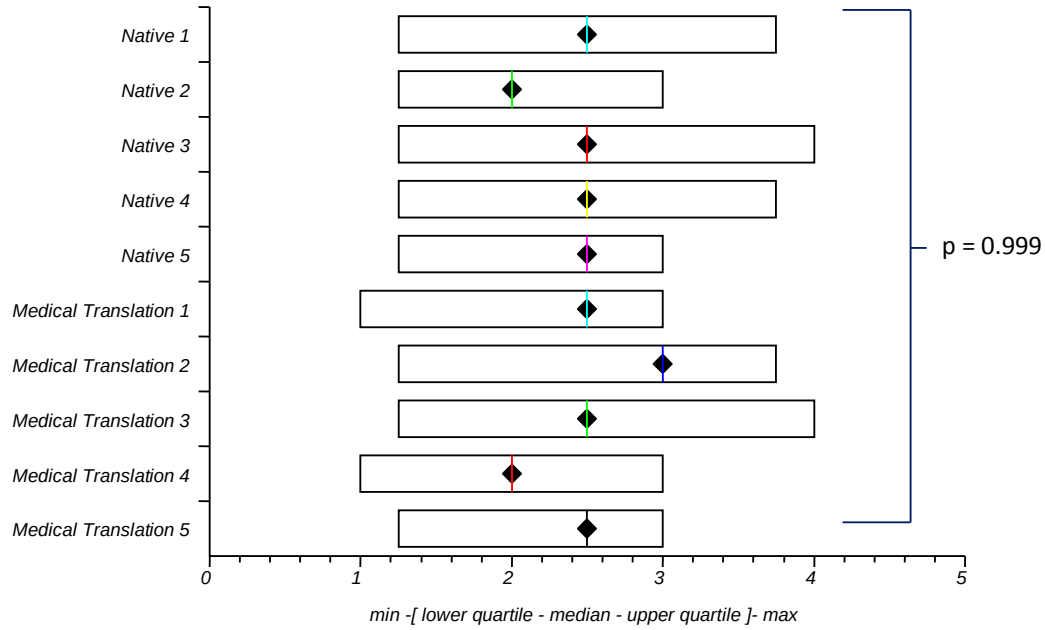


Figure 39 - Distribution of mRS scores in native language and medical translated clips (n=20). Median ± IQR. P value represents the Kruskal Wallis test of difference between distributions.

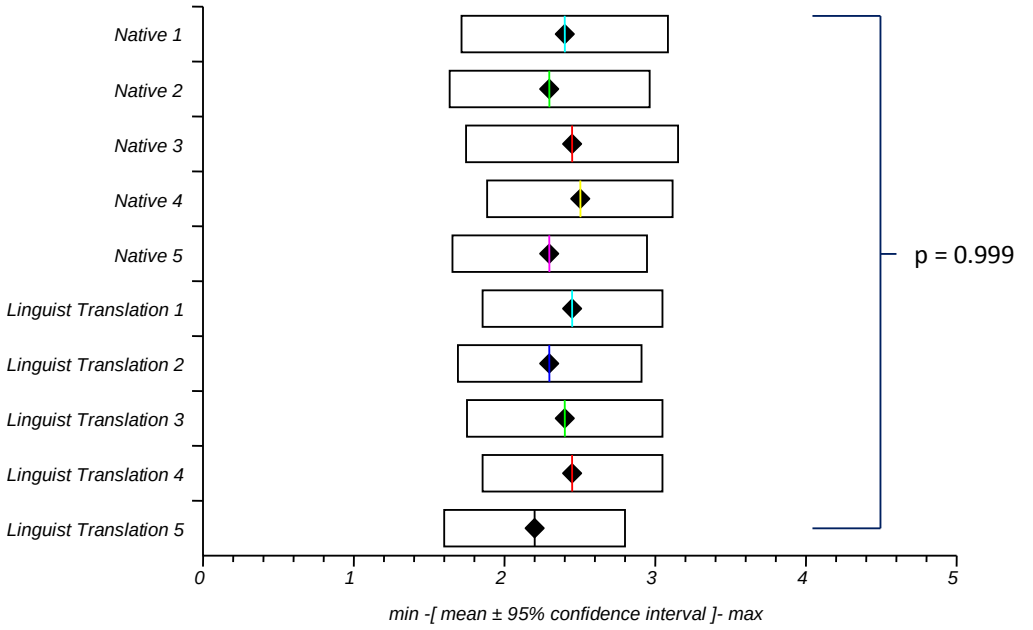


Figure 40 - Distribution of mRS scores in native language and linguist translated clips (n=20). Mean ± 95% Confidence Interval. P value represents the Kruskal Wallis test of difference between distributions.

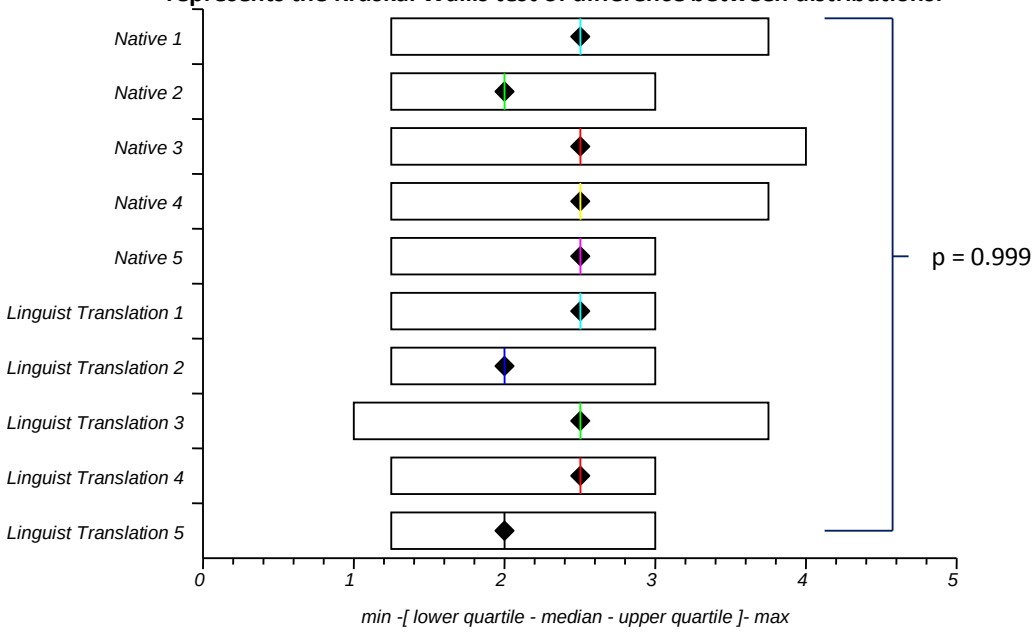


Figure 41 - Distribution of mRS scores in native language and linguist translated clips (n=20). Median ± IQR. P value represents the Kruskal Wallis test of difference between distributions.

Inter-observer reliability for native language assessment was good (n=69), κ 0.61 (95% CI 0.59-0.64), κ_w 0.91 (95% CI 0.86-0.96). ICC 0.91. Translated mRS assessments maintained good reliability (n=69), κ 0.59 (95% CI 0.55-0.62), κ_w 0.89 (95% CI 0.82-0.97), ICC 0.89. Putting an mRS trained clinician in the translation role had no demonstrable impact on the reliability of translated mRS assessments; κ 0.67 (95% CI 0.62-0.72), κ_w 0.91 (95% CI 0.81-1.01), ICC 0.91 with medical input (n=20) and κ 0.64 (95% CI 0.59-0.69), κ_w 0.91 (95% CI 0.82-1.01), ICC0.91 with linguist only transcription (n=20). Table 38 and Figure 44.

8.3.2. CARS translation sub study

Sixty mRS assessments were selected for inclusion in the CARS translation sub study (n=30 Scotland, n=30 England/Wales). Six investigators were allocated the role of "translator" (n=3 Scotland, n=3 England/Wales), each providing a modified audio summary for ten mRS video clips. Fourteen investigators were involved in providing committee scores for the modified mRS clips (n=11 Scotland, n=3 England/Wales). The translated videos were limited to include those less than 10 minutes long, range 1min 4 secs to 9mins 30secs (Mean (SD) 4mins 27secs (2mins 18secs))

All "translation" audio files were uploaded and merged with the video file successfully through the CARS translation web portal. There were no technical failures.

Median mRS in native language clips was 1.5 (IQR 1 – 2.5). Median mRS in translated mRS clips was 2 (IQR 1-3). Using the median mRS there was a trend towards lower mRS scores in the native language clips driven by two of four scores (Figure 43) but this did not result in any significant or systematic bias in the distributions of mRS scores in either group (p=0.705). Figures 42 and 43.

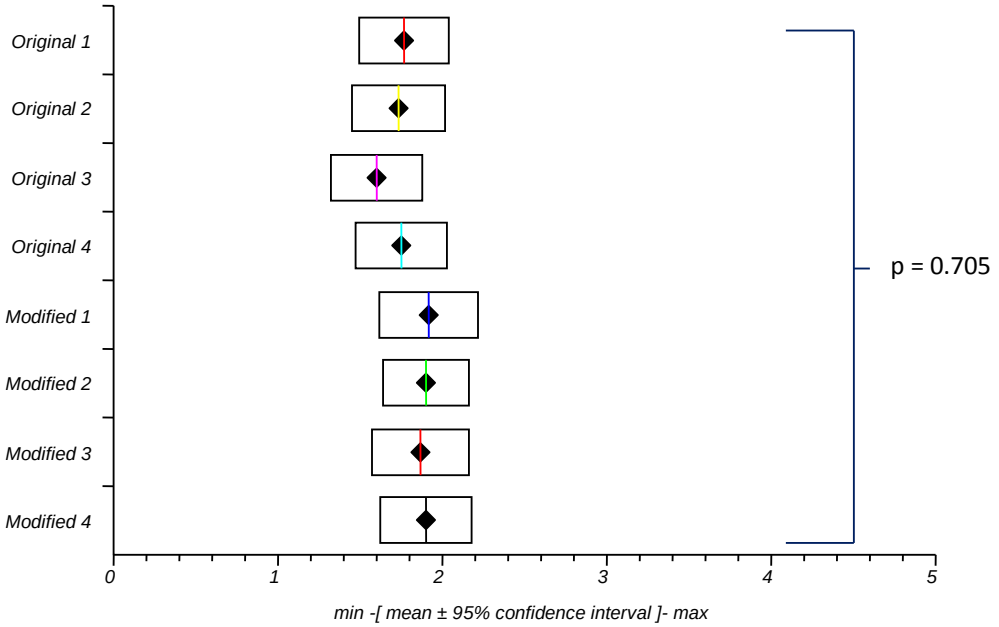


Figure 42 - Distribution of mRS scores in CARS translation sub study: Original and Modified Clips (n=60). Mean and 95% Confidence Interval. P value represents the Kruskal Wallis test of difference between distributions.

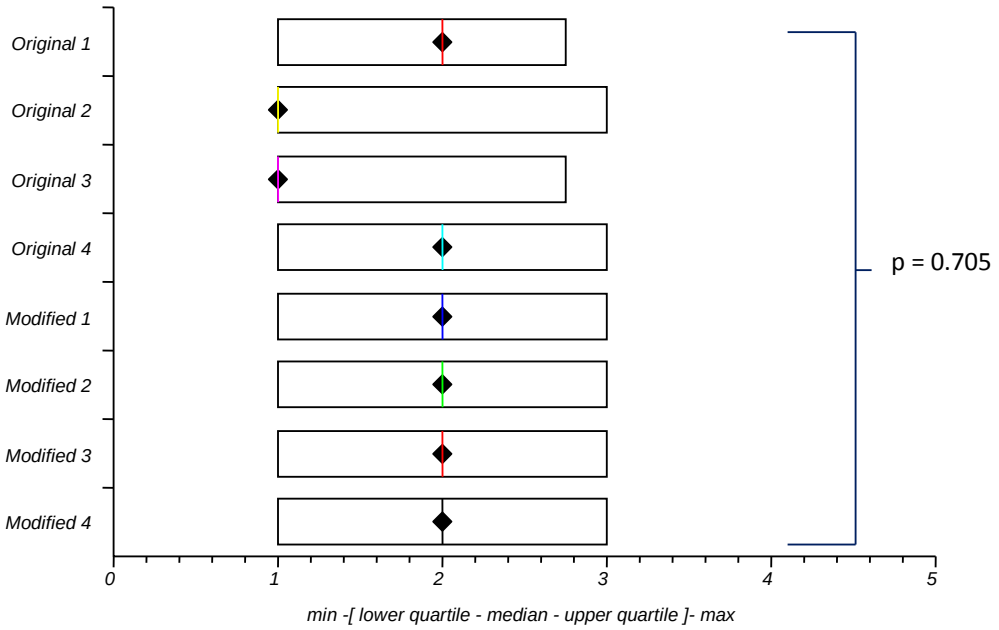


Figure 43 - Distribution of mRS scores in CARS translation sub study: Original and Modified Clips (n=60). Median and IQR. P value represents the Kruskal Wallis test of difference between distributions.

Inter observer reliability seen in the modified clips (κ 0.58 (95% CI 0.53-0.64), κ_w 0.85 (95% CI 0.74-0.95). ICC 0.85) was similar to that seen in the original video files (κ 0.64 (95% CI 0.58-0.70), κ_w 0.88 (95% CI 0.78-0.99), ICC 0.88) Table 38 and Figure 45.

Table 38 - Summary results – Inter-observer reliability of translated mRS (κ, κ_w and ICC)

	N	κ (95% CI)	κ_w (95% CI)	ICC
Translation Reliability				
All Native	69	0.62 (0.58 – 0.65)	0.91 (0.86 – 0.99)	0.91
All Translated	69	0.59 (0.55 – 0.62)	0.89 (0.82 – 0.97)	0.89
Medical Translation	20	0.67 (0.60 – 0.73)	0.91 (0.77 – 1.04)	0.91
Linguist Translation	20	0.64 (0.58 – 0.71)	0.91 (0.76 – 1.05)	0.91
Translation Feasibility				
Original	60	0.64 (0.58 – 0.70)	0.88 (0.78 – 0.99)	0.88
Modified	60	0.58 (0.53 – 0.64)	0.85 (0.74 – 0.95)	0.85

Inter observer Reliability in Native Language mRS clips (n=69)

Mandarin n=60
[Scored in Mandarin by Chinese Investigators]

English n=9
[Scored in English by UK investigators]

κ (95% CI) 0.62 (0.58 – 0.65)
κ_w (95% CI) 0.91 (0.86 – 0.99)
ICC 0.91

Dual translated mRS clips (n=20)
[Two transcripts prepared for each clip 1] by mRS certified clinician and 2] by linguist with no medical training]

Mandarin n=11
[Scored with English Transcript by UK Investigators]

English n=9
[Scored with Mandarin Transcript by Chinese investigators]

1] mRS certified transcript
κ (95% CI) 0.67 (0.60 – 0.73)
κ_w (95% CI) 0.91 (0.77 – 1.04)
ICC 0.91

2] Linguist transcript
κ (95% CI) 0.64 (0.58 – 0.71)
κ_w (95% CI) 0.91 (0.76 – 1.05)
ICC 0.91

Inter observer Reliability in Translated mRS clips (n=69)

Mandarin n=60
[Scored with English Transcript by UK Investigators]

English n=9
[Scored with Mandarin Transcript by Chinese investigators]

κ (95% CI) 0.59 (0.55 – 0.62)
κ_w (95% CI) 0.89 (0.82 – 0.97)
ICC 0.89

Figure 44 - Summary Results - mRs translation pilot study

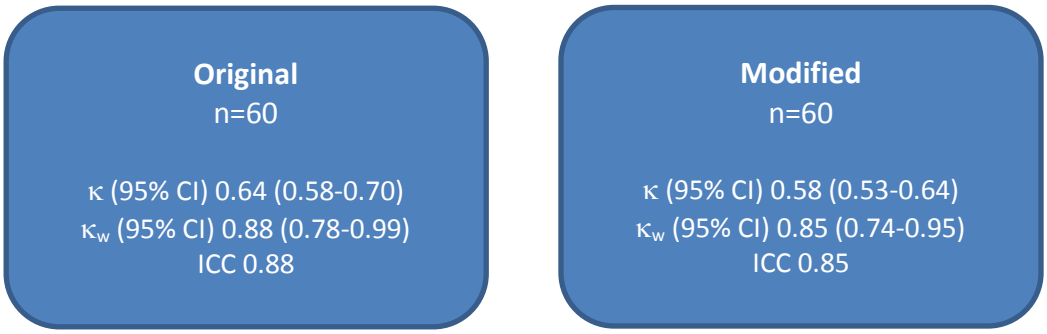


Figure 45 - Summary results - CARS translation sub study

8.4. Conclusions

Regulatory agencies such as the European Medicines Agency EMEA and US Food and Drug Administration FDA recommend the use of patient reported outcome measures in clinical research of medicinal products[205, 206]. It is acknowledged that the application of these measures in a modified form, such as use in a different population or language, may not have equal validity or utility[206]. Even within the same language, cultural differences may affect the acceptability of outcome measures in various populations[207]; for example Canadian French vs. the dialect of French spoken in France. In order to achieve the sample sizes required for successful phase III randomised controlled trials it is necessary to pool data from multinational sources. Pooling of clinical trial data from culturally and linguistically diverse groups may lead to misleading results if these issues are not considered in trial design[208]

The use of translated materials in medical research is common. When using a patient reported outcome, translation is frequently limited to the modification of a structured tool (e.g. scale or patient questionnaire) which is then administered and assessed in the native language. When using a tool such as the mRS, which is inherently unstructured and flexible, this brings added challenges if multinational, multicultural and multilingual outcomes are to be assessed centrally.

8.4.1. Translation in medical research

Four levels of language competence are described in the translation literature[209]. **Grammatical competence** – this is the goal of most beginner's language courses, to understand basic vocabulary and grammar, the ability speak and write simple sentences. **Discourse competence** – this demands a more complex understanding of vocabulary and grammar, the ability to converse and follow everyday conversations and to understand oral and written communication with complex sentence structure. **Sociolinguistic competence** – this demands the ability to express and negotiate the meaning of words and phrases according to the culture using the language, to integrate cultural norms into the communication process; for example to demonstrate appropriate politeness and respect in social situations. **Strategic competence** – this requires an ability to compensate for any lack of knowledge by using alternative vocabulary or non verbal cues appropriate to local culture. This level of linguistic competence allows the speaker to manage unexpected scenarios and social situations without being highlighted as a non native speaker. For research purposes a translator should have a minimum of sociolinguistic competence.

There are several levels of equivalence that must be obtained for a valid and reliable translation[210]. **Conceptual equivalence** – refers to constructs that exist, are relevant and acceptable in both cultures. **Semantic equivalence** – ensures that items mean the same thing to different groups. **Operational equivalence** – methods of administration of the assessment tool are appropriate for all cultures. **Measurement equivalence** – ensures that the test measures the same metrics in the source and target population. **Item equivalence** – confirms that items are not biased and carry equal weight in each culture. **Criterion equivalence** – ensures that the interpretation of scores is the same across groups.

In the translation of specific tools (e.g. questionnaires) there is a clear recommended multistep process to ensure that translations are valid and acceptable for use in medical research across countries. Regulatory agencies suggest that the method of translation is clearly and transparently described in publications of trial methods and results[206]

Details of the recommended translation procedures are widely published[211-213]. There is general agreement that the translation process must include: Synthesis of translation by at least two translators independently followed by a joint meeting to combine their translations where a written report is made to document issues and describe how these were resolved; then back translation by a further two independent translators is performed, blinded to the original version. There then must be some form of expert central committee review including clinicians and all involved translators to discuss any discrepancies. The modified tool must then be tested in a sample population with accompanying interviews before the process documentation is submitted to regulatory authorities.

The complexities involved in this process of translation would prohibit the use of this approach in translating patient interviews. There are reports of translated patient interviews used in qualitative research but the numbers are small. A qualitative study with patient translated patient interviews in Cantonese : English reported no significant differences in major categories but minor themes had some variation[214]. An alternative is to consider the use of an interpreter during the interview. This has advantages, any queries can be clarified by the researcher in real time; however the validity of this approach has been questioned[215]. An interpreter may summarise responses and without knowledge of the research field this may place limitations on data collection and threaten the content of the interview.

8.4.2. Translation in the Stroke literature

There have been several reports of translated outcome measures in the field of stroke. Translated outcome scales such as NIHSS (Italian[150], Portugese[66], German[216]), Barthel Index (Portugese[66], Persian[217], German[119]), Stroke Impact Scale (German[218, 219]), Motor assessment scale (German[220], Dutch[221]), European Stroke Scale (German[216]), Satisfaction with stroke questionnaire (German[222]) and Stroke and aphasia quality of life scale-39 (Spanish[223]) are available.

There have been two documented studies assessing a translated mRS scale; in German[216] and in Portugese[66]. These successful translation reports refer to modification of the wording

of the mRS scale, with subsequent application and scoring of this modified scale in the native language. There are no prior reports of translation of the mRS assessment itself.

We chose two linguistically and culturally diverse populations to assess the validity and reliability of the translated mRS. There has been one prior report of translation of a scale used in a Chinese population[224], but this has not been published in full.

8.4.3. mRS translation project

The challenges posed in the central adjudication of multilingual mRS assessments are clear. The number of interviews involved and the heterogeneity of content within precludes the use of recommended standardised translation techniques. The added complexity, time and expense involved in the multistep process may not enhance the reliability of results more than a single critical review by an independent bilingual observer[209]. A more pragmatic approach to translation is required and this is what we sought to investigate in the mRS translation project.

There are several factors for consideration in the application of translated, centrally adjudicated mRS assessments in clinical trials. We have demonstrated that translated mRS assessments appear to have equivalent reliability to native language mRS assessment and that incorporation of a translation step using digital dictation is feasible. However, there are limitations to the findings in this study and further research is required before we can confidently state that a translated mRS assessment is useful, valid and reliable.

The mRS pilot study included a small sample with only one language comparison. We deliberately chose two culturally and linguistically diverse populations and the focus of determining disability was noticeably different in the content of mRS interviews in each group. It is clear that the concepts relevant to health or illness are not equivalent in culturally diverse populations. In the Mandarin clips there was a strong focus on discussion regarding issues such as domestic tasks and caring for relatives where the UK clips were less likely to focus on these issues as a priority. The descriptors of independence in Chinese mRS clips included activities such as the ability to ride a pedal bicycle. This was not evident in the UK clips which were more likely to focus on driving a car or using public transport. With this in

mind it is important to highlight that the native language of the clips was not equally distributed, (English n=9 and Mandarin n=60). Each clip was scored in the native language and after translation, but the bias of Chinese clips does place limitations on the generalisability of the results. Many of the assessors in the Chinese group were bilingual. Although they performed the interviews in their native language; it is difficult to quantify any subtle changes that may have been made to the interview questions or technique as they had an English speaking audience in mind.

In the mRS pilot study each group of clips was translated by the same translator (medical or linguist), in the CARS sub study there were multiple translators. We have demonstrated that there is no significant difference in this sample using either a medically trained and mRS certified clinician or a linguist in the role of translator. This flexibility is important in the application of the model in a real clinical trial; the option of choosing a single or multiple translators with no necessity for experience in the relevant research area is a logistical advantage. Again, it must be emphasised that our sample of dual translated clips was small (n=20) and further study is indicated to ensure that these pilot study results are reproducible.

8.4.4. Summary

Including patient reported outcome measures is a challenge in multicultural and multilingual trials. There is no guidance for using translated patient interviews as an outcome measure but we have demonstrated in a small pilot study that the reliability of the mRS as an outcome measure is maintained when scoring translated assessments from two culturally diverse populations. The incorporation of a translation step into the central adjudication process seems technically feasible. Further work with multiple languages and a larger sample size is desirable to ensure generalisability of our results.

Chapter 9

Ranking within Rankin: can disability be graded within mRS grades?

9.1. Introduction

The mRS is an ordinal, hierarchical scale with broad descriptions between ranks. As we have discussed, the mRS descriptions are arguably vague and there is inherent variability in the scores allocated by independent observers. This limits the utility of the mRS as a clinical trial outcome measure.

In considering the mRS, or any other scale which contains grades, it is most useful conceptually to think of each grade as being distinct and discrete, with clear boundaries between ranks. However, in practice, the underlying distribution of subjects is likely to be more uniform, blurring the boundaries between grades. Figure 46.

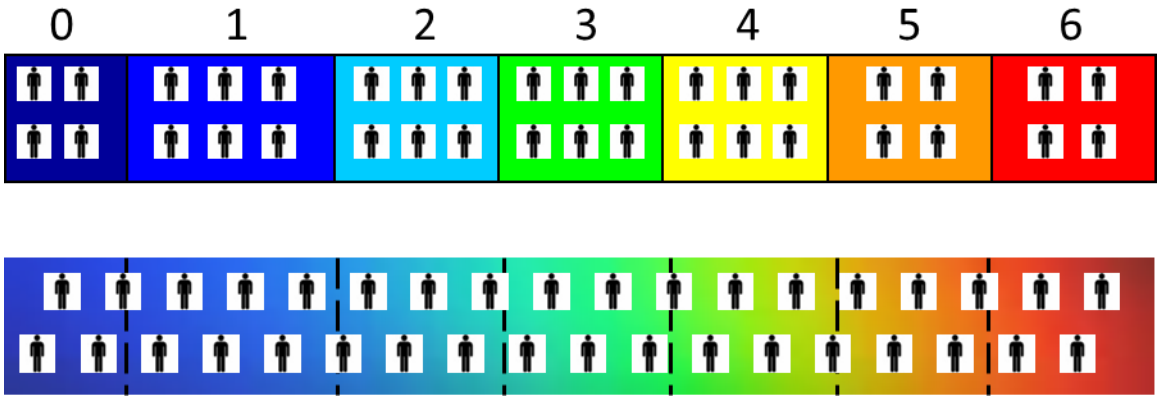


Figure 46 - Diagrammatic representation of the underlying distribution of disability within the mRS scale.

Between mRS grades, the underlying spectrum of disability within the scale is subtle, with some participants likely to sit clearly within an mRS grade where others might sit close to the line between grades. It is these participants, close to the boundaries, who are likely to contribute to inter-observer variability and misclassification.

Our aim is to accurately describe disability as an outcome measure in stroke trials. A method of detecting a difference both between and within ranks in the mRS might help us better understand the underlying "true" disability of stroke survivors and identify those who are closer to the boundary or more controversial in assigning mRS grade. Detection of discernible differences between participants within mRS grades may enrich the quality of data that are collected using the mRS as an outcome measure. We sought to assess the ability of stroke researchers to measure a more subtle effect on outcome through grading of outcomes within mRS categories. A method of identifying where in the spectrum of disability a participant lies has not previously been investigated. We used the CARS study video resource to determine the ability of stroke researchers to detect a difference in participants scored at each grade on the mRS scale.

9.2. Methods

The CARS videos represent a large sample of mRS assessments with multiple mRS scores. Each clip was scored remotely by four experienced mRS assessors and has an adjudicated consensus score. The local mRS score was not included in this analysis as it was based upon direct patient contact and therefore was not generated using identical source data.

These video assessments can be divided into two distinct groups. 1) **Classified (Agreement):** The original four adjudication committee assessments agreed and the video was automatically allocated the agreed mRS score. 2) **Misclassified (Disagreement):** There was disagreement in one or more of the adjudication committee assessments and a consensus score was allocated following group review and discussion. The degree and direction (towards greater or less disability) of disagreement can be identified by the spread of available mRS scores for each individual clip. The classified (agreement) clips can be considered "gold standard" examples of disability pertaining to that rank of the mRS, unlikely to represent a contentious mRS assessment. The misclassified (disagreement) clips exhibit a degree of controversy and can be considered to represent one of the mRS assessments that is likely to sit closer to the boundary between mRS grades.

In order to determine if mRS assessors were able to correctly identify the presence and direction of disagreement, pairs of "matched" mRS clips were compared. Ten trained and certified mRS assessors were asked to view each pair of mRS assessments for a subjective opinion regarding which participants' function was better or worse than their counterpart in the pair. We used only day 90 mRs assessments from the CARS video cohort in order to ensure that no duplicate clips containing the same participant were selected. We included two distinct samples in the comparison.

1. We selected a sample of classified (agreement) clips where there was no controversy within the original reviewing committee over mRS grade.

2. We selected a sample of misclassified (disagreement) clips, where there was more than one possible score on the mRS prior to consensus opinion. The spread of original mRS

scores in the misclassified clip was used to identify the direction of disagreement (i.e. to more or less disabled).

Each clip from the samples above was considered the “active” clip and was paired with a “control” clip. The “control” videos were all classified (agreement) clips selected from the day 90 CARS video cohort. Each pair had matched adjudicated mRs scores, generated automatically by agreed initial committee scores (in the “active” classified (agreement) and “control” groups) or by consensus following endpoint committee discussion (in the “active” misclassified (disagreement) group). See figure 47 for a summary of the study design.

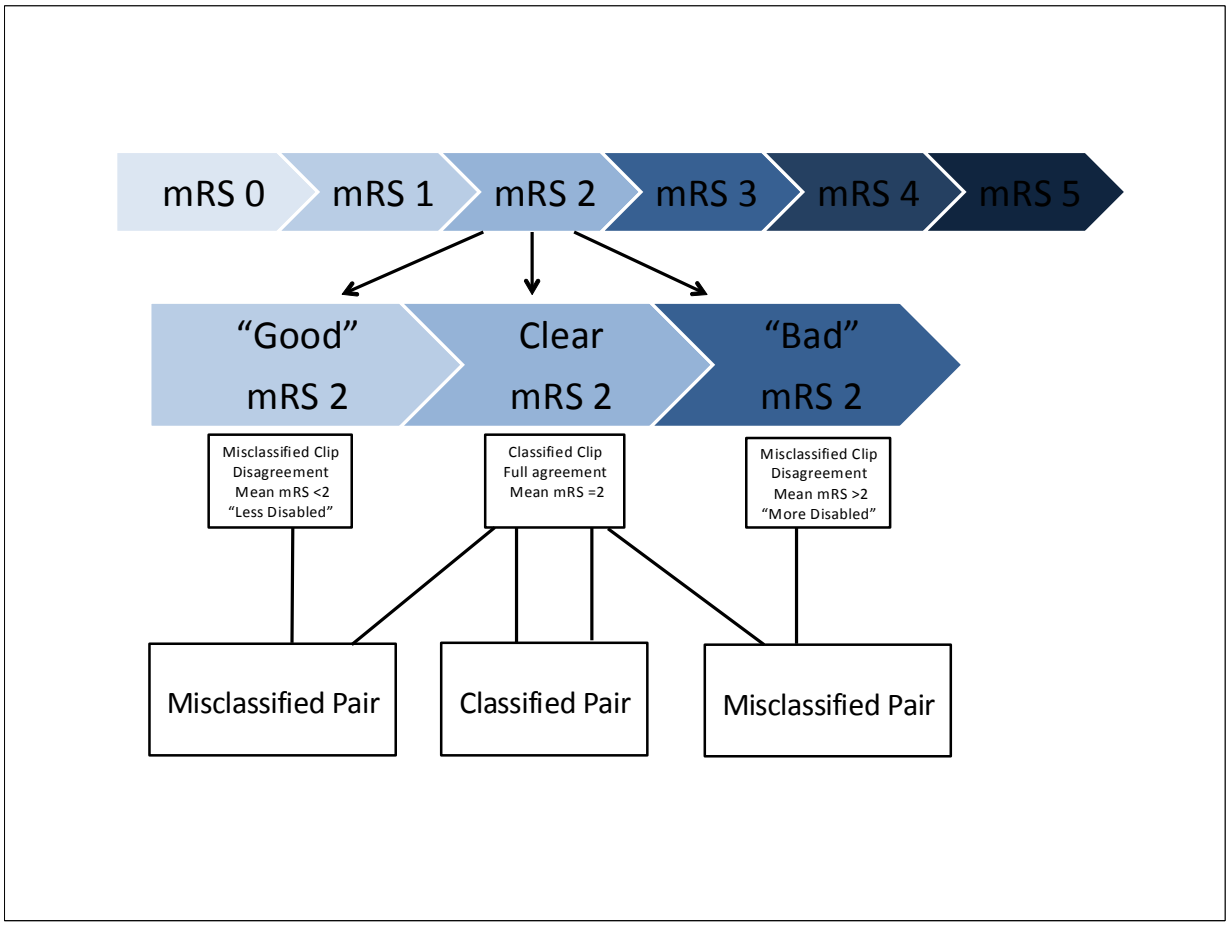


Figure 47 - "Classified" and "Misclassified" Pairs. [mRS 2 is used as an illustration, this process was repeated across the spectrum of mRS within the random sample.]

Each pair of video clips (active : control) was randomly allocated to four mRS assessors. The order in which the clips were viewed by the assessors (labelled simply as A and B) was also randomly allocated. Raters were given instruction to watch each pair of clips, in the order provided and blinded to each other's responses. After viewing each pair of mRS clips the rater was asked to note which of the participants (A or B) was "less disabled". They were not asked to provide a score on the mRS scale. We anticipated that there would be a proportion of clips (matched active classified (agreement) : control clips) where it may be very difficult to distinguish varying degrees of disability, due to the matched mRS scores involved. Despite this, raters were not permitted to provide a neutral response. This was to ensure that judgement was made in all clips and to simplify analysis.

Random sampling and allocation of clips, reviewer and order of viewing was done using R statistical software. A summary of the sampling process is shown in figure 48. Video clips were provided to each assessor on an individual CD with clips identified as A or B in the allocated viewing order. An electronic scoring sheet was provided to be completed during viewing and sent back to the study outcomes manager. The identity of the clips (control and active) was not unblinded until all assessors scores had been collated.

9.2.1. Statistical Analysis

Raters were asked to provide a judgement regarding which mRS assessment clip in each pair represented less disability. Half of the sample comprised matched pairs of classified (agreement) clips in which we anticipated this judgement would be difficult and might generate an arbitrary response or "guess". We sought to identify if the judgements were more predictable in the misclassified (disagreement) pairs and if this could be quantified. The ability of assessors to identify the direction of disagreement was quantified by comparing the clip identified as representing "less disability" to the clip with the lower mean mRS in adjudication scores. Our hypothesis was that raters would be able to identify those clips that had generated disagreement during original committee scoring by agreeing consistently that those clips with a lower mean committee mRS were "less disabled". The proportion of assessments where this was identified correctly and the rater agreed with the original mRS assessors was calculated. Equality of two proportions was tested using a test of

two proportions with a null hypothesis that there is no difference between the two proportions.

In order to determine if the agreement among assessors for the classified (agreement) clips and misclassified (disagreement) clips is more than that which would be expected by chance alone the agreement between raters was compared using kappa statistics (κ). Weighted kappa statistics (κ_w Fleiss-Cohen Weights $[1-[(i-j)/(1-\kappa)]^2$) and intraclass correlation coefficient (ICC) were not relevant in this analysis due to the categorical nature of the data.

Statistical analysis was performed using StatsDirect statistical software.

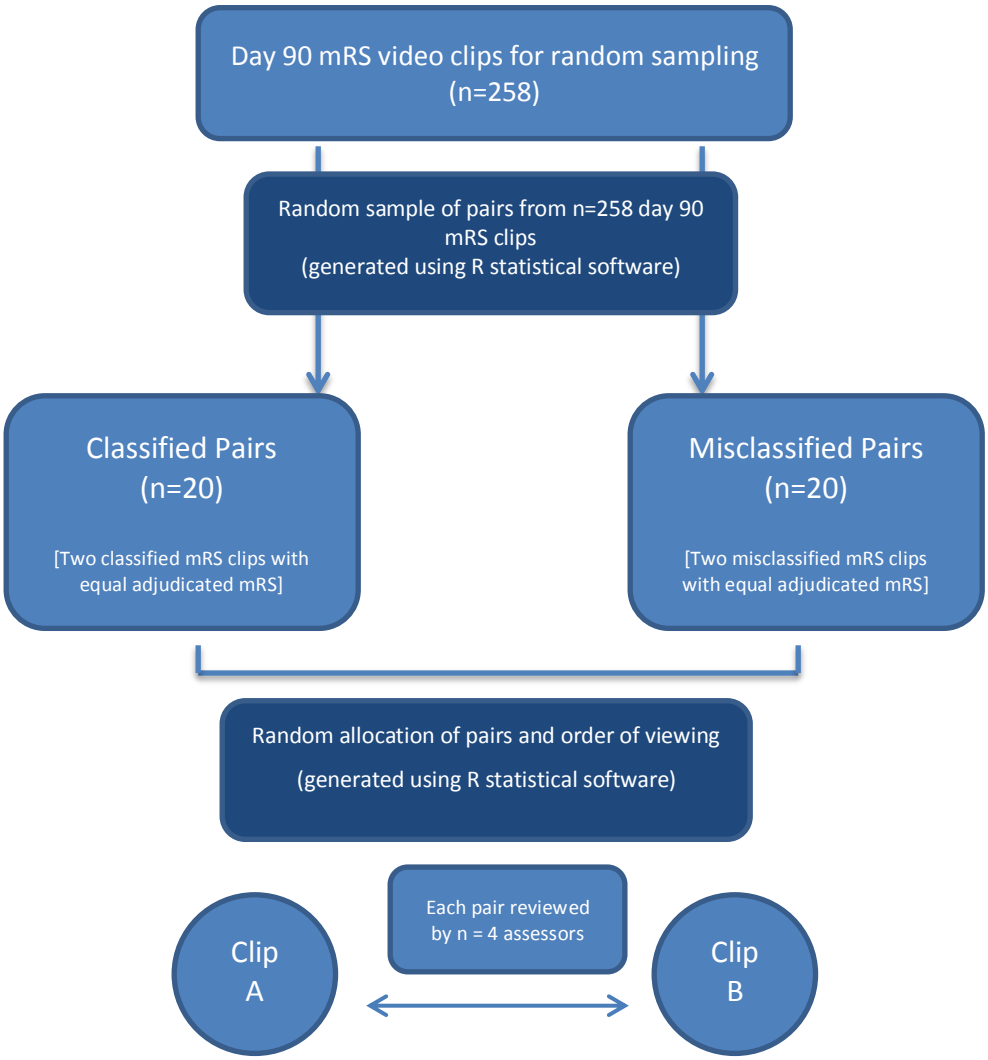


Figure 48 - Sampling of "Classified" and "Misclassified" pairs

9.3. Results

We randomly selected n=40 pairs of clips for review; n=40 active clips in each group (n=20 classified (agreement) and n=20 misclassified (disagreement) and n=40 control clips (Total clips reviewed n=80).

Ten mRS raters each reviewed sixteen pairs of clips. Each pair was reviewed by 4 independent assessors. In total we received data for 160 paired assessments (n=80 for classified (agreement) video pairs and n=80 for misclassified (disagreement) video pairs.

There was no difference between the proportion of raters who agreed which participant was "less disabled" in the classified (agreement pairs) group [54 / 80 (67.5%)] or the misclassified (disagreement pairs) group [55 / 80 (68.75%)]. There was no difference between proportions (p=>0.999) Table 39.

Table 39 - 2x2 Table displaying proportions of raters in agreement for each group. Misclassified (disagreement) pairs: agreement represents correct identification of "less disabled" clip. Classified (agreement) pairs: agreement represents chance agreement between raters.

Classified (agreement) Pairs [Agree = chance agreement of "less disabled clip due to matched scores]		Misclassified (disagreement) Pairs [Agree = correct identification of "less disabled" clip]		
		Agree	**Disagree**	**Total**
	Agree	55	25	**80**
	Disagree	54	26	**80**
	Total	**109**	**51**	**160**

The mean mRS was equal in the classified (agreement) group of paired clips due to the initial agreed mRS at adjudication review. In the misclassified (disagreement) group the difference in mean mRS between active and control clips ranged from -0.4 (suggesting less disability) to 0.8 (suggesting more disability). There was no pattern to suggest that greater disagreement (as demonstrated by difference in mean mRS) was likely to result in correct identification of the direction of disagreement at initial mRS review. Table 40.

Table 40 – Number of assessors who correctly identified the direction of disagreement in Misclassified (disagreement). Mean difference in mRS represents the magnitude of disagreement.

	Difference (mean mRS)	1 correct	2 correct	3 correct	4 correct	Total
Less Disability	-0.4	1		1	1	3
	-0.2	1		2	2	5
	0		1			1
	0.2		1	2		3
	0.4	1	2	1	1	5
More Disability	0.6				2	2
	0.8		1			1
	Total correct	3	5	6	6	20

There was no pattern to suggest greater ability to correctly identify the "less disabled" mRS clip at any level on the mRS scale. This was also seen in the control group, with no pattern in agreement among raters at any level on the mRS scale. Figure 49.

There was poor agreement between raters for all paired assessments, κ (95% confidence interval) 0.075 (-0.033 – 0.183). Agreement between assessors for the clips that were classified (agreement pairs) and misclassified (disagreement pairs) was equivalent. (κ (95% confidence interval); Classified pairs 0.077 (-0.065 – 0.218) and Misclassified pairs 0.078 (-0.084 – 0.241).

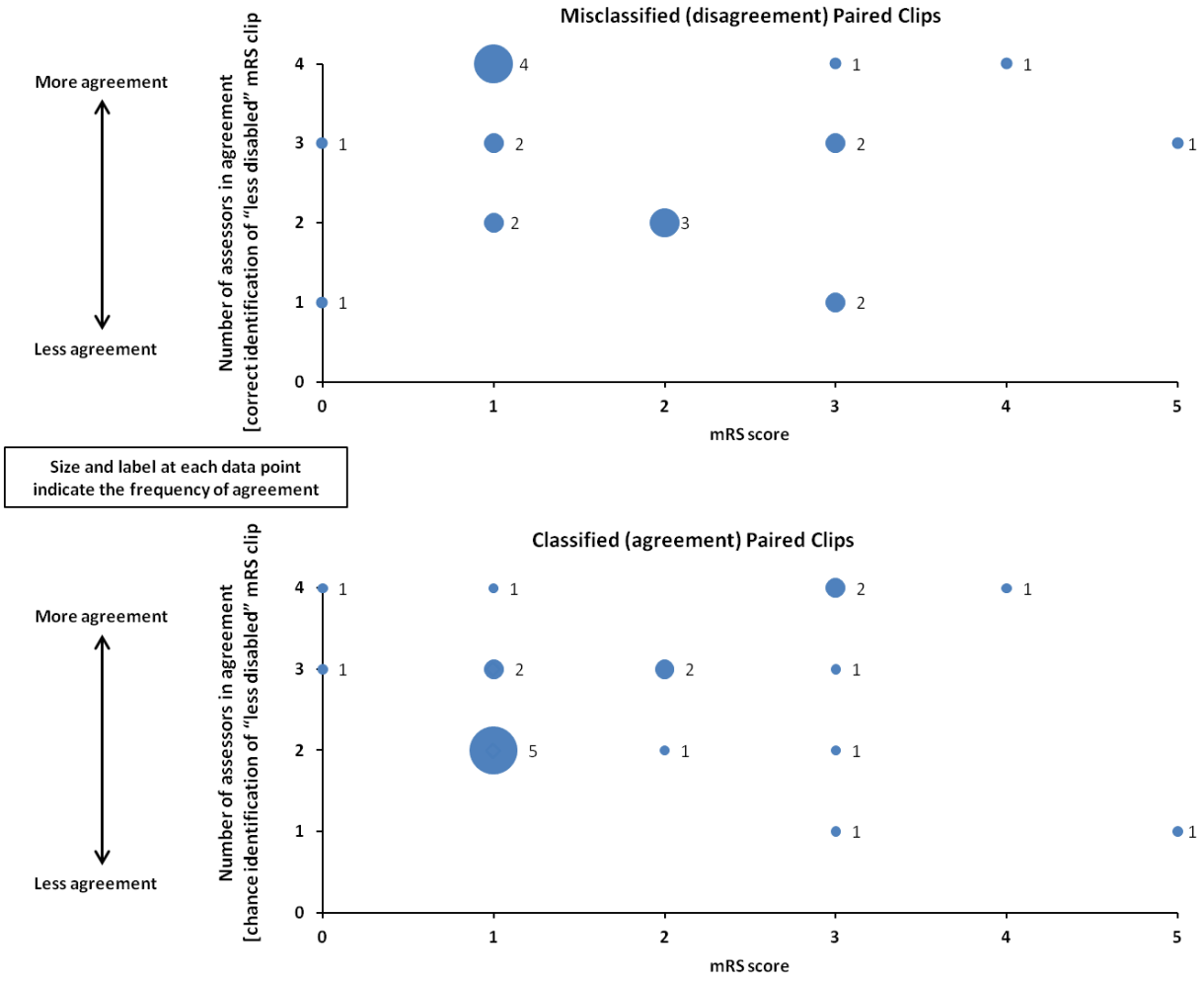


Figure 49 – Frequency of agreement between raters for Misclassified (disagreement) paired clips [indicating correct identification of "less disabled" mRS clip] and Classified (agreement) paired clips [indicating chance agreement] for each rank on mRS scale.

9.4. Discussion

The spectrum of disability seen in stroke survivors ranges widely. Converting this complexity into ranks on an ordinal scale is challenging. There is a degree of error in all measurements, but that error can be significant where it results in grouping error; i.e. a participant is placed in the incorrect outcome grade or is "misclassified".

There have been studies in the neurotrauma literature highlighting the difficulties in misclassification of outcomes. The result is reduced power and treatment effect size[144, 145]. Outcome assessment following head injury presents similar challenges to that following stroke – heterogeneous insults resulting in a wide variety of possible functional deficits.

The optimal number of outcome grades on a scale is debated. It is acknowledged that compressing a scale to a small number of grades (dichotomy, trichotomy etc.) wastes valuable information and may limit the detection of a genuine treatment effect[154]. However, there is also evidence to suggest that an increased number of ranks on an outcome scale results in greater inter-observer variability and misclassification. The Glasgow Outcome Scale (GOS) used following head injury is a disability outcome scale very similar to the mRS. In its original format there are 5 grades, an extended Glasgow Outcome Scale (GOSE) with 8 grades is also in widespread use. Versions of the scale with more ranks have been found to result in greater variability, threatening the utility of the extended scale[225, 226].

The mRS in its traditional form has 7 ranks. Ordinal analysis of the mRS provides more information than a binary outcome[104]; but even this is much less sensitive to change than a continuous outcome measure. The ability of an ordinal scale to accurately quantify outcome is determined by the width interval of each rank. Broad ranks result in the disposal of potentially useful information whilst finer distinctions between ranks may make it more difficult to distinguish between function in each group. In an interval scale, the size of the difference between each outcome group is equal or quantifiable, however using an ordinal scale there are likely to be non-uniform steps between grades.

Using a procedure developed by the WHO Global Burden of Disease Project (WHO-GBDP) the interval distances between mRS grades have been reported using disability weights[111]. Figure 50. This clearly depicts the continuous nature of the disability spectrum within the mRS. It also helps to define the difference between a change from mRS 3 to mRS 4 (broad interval) and between mRS 5 to mRS 6 (fine interval).

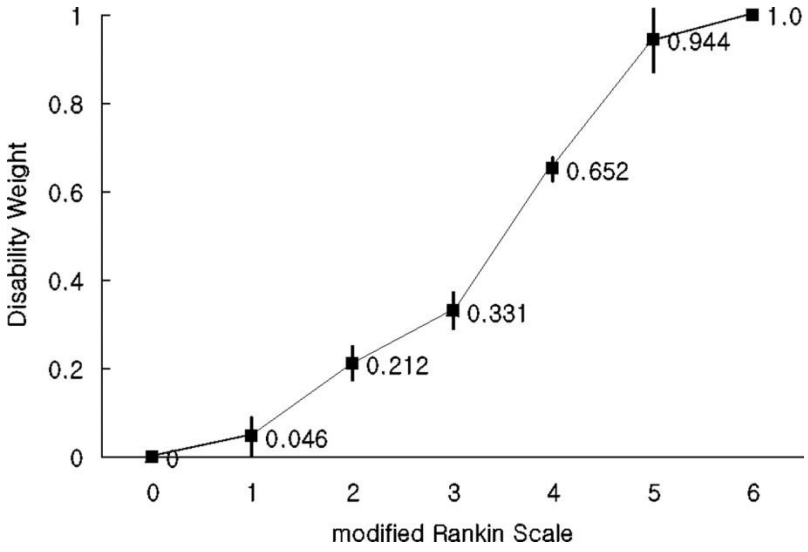


Figure 50 - Disability Weight (and 95% CI) for each grade of mRS. Disability weights generated using WHO Global Burden of Disease[111].

We aimed to explore the potential for experienced mRS raters to be able to identify these finer distinctions within groups. If it is possible to rank disability within mRS grades this could be used to display outcome data in a more continuous form, allowing for alternative statistical strategies.

We have found that our team of mRS raters could not reliably identify those clips that represent "less disabled" mRS participants. Our hypothesis was that the clips in the "active" group; i.e. those that contained a clip that resulted in disagreement paired with a gold standard agreed mRS clip; would be identifiable to the raters as ones which clearly depicted a participant at the "good" or "bad" end of an mRS grade in comparison to one in the middle

of an mRS grade. The results suggest that comparison of the "active" and "control" pairs were similar enough to prevent identification of the "active" pairs. There was no pattern to suggest that raters were more likely to identify pairs with a greater change in mean mRS score and no pattern to suggest that identification of the "active" pairs was more likely at any level on the mRS scale. From the disability weights data we could postulate that it should be easier to define the difference between mRS participants in the middle grades of the scale than at the more extreme end of the disability spectrum but we cannot support this hypothesis using our data.

The grading of the mRS with "good" or "bad" forms of each grade is not reliable on the basis of this exploratory study. There are limitations to this study due to the small number of clips and assessors involved, however the results are equivocal in every sense. Significance testing and kappa statistics indicate that there is no agreement between assessors beyond that which we could expect by chance alone. Perhaps alternative methods of converting the ordinal ranks of the mRS scale into a more continuous distribution should be investigated; such as the use of a mean mRS score following multiple mRS ratings.

Chapter 10

Discussion and Conclusions

The outcome measures available for use in stroke trials are large in number and diverse in content. The choice available is testament to the fact that there has been a lack of uniformity in the outcome measures used and that no one tool is infallible. The aim is to identify a tool that is universally accepted and can be used in a standard form in most stroke trials, aiding interpretation of the relative risks and benefits among treatment strategies.

The mRS has become the preferred primary outcome measure in most stroke trials and has gained favour for many reasons[29, 227]. Experience and familiarity with the mRS is widespread; there are accepted procedures to complete training and gain certification in its use. A 90 day mRS outcome is available for the majority of stroke trials conducted over the last decade, although not consistently as a primary outcome measure. The availability of this data is useful in comparing and pooling the results of clinical trials, for example using the Virtual International Stroke Archive (VISTA)[228]. The mRS is also a useful measure in terms of quality of life[110, 111] and economic measures[94]; metrics which are crucial in the application of novel treatment strategies.

This thesis has examined the potential benefit of improving reliability of the traditional mRS in stroke research and has detailed a novel assessment technique utilising a central adjudication model.

Statistical modelling techniques using real stroke trial data distributions have demonstrated that there are meaningful gains in sample size to be realised if we can reduce variability in mRS scoring and improve the reliability in stroke trials. The reliability of the standard mRS is likely to be poor based on available estimates. We have reported a possible reduction in sample size of 20% to 25% with improvement of mRS reliability from baseline κ 0.25 to κ 0.5 or κ 0.7.

This may translate to important financial and ethical benefits for trialists. The ethical benefits in reducing the number of participants in clinical trials are clear; we must strive to randomise as few participants as possible in order to gain the required evidence to guide optimal treatment. The financial gain in including fewer participants is less clear and less immediately justifiable. The cost analysis of every clinical trial is individual to that trials aims, however, a reduction in participants by 25% would translate to significant saving in terms of cost and time to completion of recruitment.

We do not propose that mRS reliability can be included in sample size calculations, the complexities involved clinically and statistically are too great. However, we have illustrated that there is an inherent degree of inter-observer variability in mRS assessment that impacts upon trial power and strategies to reduce variability would be of benefit in general terms.

We have found that the use of remotely assigned mRS scores may be feasible in a multicentre acute stroke trial. Our web based outcomes portal was successful with very few technical failures. Of course in a clinical trail the loss of any data through technical failure is unacceptable, scientifically and ethically. However, it is important to note that with the CARS trial design there would always have been a locally assigned mRS score to be used in the event that a centrally assigned score was unavailable. The video equipment that we used has long since been superseded by smaller, cheaper versions that are easily connected to the internet through wireless technology. This can only offer advantage to our model in the future.

The central adjudication model was an entirely new concept to investigators and participants. We found no reluctance from investigators to be involved in the video process.

Some sites were more active in recruiting than others, largely reflected by the research staff available. Some investigators required a little more guidance than others in negotiating the technological aspects of the trial but we found a clear learning effect as the trial progressed.

Participant withdrawal was substantial and 17% of study visits were missed. Although the majority of withdrawals took place before any study procedures we cannot conclude that the video process did not have an effect on this withdrawal rate. We must be cautious in drawing the conclusion that this model is feasible and acceptable, in a real RCT missing outcome data on this scale would have a substantial effect on analysis and the interpretation of results. Our data suggest that central adjudication may be acceptable and accessible to investigators and participants but further study is necessary in the context of a true RCT. Investigation of the model with larger scale application must be undertaken; our findings may not be applicable in other countries or cultures where infrastructure and health care facilities differ.

The real challenge in documenting mRS outcomes is that we are trying to use an ordinal scale to accurately document function on a continuous spectrum of disability. From our data we cannot conclude that a remotely allocated score more or less accurately reflects the "true" disability. There are no means by which we can assign a numerical score with the sensitivity and precision required to score "true" disability. This unknowable and unquantifiable concept is an unrealistic target for comparison. The best alternative is to compare assignment of outcome to current gold standard practice; which is local mRS assessment by a trained and certified investigator. We have found that remotely assigned mRS scores are valid, in comparison to current best practice and factors known to be related to outcome after stroke. In the future validity could be further tested in comparison to a standard mRS assessment in the detection of a treatment effect that we have previously recognised (such as rtPA) or may find to exist in ongoing or future clinical trials.

We have demonstrated reliability in the CARS study that is favourable in comparison to current estimates of mRS reliability. Reliability among the adjudication panel was κ 0.59. In light of the published estimate of mRS reliability in a multi-centre trial[100] and the result of our simulations, this may indeed translate to a reduction in sample size of between 20% and 25%. Our translation pilot and sub-study have provided encouraging data to support the use

of central adjudication in an international, multicultural and multilingual trial. Further study with several languages is warranted.

We have investigated methods of identifying the "difficult" mRS assessments likely to contribute to inter-observer variability. Our data do not suggest that there are any consistent predictors of variability in mRS scoring and that the difficulties in scoring individual clips arise through factors that are individual to that participant, that investigator and that interview. Analysis of paired mRS clips to assess raters ability to identify "good" or "bad" examples of each mRS grade was also unsuccessful.

Where the hypothesis had been that we might be able to pre-determine those assessments likely to need assistance in mRS grading, our data suggest that any strategy to improve reliability in mRS grading must be applied to all outcome assessments. Our central adjudication model was to provide multiple scores for each assessment, allowing estimation of the reliability among observers. This has been feasible but labour intensive for the adjudicators involved. In practice there might be alternative and more pragmatic approaches to providing central adjudication and quality control without full committee review of all clips. A single independent adjudicator might review each clip and only proceed to full committee review where there is disagreement. These strategies have not been formally tested and require further validation.

The major opponent to the use of centrally adjudicated mRS outcomes are the numerous structured forms of mRS assessment. Several structured mRS assessment tools have been proposed to limit variability in mRS outcomes. A simplified modified Rankin questionnaire has found very good reliability with a very fast administration time (average 1.67 minutes) in a sample of 50 paired mRS assessments[132]. A simple nine question interview only allowing binary "yes/no" responses demonstrates excellent reliability[131]. A more comprehensive structured mRS interview was originally developed by Wilson et al[105] and has been adapted to the Rankin Focussed Assessment (RFA) by Saver et al[133]. This structured four page form is designed to be completed in conjunction with interview of the participant, relative or caregiver and review of participant medical records. The RFA reports excellent inter observer

reliability (κ_w 0.99) in a sample of 50 paired mRS assessments and has been prospectively validated as part of the ongoing FAST-MAG trial.

There is currently much uncertainty regarding the relative benefits of a traditional mRS approach and that of the many structured mRS interviews available, such as the Rankin Focussed Assessment (RFA). Despite the improved reliability quoted for the structured mRS tools there are some advantages to the traditional mRS model. There are formal training tools with rigorous assessment and a freedom within the assessment tool to explore the complexities of symptoms and limitations specific to each participant. This may in part be why previous study has not convincingly shown benefit of structured interviews[99]. Initial positive results of structured mRS tools have not been replicated in independent cohorts and further independent validation is required before any of the tools become the method of choice for documenting outcome in acute stroke trials. In a randomised evaluation of traditional mRS vs. Wilson's structured mRS tool the promised benefits in reliability were not replicated[199]. The global and unstructured nature of the traditional mRs is a great advantage; without relying on individual activities of daily living there are no floor or ceiling effects in its application which are common to structured instruments. Any potential benefit of the structured interview is at the cost of increased interview time and complexity. However, the same is likely to be true of novel interventions currently being piloted, such as video based mRS or off-line group assessment. In a large-scale clinical trial that may recruit hundreds of patients, the cumulative effect of even minor increases in interview complexity could have major effects on overall costing and time to completion. Further study of each approach to mRS assessment is required in large multicentre trials. The benefit of a structured assessment such as the RFA and central adjudication together may hold potential. An advantage of the central adjudication model is that it can be used regardless of the method of mRS adopted – traditional, structured or questionnaire based. It could readily be used alongside a shortened Rankin assessment to ensure consistency and quality.

Beyond reliability, the benefits seen with central adjudication are numerous. Any central adjudication panel allows a degree of "expert" review and we have demonstrated that no important data are lost in the process of video recording and remote assessment. Blinding is

crucial to the integrity of trial outcomes and a remote group adjudication approach may be of use in these circumstances where this is difficult e.g. neurosurgical interventions or complex rehabilitation trials. Central adjudication provides a method of ensuring quality control; repeat assessment or further information can be gathered if an assessment is inadequate or below standard. In these circumstances a group review approach may prevent a potentially erroneous outcome score being recorded. The video approach also allows storage of a "hard copy" of the outcome assessment allowing trialists to re-examine functional outcome data where there are data queries. Finally, it offers remote source data verification of the patient's existence and consent in a way that no document can offer.

A criticism of the video mRS interview would be that not all information gathered by an investigator in their contact locally with a participant can be captured on film and transferred to the adjudication centre. Intuitively it seems likely that some information is lost where the mRS score is based only on the video clip. There may be clues to function and disability in each encounter that are not captured on the clip. i.e. How did the participant travel to the hospital, did they struggle to walk to the consultation room etc. Pilot work comparing face to face interview and video interview did not suggest this was a limitation of video mRS. It is important to acknowledge that in its definition, the mRS score is based upon a patient interview. A properly conducted interview should contain all the detail required. It is recognised that for scales clinicians may not perform a comprehensive assessment but will estimate the results based upon initial meeting, watching the participant in a consultation and clinical intuition[229]. This is not accurate for mRS[199]. Any assumption based on visual clues should be clarified and explored by a diligent reviewer on questioning. Where there were non-verbal cues we encouraged interviewers to highlight these in their interview where they felt these might be relevant. No guidance was given to investigators regarding the length of the video clip, where there were more complex issues we encouraged full discussion of these to aid in scoring. One piece of information that is concealed in the video clip is treatment assignment: subtle prejudice that could cause bias in scoring for an open label or PROBE trial can be prospectively excluded.

For many reasons the mRS is sub-optimal; however the traditional mRS is likely to remain the outcome measure of choice for trialists and regulators. An imperfect tool applied consistently is preferable to multiple versions of that tool with individual modifications. There is a temptation to modify or alter an existing scale to improve its efficiency or reliability; for example to add, omit or re-write certain items. This adds complexity and may reduce the utility of single outcome measure applied constantly in clinical research. To pool data, such as the pooled analysis of the landmark thrombolysis trials[44], it is necessary to have a shared, consistent and identical primary outcome measure. For this reason we should accept the mRS in the format that it is commonly applied and work to improve its use in trial design.

We have demonstrated that mRS assessment can be performed remotely via a method that is feasible, acceptable to participants and investigators and is valid and reliable in comparison to current known metrics of disability assessment.

There is now a need to apply our approach to real world intervention trials. Based on our encouraging initial experiences, central adjudication using our infrastructure is already being used in the following trials: CLEAR-3 Clot Lysis Evaluating Accelerated Resolution trial of intraventricular thrombolysis in the treatment of intraventricular haemorrhage (NCT00784134, NIH funded), EuroHyp-1 trial of therapeutic cooling after ischaemic stroke (EU FP7 grant), MISTIE-III Minimally Invasive Surgery plus rtPA for ICH Evacuation Phase III (NCT01827046, NIH funded) and the SITS-OPEN trial investigating the safety and efficacy of thrombectomy after initiation with intravenous rtPA in ischaemic stroke (charitable and industry funded).

Appendix A

Written Documentation provided to local investigators to supplement face to face training session

The Central Adjudication of Modified Rankin Scale Disability Assessments in Acute Stroke Trials Study

The CARS Study

Contents

1. Study Contacts

Please feel free to contact any of the study team if you need to.

Chief Investigator

Professor Kennedy Lees
Professor of Cerebrovascular Medicine
Dept Medicine and Cardiovascular Sciences
Western Infirmary
Dumbarton Road
Glasgow G11 6NT
Tel: 0141 2112176
Email: k.r.lees@clinmed.gla.ac.uk

Co-Lead Investigator

Dr Jesse Dawson
Lecturer in Medicine and Clinical Pharmacology
Dept Medicine and
Cardiovascular Sciences
Western Infirmary
Dumbarton Road
Glasgow G11 6NT
Tel: 0141 211 6395
Email: j.dawson@clinmed.gla.ac.uk

Trial Manager

Mrs Pamela MacKenzie
Trials Manager
Dept Medicine and
Cardiovascular Sciences
Western Infirmary
Dumbarton Road
Glasgow G11 6NT
Tel: 0141 211 2176
Email: pcn1w@clinmed.gla.ac.uk

2. Introduction

Acute stroke require a robust measure of functional outcome. At present, the modified Rankin Scale (mRs) is the most popular outcome measure (table 1) and is an ordinal scale with 6 categories ranging from zero (no symptoms) to five (complete physical dependence). A sixth category can be added to signify death. Despite being the most commonly used assessment, there are some concerns. Considerable inter-observer variability is recognised meaning that observers often disagree even when assessing the same patient.

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

We have already established that agreement between observers can be worryingly low when they assign scores to the same patients. This raises the possibility, should disagreement be across an endpoint cut-off point in a clinical trial, that patients could be placed in the wrong outcome group. This reduces statistical power and could compromise a clinical trial.

Digital video recording of mRs assessments in a large clinical trial could address this concern. It will limit the effect of inter-observer variability by allowing central "off-line" scoring by a small number of expert investigators. It will also permit validation and re-scoring of initially misclassified patients, or in situations where disagreements occur (there will always be some disagreement but a consistent approach to these subjects is crucial). It will also help ensure quality of data (via source data verification and by ensuring adherence to interview procedures) and improve blinding of endpoint assessment in studies where this is difficult (such as neurosurgical studies). It may also afford examination of more subtle gradations of disability.

3. The Aim of The CARS Study

The aim is to evaluate use of digitally recorded and centrally adjudicated mRs assessments in a multi-centre acute stroke trial. Before digital recording of outcomes and central adjudication could be widely adopted, it must be rigorously assessed. Even though this study is based upon an adaptation of an already commonly used method there are several areas we must note. First, the mRs by nature is subjective and whether extra information (such as how the patient travelled to hospital or other background details) contributes and by how much is unclear. Further, this approach will add to complexity of trial design and although we feel this additional complexity is marginal, the technique must yield benefit before it could be deemed worthwhile

4. Observer Training

All investigators must be trained in mRs and NIHSS assessment using a validated web-based training programme. A link to the training web-sites can be found in the training section of the web portal. You require to register your own account on each site. For mRS training, the link is **http://trials-rankin.trainingcampus.net** . **For NIHSS training, the link is** **http://learn.heart.org/ihtml/application/student/interface.heart2/index2.html?searchstring=583** . You will be unable to upload any assessments if training has not been completed.

All will also be shown how to operate the video camera and given a practical demonstration on video upload procedures and use of the Rankin Outcome Adjudication web portal.

The reference booklet can of course be referred to and the co-ordinating centre contacted at any point.

5. Getting Started

Before starting the study there are some simple tasks we ask you to follow (as well as completing training). These are to make sure you are comfortable using the equipment, that you have performed a test upload and that you have completed all relevant training.

We also require that you install AVS Video Converter version 6 software on your computer (we will provide discs and activation codes). You may need to liase with your local IT department for this to be performed. You should also make a folder on your computer hard drive called "CARS Assessments." This can be located anywhere as long as you know how to find it.

6. The Study Population

The aim is to recruit a minimum of 300 patients from between 5 and 10 centres. The inclusion and exclusion criteria are detailed below. We hope each centre will recruit at least 15 patients.

Main Inclusion Criteria
Diagnosis of acute stroke (ischaemic or haemorrhagic)
Onset within 48 hours of ictus
Demonstrable deficit on the National Institutes of Health Stroke Scale (NIHSS)

Main Exclusion Criteria
Pre-morbid modified Rankin score of ≥3

7. Identification of Participants and Consent

Potential participants should be identified by a member of the treating clinical team as soon as possible after admission to the Stroke Unit. Suitable patients and/or their nearest relative should be approached by the study clinician or nurse and will have the project explained briefly to them. Those who are willing to consider it further can then be approached by a member of the research team (who may also be part of the treating clinical team) to have the study explained in more detail and consent will be sought for participation.

Since no study specific procedures that differ from usual care are required in the first days of participation, the patient and relatives will be allowed at least 48 hours to decide. However, consent can be obtained earlier, including less than 24 hours, if participants are willing to proceed on that basis. The first video interview does not take place until one month later and of course participants can decline involvement at that stage.

Participants should be asked to sign the consent form. Two copies will be signed (one each for the participant and the site file). The consent form should be copied with the copy placed in the case notes. Consent can be taken by one of the investigators or by a study research nurse (in which case it will be countersigned by an investigator).

Participants who are unable to consent for themselves can be included; assent from a relative / welfare guardian will be sought. In Scotland this is done under the Adults with Incapacity (Scotland) Act and similar provisions are in place in England and Wales under the Mental Capacity Act. In practice, the early procedures are indistinguishable from normal care and it is only the later video interviews that are "intrusive". Those who are able to consent at this stage will be asked to do so at this stage. For those who remain unable to consent at this stage due to severe disability, continued assent / agreement from the relevant relative or welfare guardian will be accepted.

No participant will be included in England and Wales or Scotland against the wishes of a relative / next of kin. Importantly, in the case of severely affected participants it is more likely to be a member of staff or a relative/carer who will be interviewed, though the patient may be seen in the video recording.

8. Trial Design and Conduct (figure 1)

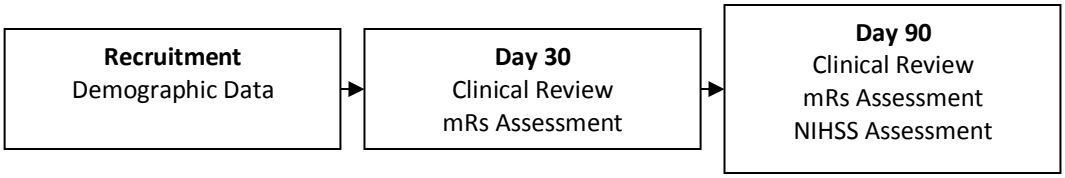


Figure 1 – Participant Flow Chart

Baseline Assessment
At recruitment, baseline demographic details including a measure of stroke severity (NIHSS) should be gathered. No intervention or change to normal routine care will occur during this study. These data should be entered into the paper case report form (CRF) and the electronic (e) CRF at the time of the visit.

Day 30
A digitally recorded mRs assessment should be performed. An assessment for any serious adverse event should also be made. This visit should last under 15 minutes. Data from this assessment should also be entered into the paper case report form (CRF) and the electronic (e) CRF at the time of the visit. An mRs score must be assigned by the local observer at this point and any comments regarding dysphasia or other problems noted in the CRF.

Day 90
A digitally recorded mRs assessment should be performed, as should a recorded NIHSS assessment. An assessment for any serious adverse event should also be made. This visit should last under 30 minutes. Data from this assessment should also be entered into the paper case report form (CRF) and the electronic (e) CRF at the time of the visit. An mRs score must be assigned by the local observer at this point.

9. Guide to Performing The mRs Assessment

These will be performed on survivors in standard fashion according to each centre's normal practice, although guidance is available here. The assessment should ideally be performed in a quiet and private clinic room, or if needed by a patient's bedside with the curtains drawn or at home if they are unable to attend the hospital. Before an assessment is performed outwith a hospital site, it should be ensured that a local Lone Worker Policy is in place and that this is followed.

Whenever possible, the assessor should remain constant across the follow-up period for a given patient. We recognise these restrictions may sometimes be impractical.

The main mRs assessment must be recorded using a digital video camera, **unless the participant clearly has a mRs score of 5** where a proxy should be interviewed on video in their stead. A suitable proxy is a relative, member of nursing staff or other carer.

Note that only symptoms arising since the stroke should be considered. Walking aids or other necessary mechanical devices are disregarded provided that the patient can use these without external assistance.

The score of 0 is awarded to patients who have no residual symptoms after their stroke, not even minor symptoms.

If patients have any symptoms resulting from the stroke, whether physical or mental, then they should be scored at least 1 on the Rankin scale. For example, if they have any new difficulty in speech, reading or writing,

in physical movement, sensation, vision or swallowing, or any change in their mood that does not limit their activities, they still should score 1. Patients in this category can continue to take part in all of their previous work, social and leisure activities. For this purpose, "usual" is regarded as any activity that they used to undertake for a monthly basis or more frequently.

If there is any activity that they used to undertake that they can no longer do since the stroke, whether because of a physical limitation or because they have chosen to give up the activity as a result of the stroke, then they should be scored 2 on the Rankin. In this category the patient has slight disability and is unable to carry out all his previous activities, but he is still able to look after all of his own affairs without any external assistance. For example, a patient would be scored in this category if he used to drive a car and is no longer able to do so, or if he used to have a job whereas he now no longer works. The patient should still be able to look after himself without any daily help. In other words he will be able to dress, move around, eat, go to the toilet, prepare simple meals, undertake shopping and make short journeys by himself. He will not require any supervision from other people and could safely be left at home for periods of a week or more without any concern.

Rankin category 3 is for patients who have moderate disability. These patients require some external help for daily activities but are able to walk without assistance. They may use a stick or a frame for walking but the assistance of another person is not required for this. They will be able to manage daily activities such as dressing, toileting, feeding etc, but will need help for more complex tasks such as shopping, cooking or cleaning or will need to be visited more often than weekly for some other purpose. The external help may simply be advisory, for example supervision for their financial affairs.

Patients with moderately severe disability who are unable to walk without assistance and unable to attend to their own bodily needs by themselves are given a score of 4. These patients are not independently mobile and will need help with daily tasks such as dressing, toileting or eating. They will need to be visited at least daily or will need to live in close proximity to a carer. To discriminate patients in category 4 from those in the most severe category, consider whether the patient can regularly be left alone for moderate periods of a few hours during the day.

Patients who cannot be left alone even for a few hours should be given the score of 5. Patients in category 5 have severe disability and are usually bedridden, incontinent and require constant nursing care and attention. Someone else will always need to be available during the day and at time during the night, although this will not necessarily be a trained nurse.

Thus, in summary, to distinguish between patients in category 0 or 1 consider whether the patient has any remaining symptoms. To distinguish between categories 1 and 2 consider whether the patient can undertake all of his previous activities. If the patient is independent of others in activities of daily living, then he should be scored 2 rather than 3. To distinguish between category 3 and category 4 the crucial question is whether the patient can walk without the assistance of other people. Finally, a patient who can be left by himself for a few hours during the day would be given a score of 4 rather than 5.

It is important to note that patients do not always fall neatly into one category and some judgement is usually required when scoring them. When in doubt between 2 categories, always stick to the key discriminators of the scale. Thus if the patient has remaining symptoms he scores at least 1. If the patient is unable to undertake previous activities he scores at least 2. If he is dependent upon others in activities of daily living he must score at least 3. If the patient is unable to walk without assistance he must score at least 4 and if the patient is bedridden and requires constant nursing care he will score 5. **Finally, if there is still some doubt between two alternatives on the scale, and both options appear equally valid, then the worse option should be chosen.**

There are some key discriminating questions that should be considered when using the modified Rankin scale. These are shown in more detail below (the official definitions of each category are shown below in bold and the italicized text provides guidance that may reduce interobserver variability, without requiring a structured interview).

0. **No symptoms at all**

The patient should be unaware of any new limitation of symptom caused by the stroke, however minor.

1. **No significant disability despite symptoms; able to carry out all usual duties and activities**

The patient has some symptoms as a result of the stroke, whether physical or cognitive – for example affecting speech, reading or writing; or physical movement; or sensation; or vision; or swallowing; or mood – but can continue to take part in all previous work, social and leisure activities. The crucial question to distinguish grade 1 from grade 2 (below) may be, 'is there anything that you can no longer do that you used to do until you had the stroke?' As a guide, an activity that was undertaken more frequently than monthly could be regarded as a 'usual activity'.

2. **Slight disability; unable to carry out all previous activities but able to look after own affairs without assistance**

The patient will be unable to undertake some activity that was possible before the stroke (e.g. driving a car, dancing, reading or working) but is still able to look after him/herself without help from others on a day to day basis. Thus, the patient can manage dressing, moving around, feeding, toileting, preparing simple meals, shopping, and travelling locally without needing assistance from anyone else. Supervision is not necessary. This grade assumes that the patient could be left alone at home for periods of a week or more without concern.

3. **Moderate disability; requiring some help, but able to walk without assistance**

At this grade the patient is independently mobile (using a walking aid of frame if necessary) and can manage dressing, toileting, feeding, etc but needs help from someone else for more complex tasks. For example, someone else may need to undertake shopping, cooking or cleaning and will need to visit the patient more often that weekly to ensure that these activities are completed. The assistance can be advisory rather than physical: for example, a patient who needs supervision or encouragement to cope with financial affairs would be in this grade.

4. **Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance**

The patient requires someone else to help with some daily tasks, whether walking, dressing, toileting or eating. This patient will be visited at least once and usually twice or more times daily, or must live in proximity to a carer. To distinguish grade 4 from grade 5 (below), consider whether the patient can regularly be left alone for moderate periods during the day.

5. **Severe disability: bedridden, incontinent, and requiring constant nursing care and attention**

Someone else will always need to be available during the day and at times during the night, though not necessarily a trained nurse.

10. Recording the mRs Assessment

A Canon HF100 video camera will be used. The camera records direct to an internal memory card. In conjunction, a desktop omni-directional condenser boundary microphone will be used (ATR97, Audio-technica, Ohio USA; Specifications: Frequency response: 50-1500Hz). An easily portable desktop tripod will be used to mount the video camera (Hama Minipod).

It is crucial to ensure that the microphone is correctly connected and switched on before commencing the recording. Also ensure there is enough battery power in the camera or that it is plugged into the power supply. It is also recommended that, prior to any recording, that a do not disturb sign is placed on the outside of the door and that ward or clinic staff are informed recording is taking place to minimise external noise.

Please do not alter the settings on the camera as these have been adapted for the purposes of this study. For your information, these settings will be shown in the on-line appendix.

It is recommended that the assessment is performed with the assessor, participant and equipment positioned as shown below (figure 3).

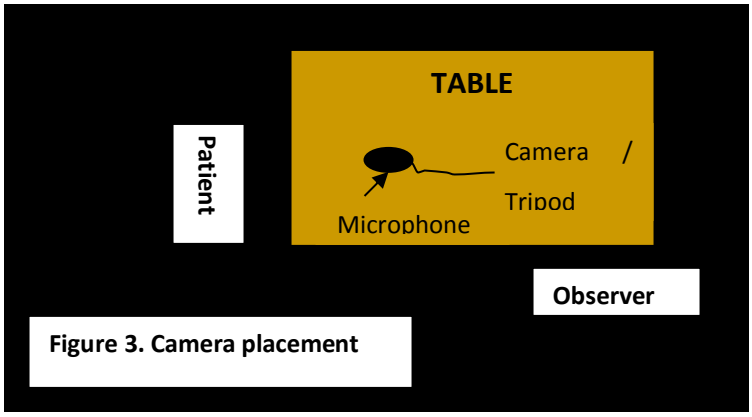


Figure 3. Camera placement

The camera should be mounted on the desktop tripod.

Figure 4. The Hama Mini-pod

The desktop tripod has extendable legs (1) which may not be required, although can be used to optimise the height of camera and views of patients obtained. It also has an adjustable lever to change angulation of the camera (2).

The camera is attached by aligning the in built screw on the tripod (3) with the corresponding area on the lower surface of the camera and then turning the circular disc located just below the camera mount on the tripod (4).

To give maximum stability, it is recommended that the camera lies directly above and not between one of the tripod legs.

Once the camera is mounted, it should be turned on and it is also recommended that the power supply to the camera be connected to maintain full charge.

This is the side view of the Canon HF 100 camcorder. The LCD screen can be gently opened the camera at the finger groove (1).

The on-off switch is located on the superior aspect of the camera (2)

Figure 5A. The Canon HF 100 camcorder.

This is the rear view of the Canon HF 100 camcorder. The record switch can be seen (1). The microphone and power sockets are located behind a protective cover (2 and 3) which can be gently opened.

The menu on the LCD screen is navigated by using the small knob at its side (4).

Figure 5B. The Canon HF 100 camcorder.

The boundary microphone should be connected to the "microphone in" socket on the camcorder (shaded red. The headphone socket which is not to be used for the microphone is green). The boundary microphone must be switched on. The switch is located on the bottom of the microphone. This should then be placed on the table between the patient and the assessor as shown above.

The on-off switch is located on the lower surface of the microphone. It must always be confirmed this is switched on before recording commences. This should also be switched off at the end of recording to preserve battery life.

Figure 6. The ATR 97 Omnidirectional Microphone

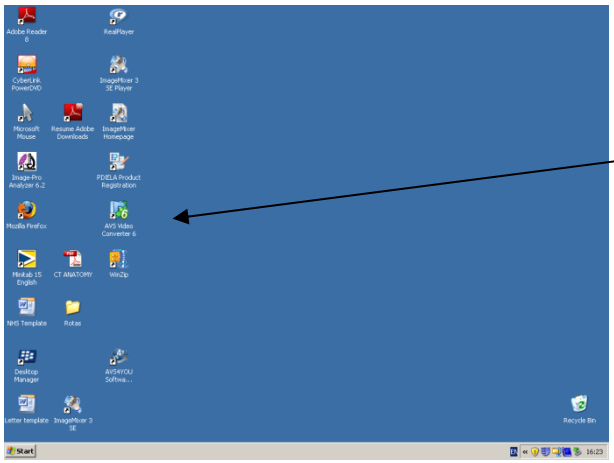
Once the equipment is set up, the LCD screen should be opened and the position of the camera adjusted so that the patient's face and trunk are seen. Once this has been done, the recording can begin. Either the remote control or the record button (shown above) on the camera can be used. **Always ensure the red light on the LCD screen is present (which means the camera is recording) before starting your assessment.**

After the assessment is complete, remember to stop recording and we recommend that you switch off the microphone.

11. Saving the Assessment and Converting the File

First, the camera needs to be connected to a USB port on your computer. The computer should automatically recognise the camera. Sometimes a window of files located on the camera will automatically open. This should be closed. The recorded clips are very large (at least 200 to 400 megabytes) and require to be converted to a smaller size prior to upload to the study web portal. However, a copy of the full size clip should be recorded to compact disc and archived locally along with other source data. This can be done as normal and details of how to locate the file and how to name it are given later.

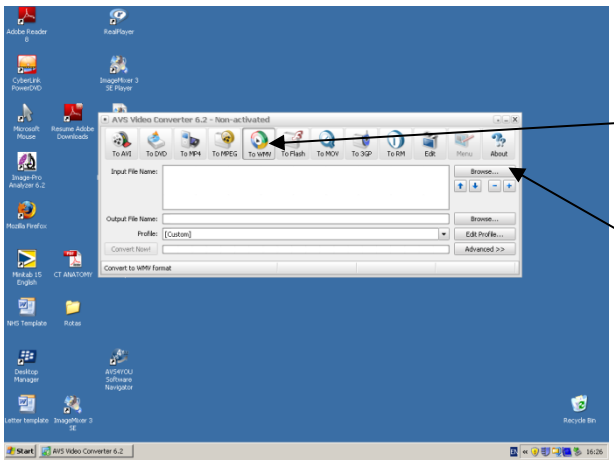
A step by step illustrated guide of how to do convert the file is shown below. The converted files should range between 10 and 30 megabyte in size.



The file format needs to be converted to reduce file size.

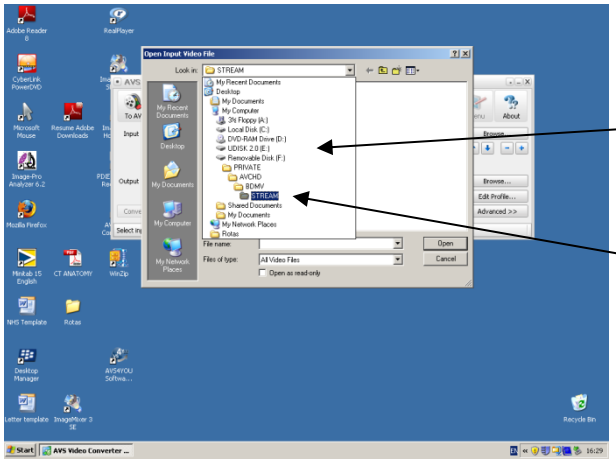
We will use AVS video converter version 6 software to do this.

Double click on this icon to open the software.



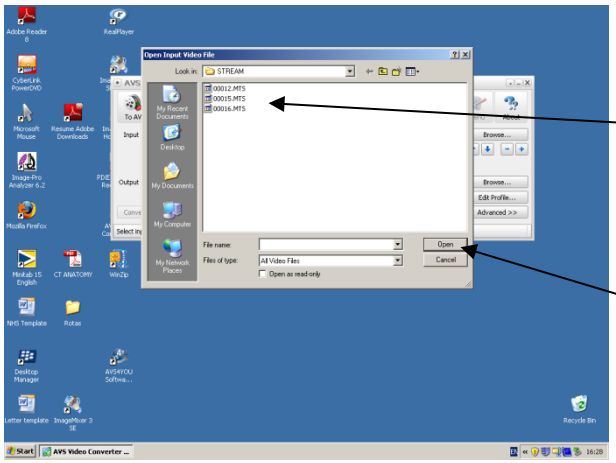
When the software is opened, this window will appear and it is important to ensure that the WMV icon is highlighted.

The first step is to locate the file that you wish to convert (the mRs assessment you have just recorded). Do this by clicking “browse” next to the input file name bar.



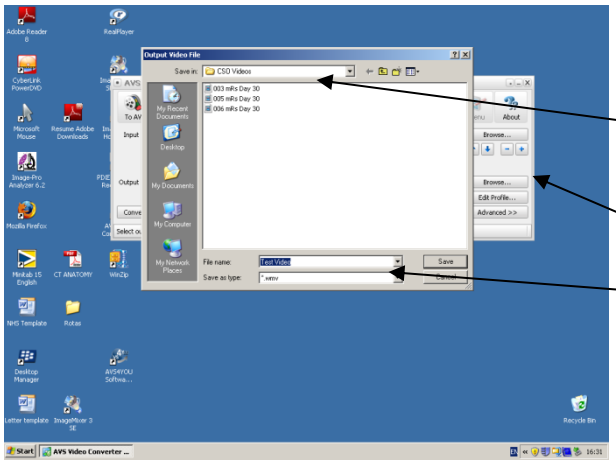
The camera will be identified as a removable disc (usually by the letter “E” or “F”) and the files are located in the “STREAM” folder.

To open this first click on the removable disc that represents the camera then “PRIVATE” then “AVCHD” then “BDMV” then “STREAM”.



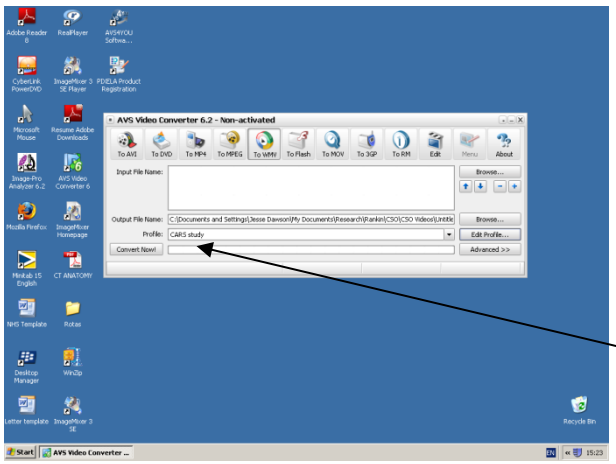
The camera automatically saves and labels files as a number. The last video recorded has the highest number so should therefore be the last mRs assessment performed.

Once you have highlighted the files you wish to convert, click open.



You then need to identify the folder you wish to save the assessment to. We recommend you use the CARS assessments folder that you have created.

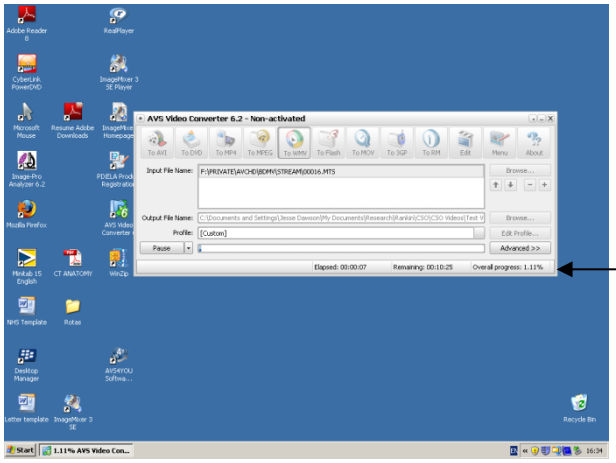
To locate this, click on browse and this screen will appear. Then type in the file name which should start with the participant number then the type of assessment (for example “003 mRs Day 30”). This name should also be used for the copy you save to compact disc.



The profile refers to the conversion parameters and this will be automatically determined during the site initiation visit. The profile will be named (CARS study).

You must ensure this has been selected before beginning the conversion.

This can be selected from the drop down menu.



Once all these steps have been followed, you are ready to perform the conversion which takes about 10 minutes. To start this, click on "convert now" and this screen will appear.

The bar at the bottom will tell you the progress of the conversion. When conversion is finished you will be asked if you want to "open the folder" in which the new assessment is located. This window can be closed.

11. The Rankin Outcome Adjudication Web Portal and File Upload

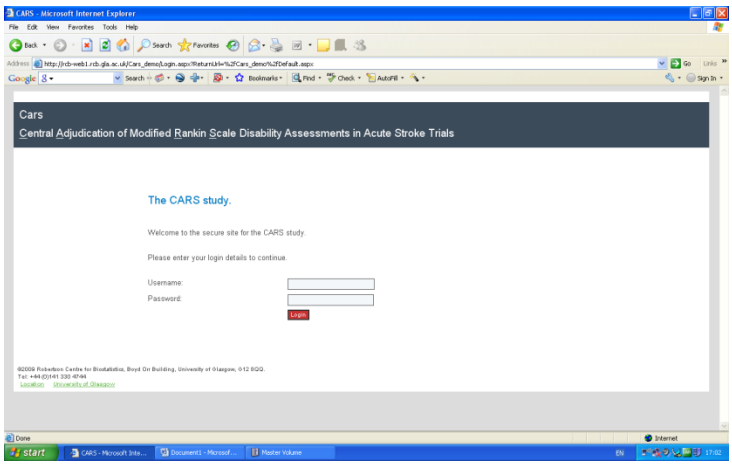
The web portal provides tools for investigators to enter their subjects' modified Rankin Scale assessments and upload accompanying videos. Demographic and other outcome data should also be entered via the web portal. The portal is administered by the Robertson Centre for Biostatistics in Glasgow.

The web portal is secure and end-users access the system by entering a username and password. On first use, users will be asked to change their password. Smart passwords will be required and users will be prompted to change these routinely. Access to the portal is restricted to named co-investigators.

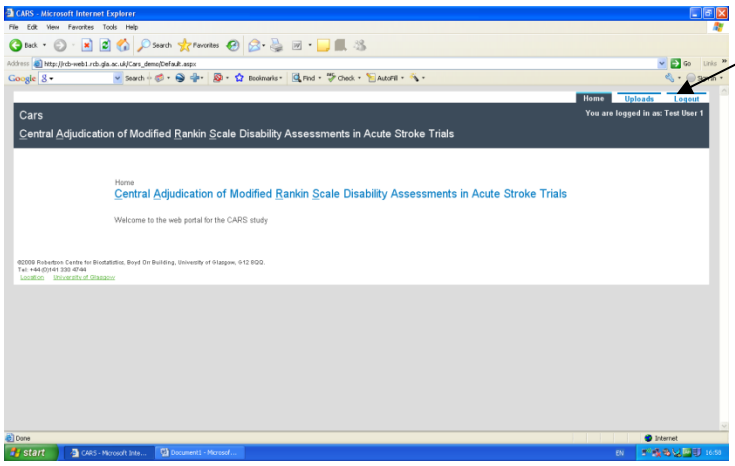
The web portal also includes a system that will make videos available to the co-ordinating centre for quality checks and pre-review editing and transcription. In addition, it will allocate clips to investigators for review and mRs scoring. For confidentiality reasons, such access will be restricted to those at the outcome coordinating centre at the Western Infirmary Acute Stroke Unit, Glasgow, UK.

A step by step guide of how to use the portal is shown below.

The web portal can be located at the web address www.glasgowctu.org/cars

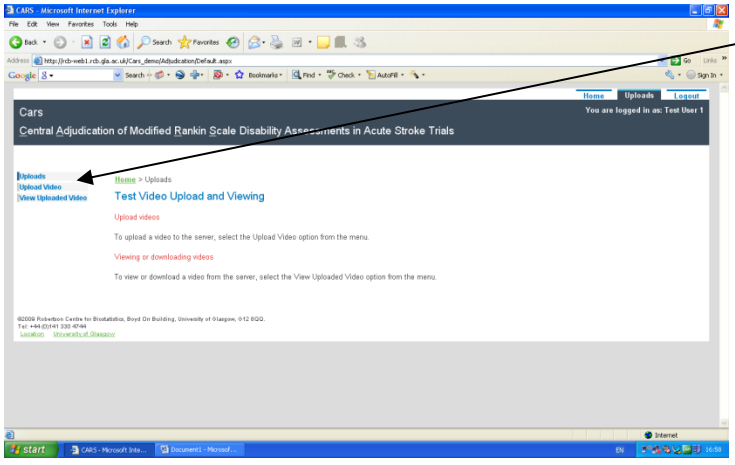


To enter the portal you are required to enter your personal user name and password

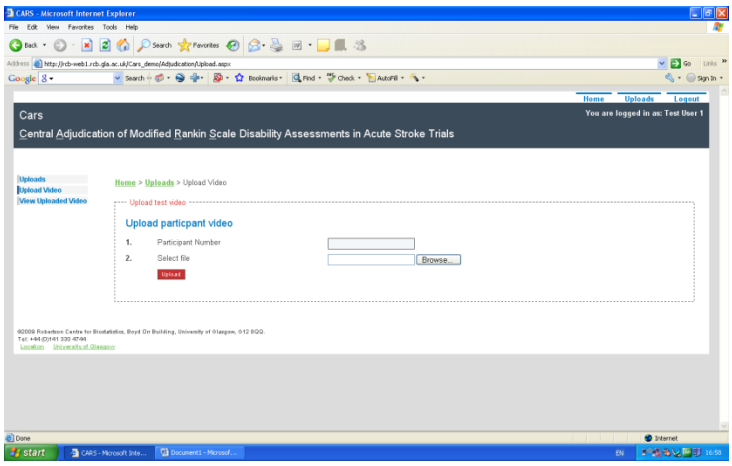


Once you are inside the portal, you will see the menu bar at the top right of the screen. This will contain links to study resources, CRFs and the clip upload section.

To upload a clip, clip on the “uploads” link.

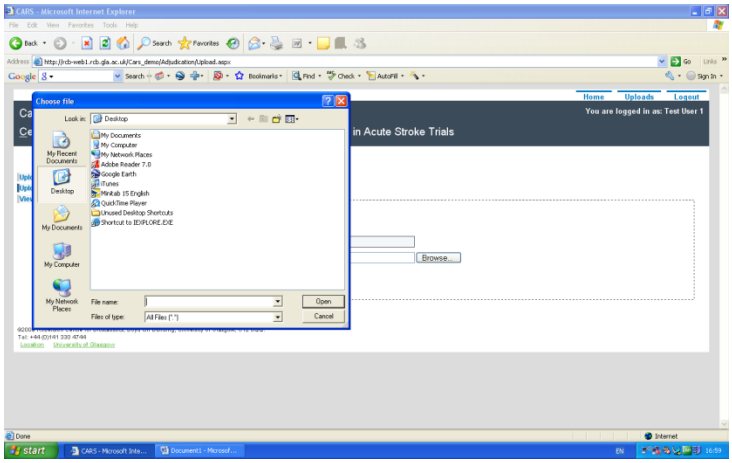


To upload a clip, clip on “upload video”.



You will then be required to enter the participant number and the assessment day (day 30 or 90).

Then you will need to locate the file for upload by clicking browse.



The file should be located in the CARS assessment folder you created.

Once the clip has been identified, click on open and then upload. The duration of the upload will depend upon the size of the file and the network speed but should be under 5 minutes. The portal will tell you if the file has been successfully incorporated into the database.

12. Video NIHSS Assessments

A day 90 video NIHSS assessment will also be performed and uploaded to the web portal in the same fashion as the mRS assessment. To capture the full NIHSS assessment, it may be helpful to place the camera approximately 10 feet from the patient, near the foot of the bed and on the opposite side of the bed from the examiner.

13. Recording of Participant Events

Although this is not an interventional clinical trial, we hope to capture serious adverse events. This will allow us to better identify individuals where the video technique is unsuitable or problematic should that scenario arise. Events do not require to be reported to the sponsor but should be recorded in the paper and eCRF.

A serious adverse event (SAE) is defined as any event that;

a. results in death
b. is life threatening
c. requires hospitalisation or prolongation of existing hospitalisation
d. results in persistent or significant disability or incapacity
e. consists of a congenital anomaly or birth defect.

Appendix B

Written information given to translators and assessors in the CARS translation substudy.

CARS TRANSLATION SUBSTUDY

BOOKLET OF INFORMATION AND INSTRUCTIONS

Contents of the Translation Guide

Contacts:

For new login details for the CARS portal please contact the CARS helpdesk at CARS@glasgowctu.org or Kate McArthur at kate.mcarthur@glasgow.ac.uk

For any other queries contact Kate McArthur at kate.mcarthur@glasgow.ac.uk

Summary

The translation substudy is designed to test the feasibility and validity of introducing a translation step into the central adjudication model. Most large stroke trials are international and multilingual. A method for dealing with this is necessary if central adjudication is to be feasible in a typical RCT.

The CARS study has been performed exclusively in English, with various different local accents and colloquialisms. Despite the common language used, we do have an arbitrary "border" between Scotland and England/Wales which we can use for these purposes as a language boundary.

A selection of clips (n=60) has been randomly sampled from the pool of adjudicated 90 day CARS videos. There are 30 clips from either side of the "border" and these will be translated by investigators from the other side of the UK. I.e. Scottish clips will be "translated" by investigators in England/Wales and vice versa.

The translation will be done with a digital dictaphone and the file uploaded to the CARS web portal. This will be merged with the original video file to create a dubbed clip which can then be re-scored on the mRS.

The scores generated from the translated clips will be compared to the original "native language" scores to see if there are any significant differences.

The "translated" video file:

The allocated translator of each clip will watch the video and dictate a "dubbed" version of the sound on the clip. The translator will find that they need to start and stop the video and may rewind and overwrite bits of their dictation – this is expected and a less smooth version of the sound file (in comparison to the original) is inevitable. We do not wish translators to spend a large amount of time getting the translation word perfect - simply a close copy of the conversation between the assessor and the patient (and / or proxy).

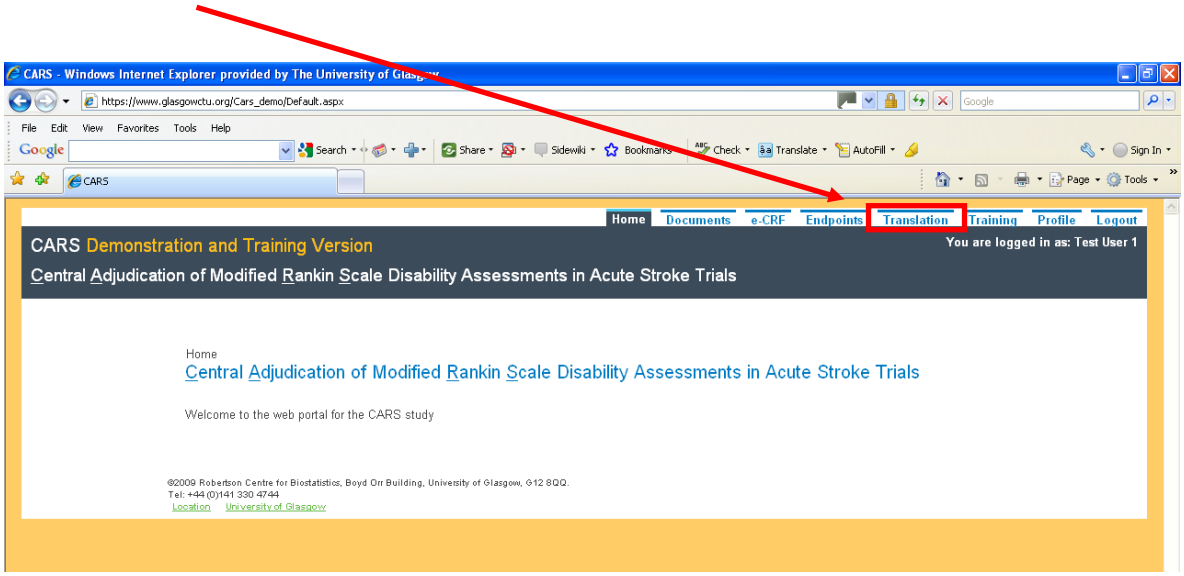
If there are more than two people in the clip they can identified by voice on the dictated file before each person speaks. i.e. "patient says…….", "assessor says…….", "relative says……." throughout. Don't worry it if sounds a bit more disjointed, we expect this and are testing to see if these "translated" clips are still valid despite the added sound file.

The dictated file may end up longer or shorter than the original clip, don't worry about this, the automated file merging system will still dub them together successfully. The scorer may find that the sound continues to play after the video stops or vice versa – this is expected. The sound file will not dub perfectly with each word / sentence. We anticipate that the merged videos will appear disjointed, however visualising the patient / proxy and listening to the content of the interview should allow a score to be allocated.

CARS online portal

The translation portal has been built into the CARS portal which you will have used for the main study.

A new tab has been added in the top right of the screen to direct you to your translation worklist, you will not need to use any other part of the portal for this substudy.



For new login details for the CARS portal please contact the CARS helpdesk at CARS@glasgowctu.org or Kate McArthur at kate.mcarthur@glasgow.ac.uk

There are two roles in the substudy:

Translator:

In this role you will be assigned clips for translation and clips for scoring. You will be asked to enter an mRS score for every clip you view, either as translator or as a committee assessor. You will not be allocated your own translated clips for review.

What you will need:

- Computer with internet access, a USB socket and speakers / headphones
- Login details for the CARS online portal
- Phillips Digital Pocket Memo Dictaphone (LFH9370/5)
- Phillips Speech Exec Dictation Software

Before starting the translation work it is a good idea to have a few practices with the dictaphone and get used to using the controls.

Committee Assessor:

In this role you will be assigned clips for review and scoring only – these will previously have been translated.

What you will need:

- Computer with internet access and speakers / headphones
- Login details for the CARS online portal

Philips Digital Pocket Memo Dictaphone

Contents of the box:

- Dictaphone
- Leather case
- Digital Memory Card
- Batteries
- USB cable
- Speech Exec software CD
- Instruction Manuals

The digital Dictaphone is easy to use.

The pictures below highlight the relevant functions and buttons.

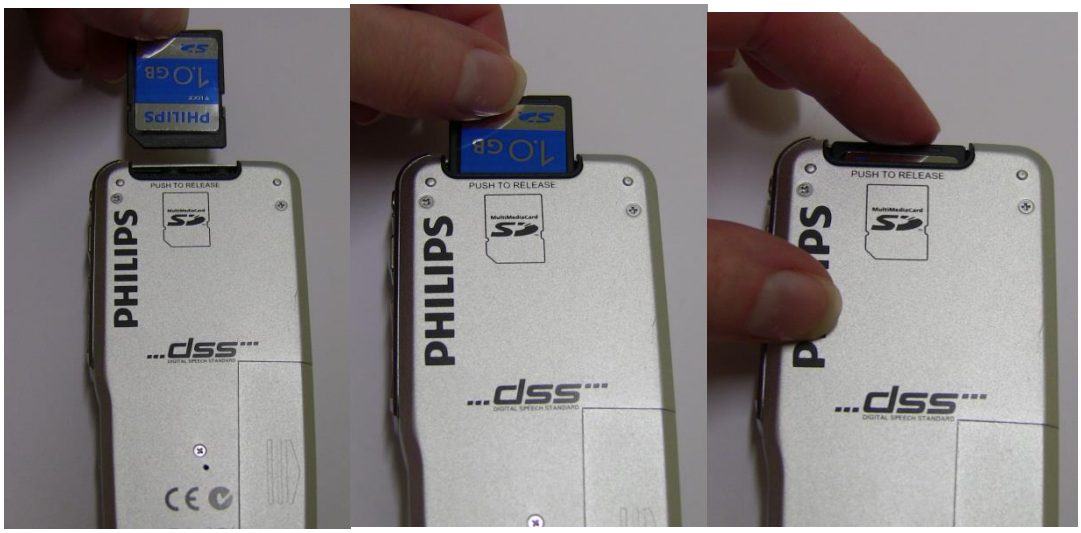
Record Button
Press once to start and once to stop recording

Forward/Rewind
This can be used to rewind and over-write any errors

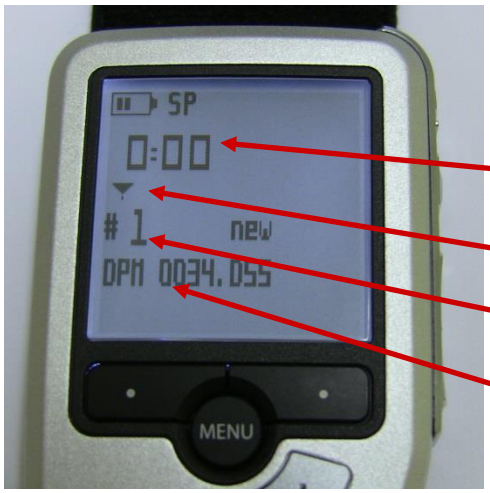
On / Off switch

Getting Started:

Insert the memory card and batteries as shown below.



Recording:



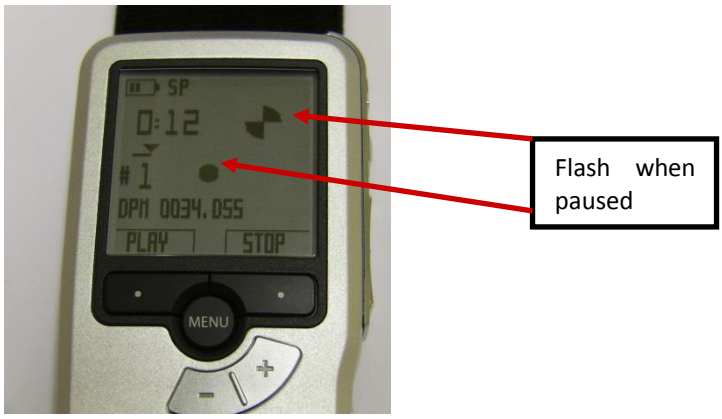
When you switch the dicatphone on, ready to start recording, the screen will look like this.

- Length of recording
- Arrow to denote progress
- File number on the memory card
- File name on memory card

When you press the red record button on the right hand side the recording will begin. The length of your recording will be seen on the clock.

You will know that it is recording due to several indicators.

- Red light on top right corner
- Clockwise spinning wheel
- Solid recording spot
- Play and Stop buttons appear



You can ***pause*** recording by pressing the red record button again.

If you do this the spinning wheel will stop moving and the recording spot will flash to indicate that you have paused.

The timer will pause until you resume recording by again pressing the red record button.

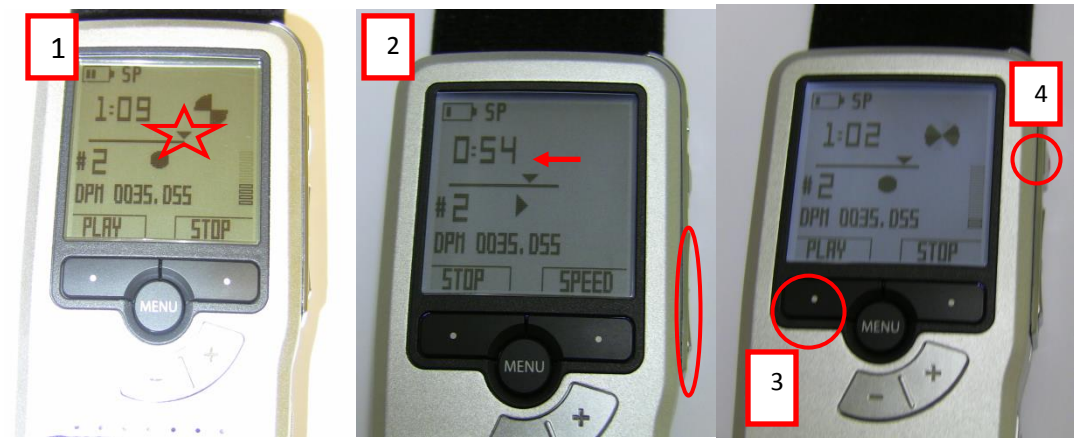
To rewind and listen or overwrite:

This arrow shows where you are in the recording (1)

In the above pictures you see that it can be moved back if you want to rewind and listen to your dictation or record over an error.

To do this press the upper part of the forward/rewind button on the side panel and watch the arrow move backwards. The timer will move backwards to demonstrate how far you have gone. (2)

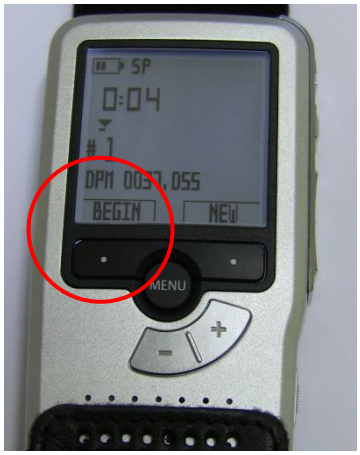
To listen to your dictation press the play button in the bottom left of the screen (3)
To overwrite an error simply press the red record button again and re-recored over it (4)



To finish a recording:

To finish a completed recording press the stop button at the bottom right of the screen.

To listen to a finished recording:



Press the Begin button at the bottom left hand corner of the screen

To start a new recording file:

We recommend that you only dictate one file at a time to avoid confusion. As you cannot rename the files while they are on the Dictaphone there is a risk that the wrong audio file may be merged with the incorrect clip. We would suggest dictating one at a time, uploading it (which will automatically wipe it from the Dictaphone) and then starting a new file for the new clip.

If you have to start a second file for dictation this done as follows:

Press the New button at the bottom right hand corner of the screen

The screen will refresh with a new timer and #2 displayed on the screen

Connecting the Dictaphone to the computer:

This USB cable is contained within the Phillips box

There is a small connector for the Dictaphone and a regular USB connector for the computer

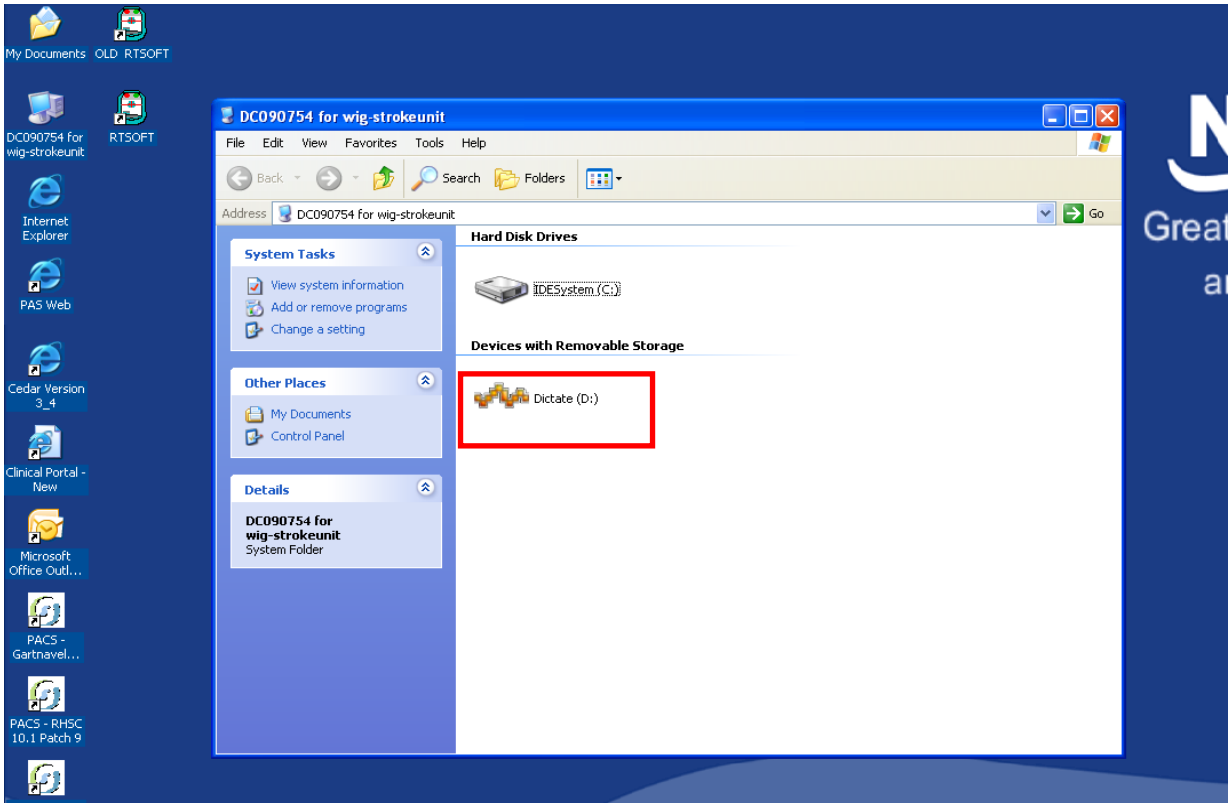
The two ends of the USB cable connect to the devices as shown

When the cable has been inserted correctly the screen on the Dictaphone will demonstrate a successful connection

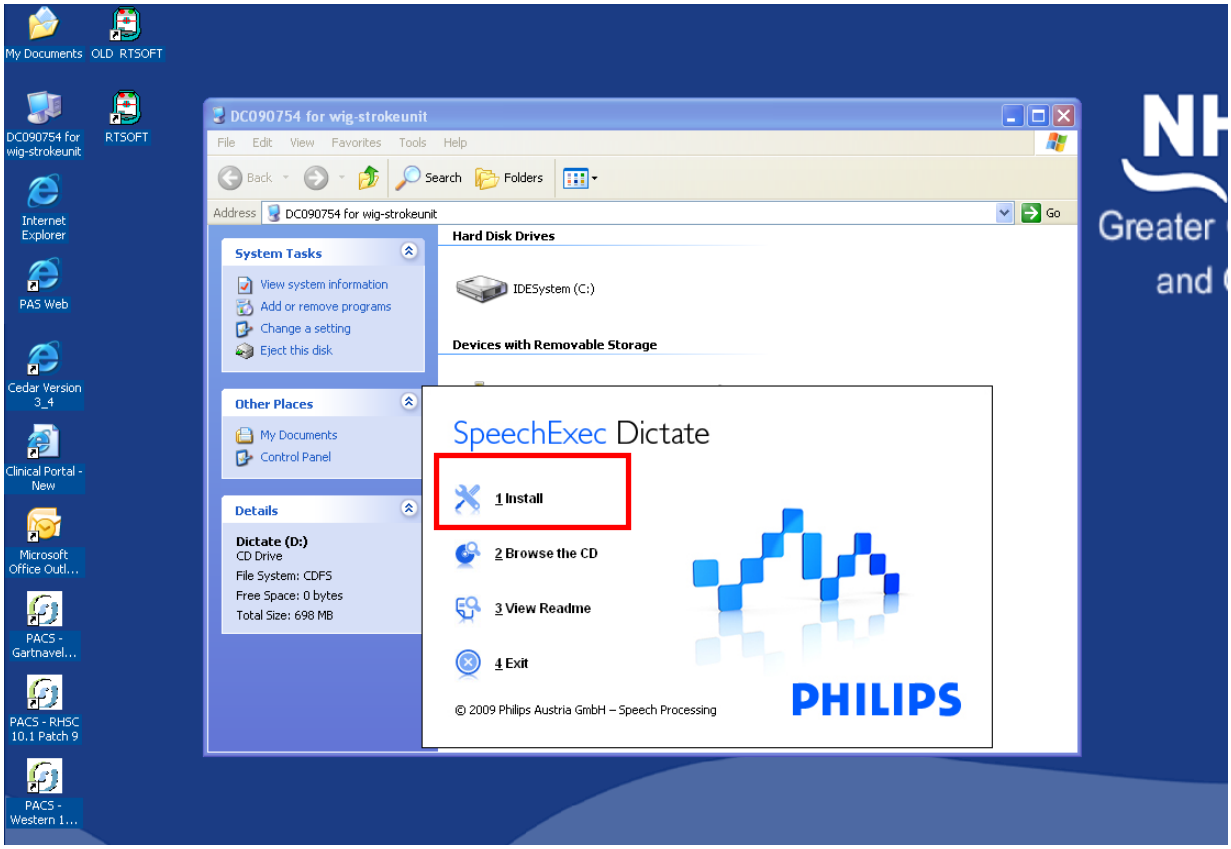
Phillips Speech Exec Software

The SpeechExec software provided with the dictaphone will be either version 5 or version 7. Both work in exactly the same way. Before use the software will need to be installed on your PC. This is straightforward to complete with the following instructions but may need to be sanctioned by the local NHS IT dept.

Enter the SpeechExec disc into the CD drive of your PC. Open “My Computer” to locate the disc on the screen. Click on the Dictate (D:) icon to open software.



Once the software has opened, click on option 1 – Install to start up the installation wizard.

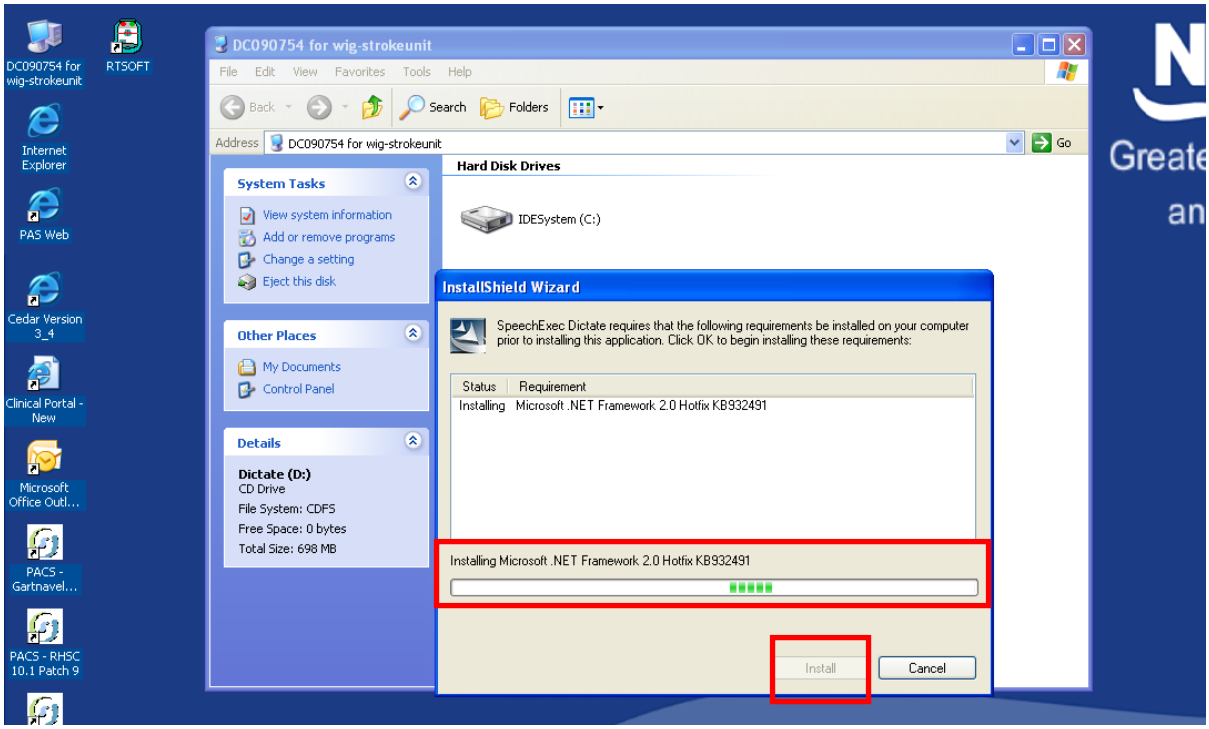


A box will appear requesting that you pick a setup language from the drop down list. Select English (United States).

Unfortunately it doesn't offer a UK English option.

Next click Install at the bottom of the open box to begin the process.

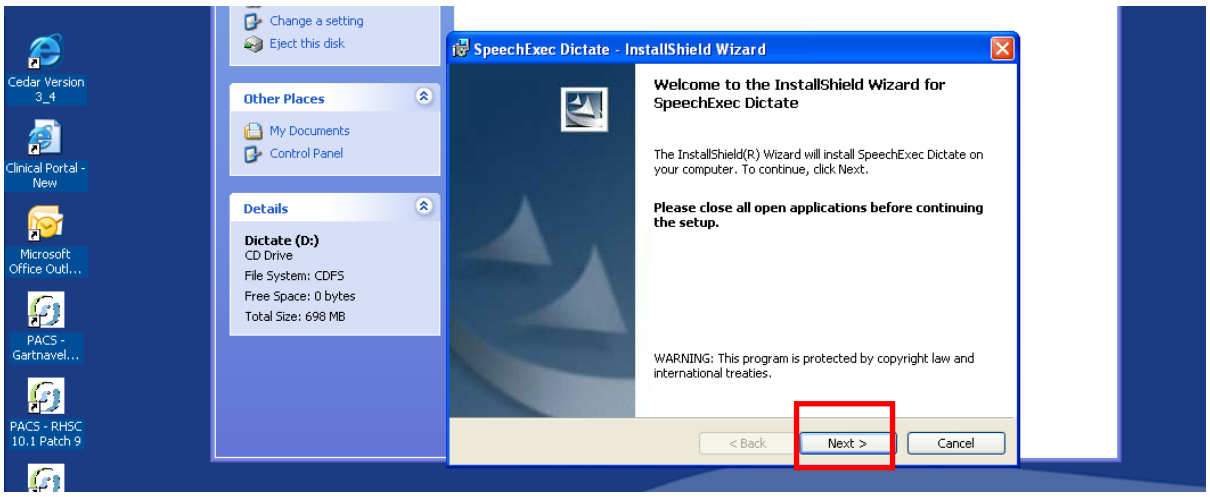
A green progress bar will move along the bottom of the box to indicate that the software is being accessed.



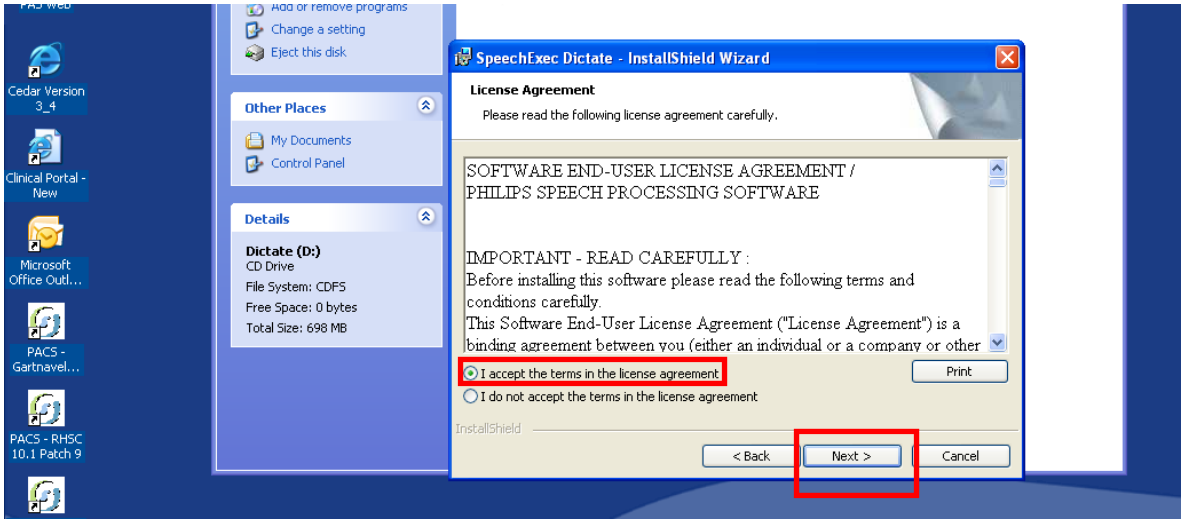
The SpeechExec Dictate – InstalShield Wizard will open once the green progress bar has finished moving.

Close all other open applications (e.g. email inbox / word files) before progressing

Click next to begin installation.

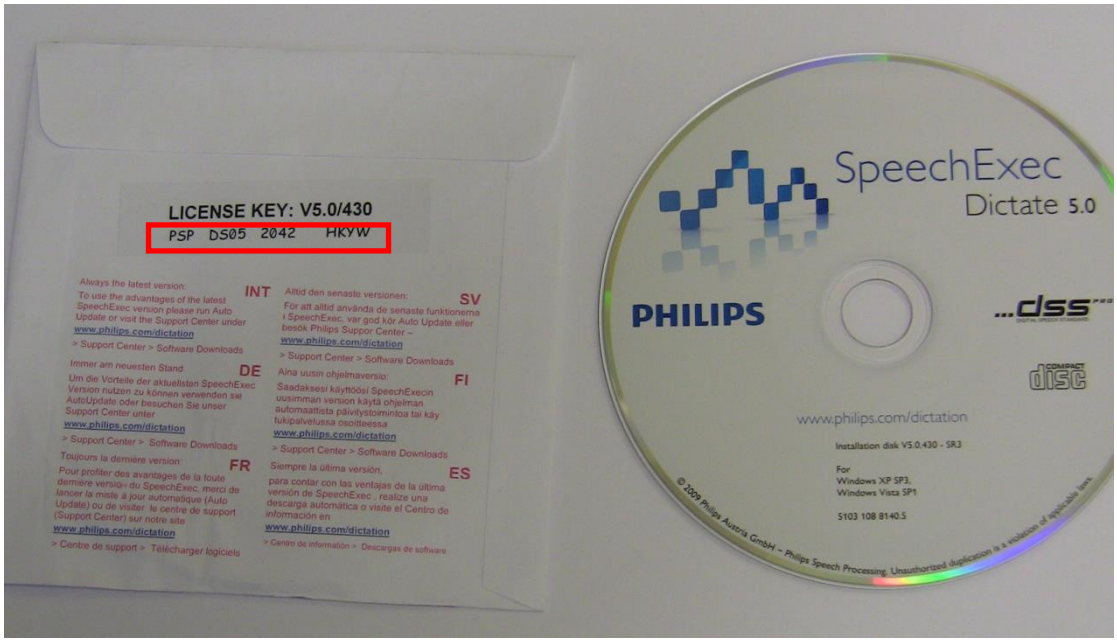


Click to accept the terms of the license agreement and again proceed by clicking Next.

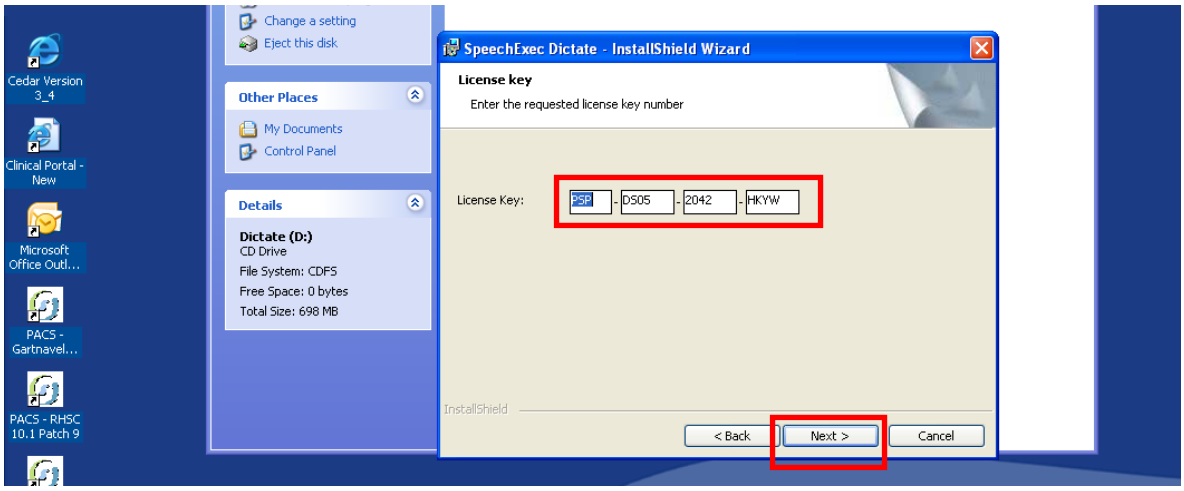


You will now be asked to enter the license key code that is printed on the back of the CD sleeve. In this example the Licence Key is *PSP DS05 2042.*

SpeechExec Version 7 sleeves will have a longer licence key in a similar format.

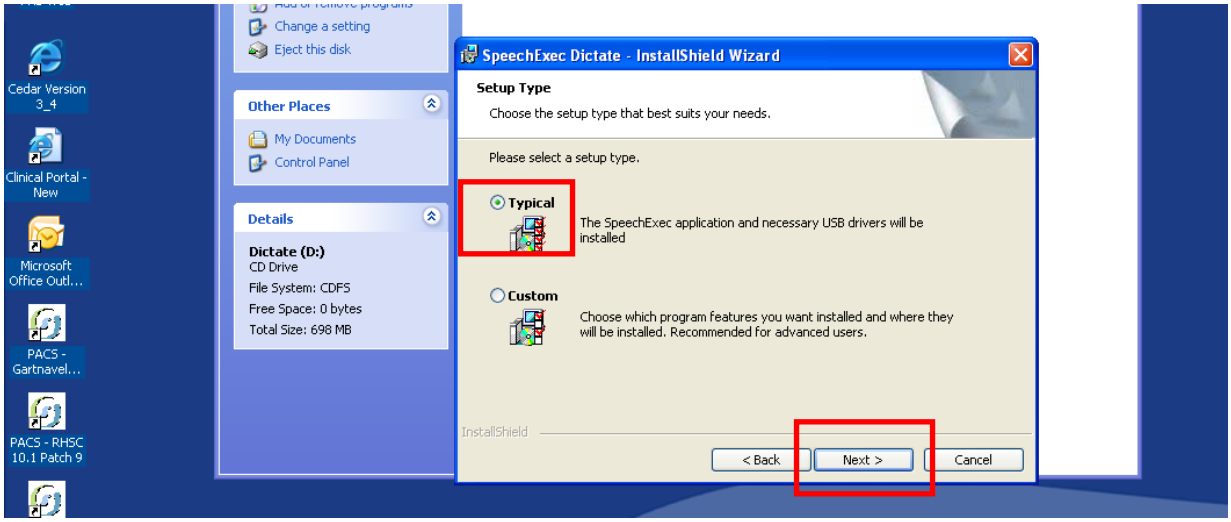


Enter the Licence Key as shown and click Next. These are not case sensitive.

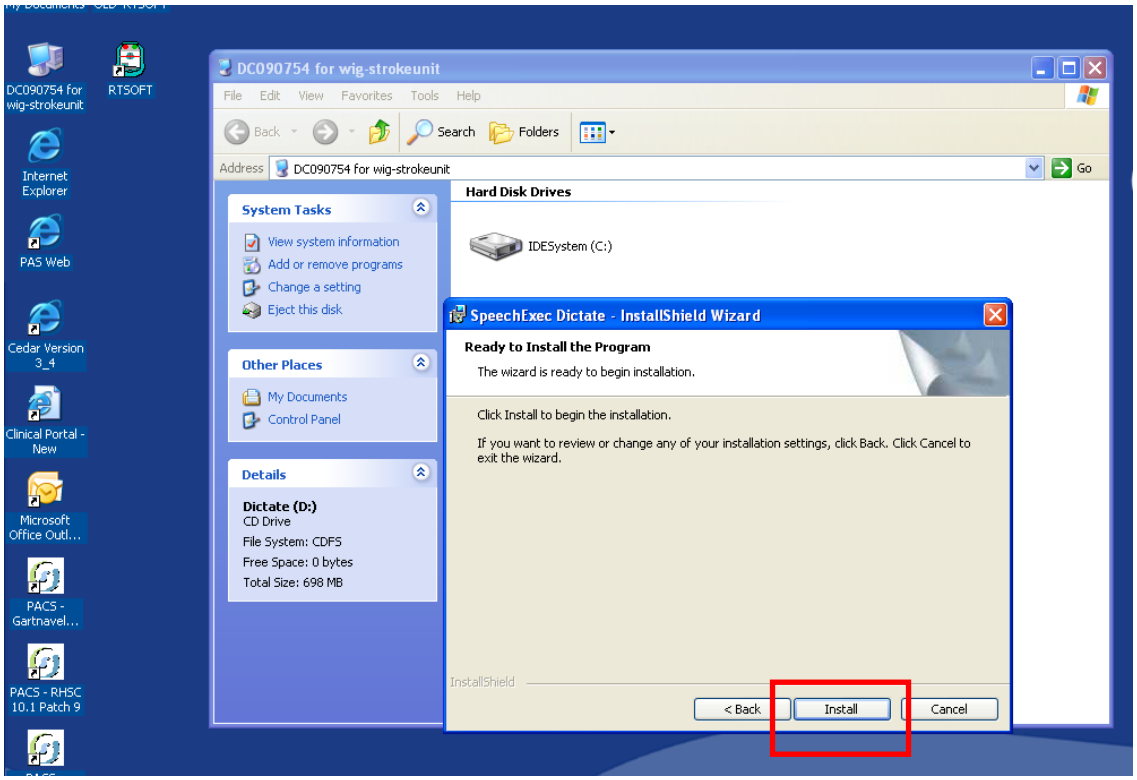


In Version 7 you will be prompted to enter your details. Enter your name but in the institution please enter University of Glasgow as this is where the licence was purchaced. Do not enter an email address, this is unnecessary and will avoid junk email. NB This step is not necessary in Version 5.

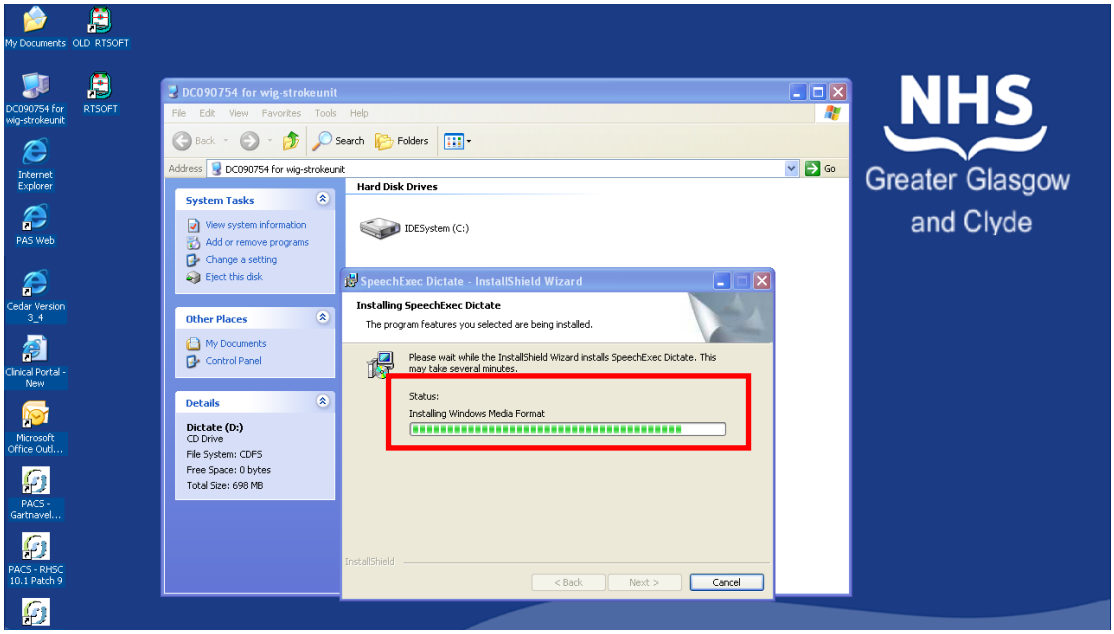
The software will offer you an option to customise your user profile. This is not necessary. Click "Typical" and Next to progress through installation.



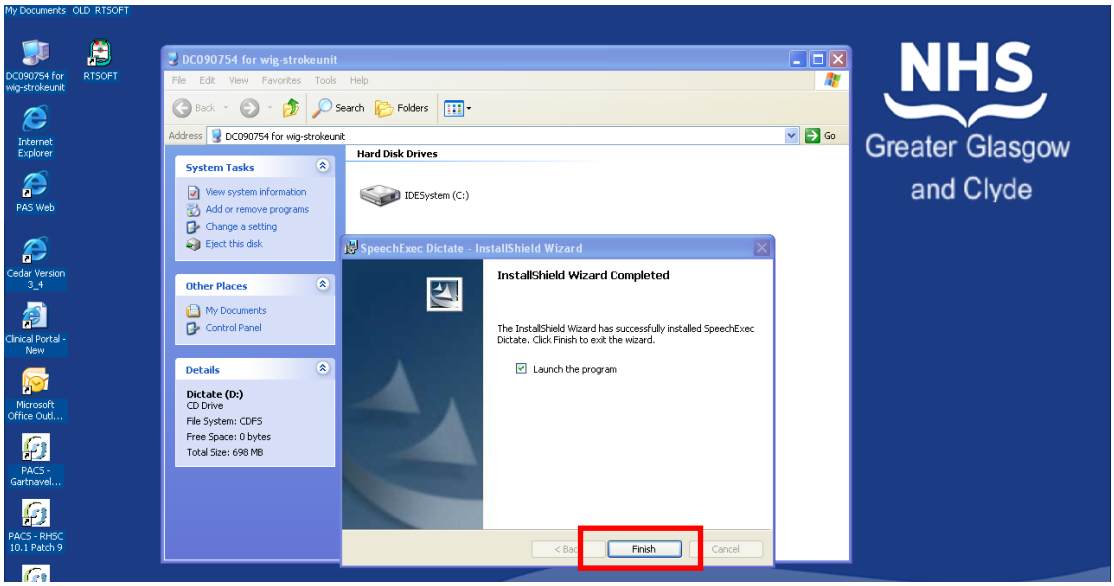
You are now ready to install SpeechExec.



Click Install to begin and again several green progress bars will appear in the installation box to indicate that the software is being installed successfully.

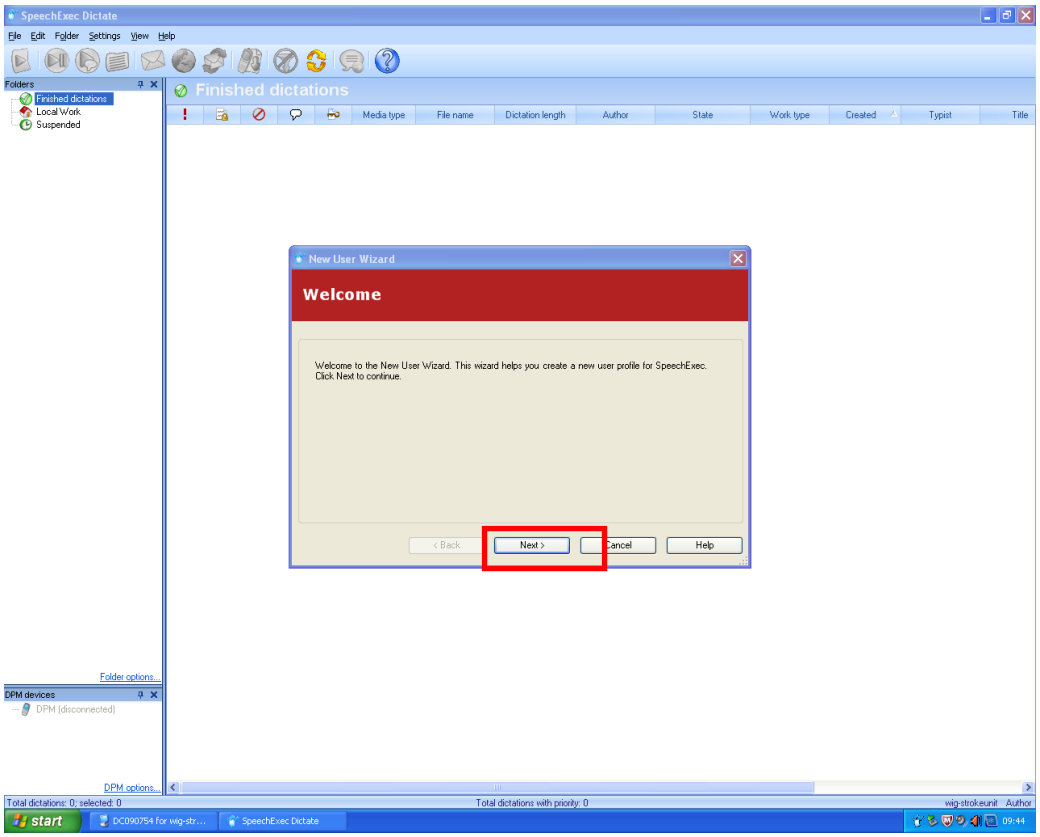


Once complete the following screen will appear to indicate that the process is complete. Click Finish.



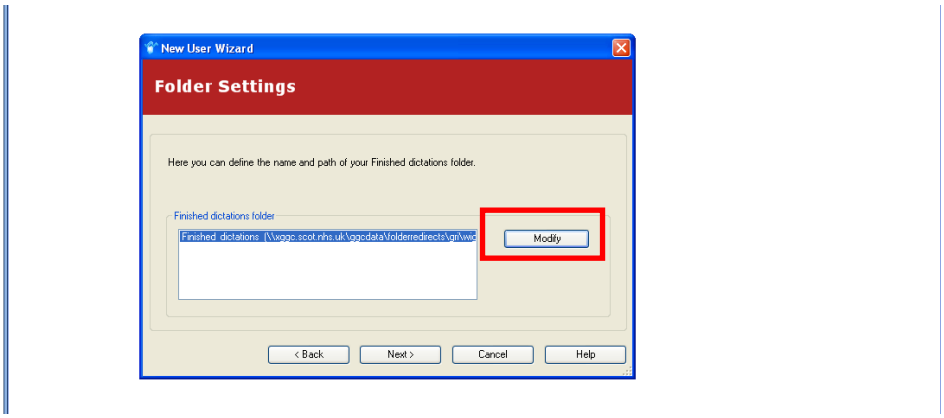
The SpeechExec Software may now open automatically. If not, open the software by clicking on the new icon on your desktop.

You will be prompted to set up your user profile before you begin to use the software. A New User Wizard will open automatically, click Next.



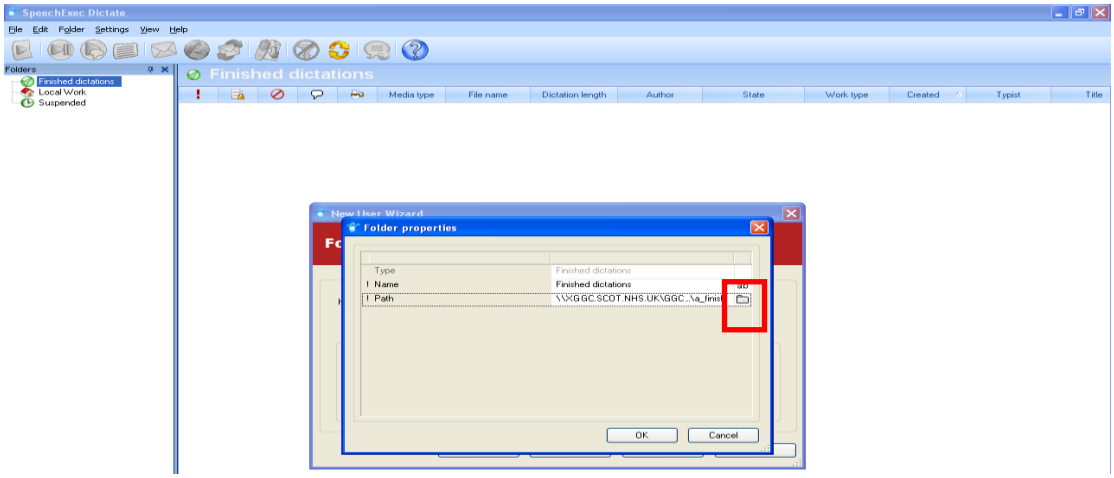
First it will automatically enter your PC username as the SpeechExec user name. This is designed for professional dictation purposes so that typing pools are able to identify where a piece of work has come from. This is less important for our purposes. Accept the automatic username offered and proceed using the next button.

You are then requested to confirm where the completed dictation will be saved in your files. Click on the Modify button.

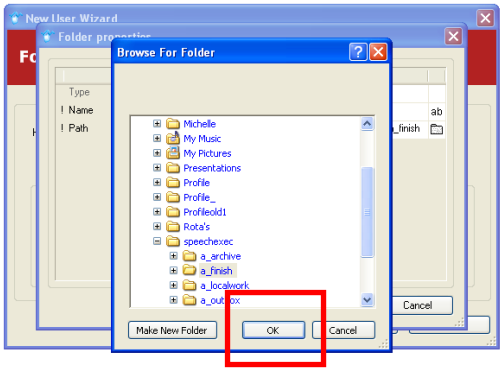


This will open a further box with two fields.

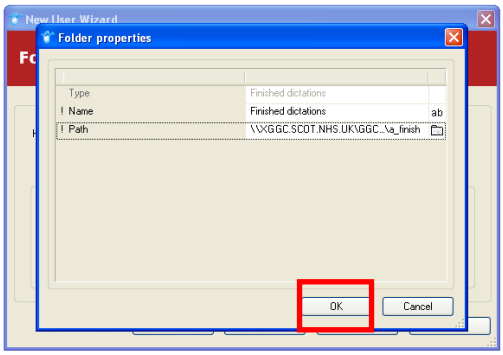
Name: Finished Dications
Path: This will display the location where your completed dictation will be saved when you connect the dictaphone to the PC. Click on the small folder icon as shown below to see where the files are sent.



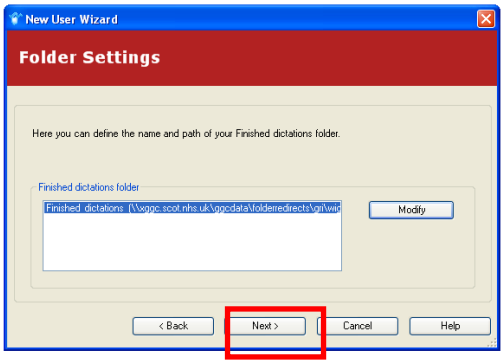
The package will automatically set up a folder in My Documents. We recommend that you accept this folder (unless there is a local IT reason that it is best saved elsewhere) and click OK



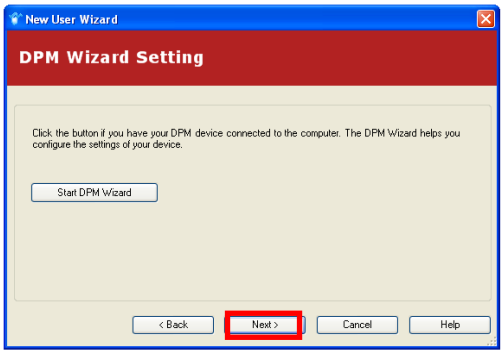
This will bring you back to the previous box, click OK to proceed.



Finally you will be returned to the Folder Settings wizard box, click Next to continue.

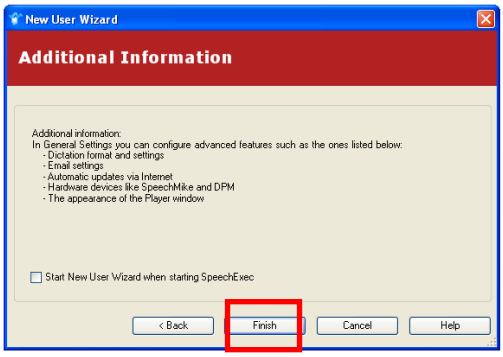


The wizard will then offer the option of setting up the device, this is not necessary so simply click “next” to move on.



Do not check the box in the Additional Information section.

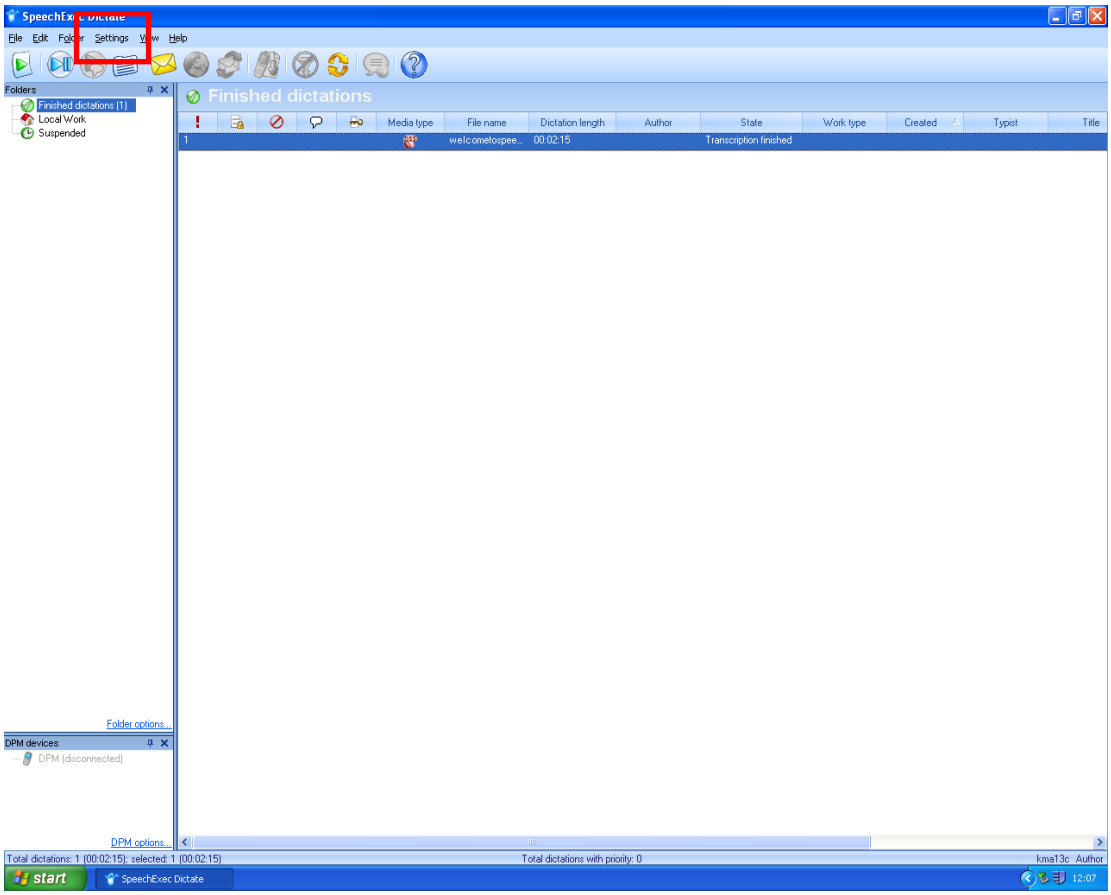
Next click Finish to complete the process.



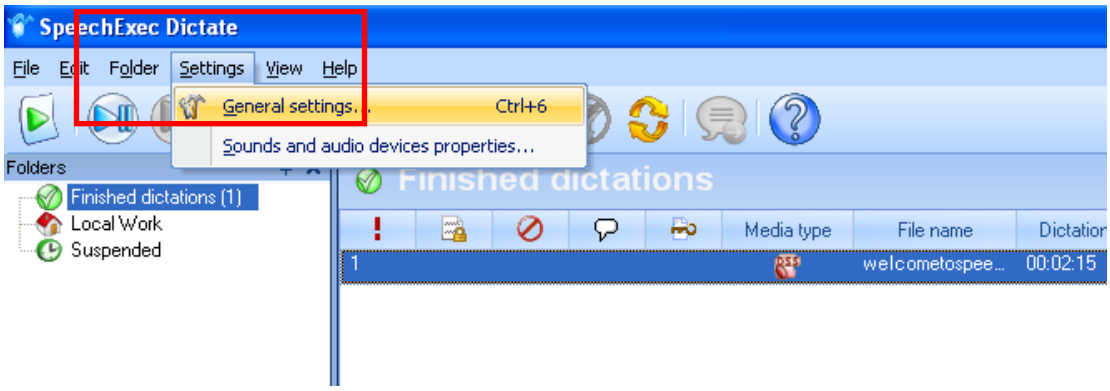
There is one change necessary to the settings of the SpeechExec software for use in the CARS translation portal.

For the portal to accept the files the audio format requires to be changed, this can be configured through the settings menu at this stage so that each time you upload an audio file there is a one click conversion process.

Open the SpeechExec software via the start menu or the desktop icon.



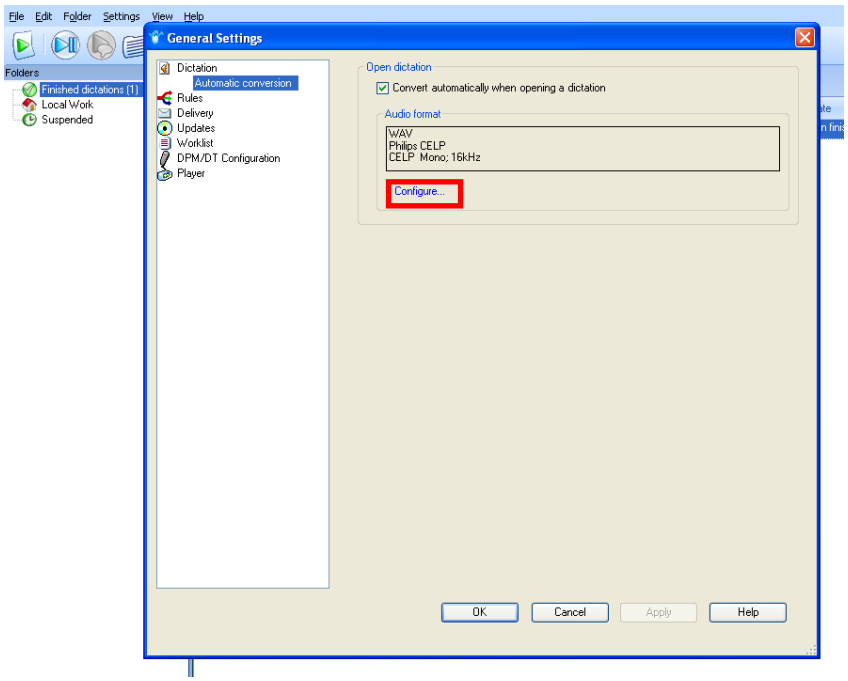
Open the “Settings” Menu and click on “General Settings”



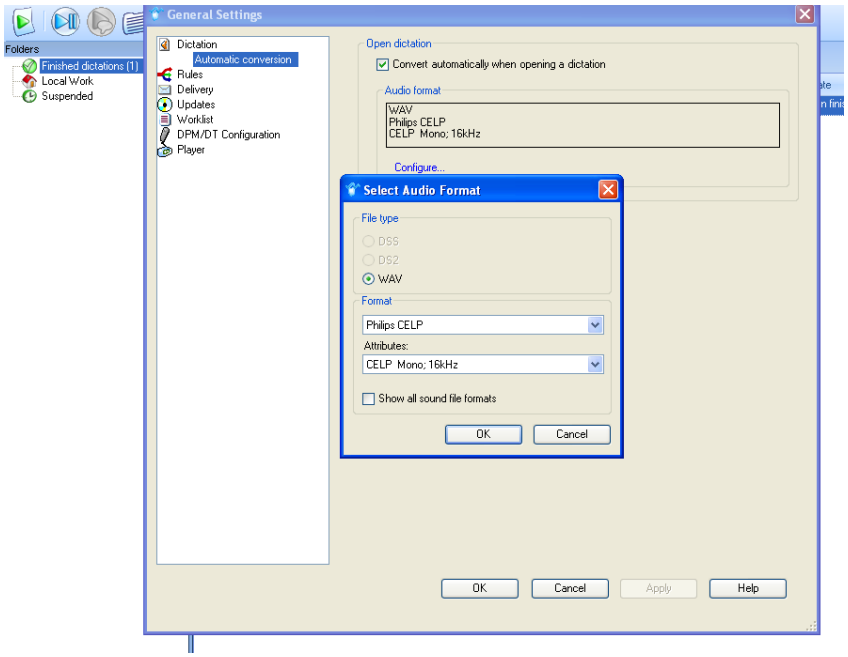
This will open a “General Settings” box which displays the audio format of the uploaded dictations.

The format should read:

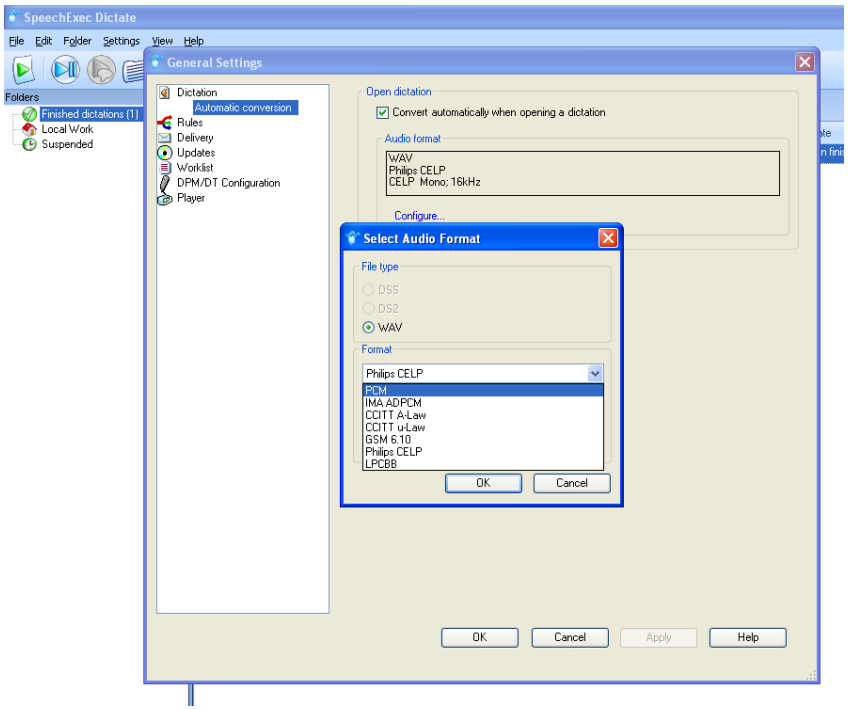
WAV Philips CELP CELP Mono, 16kHz



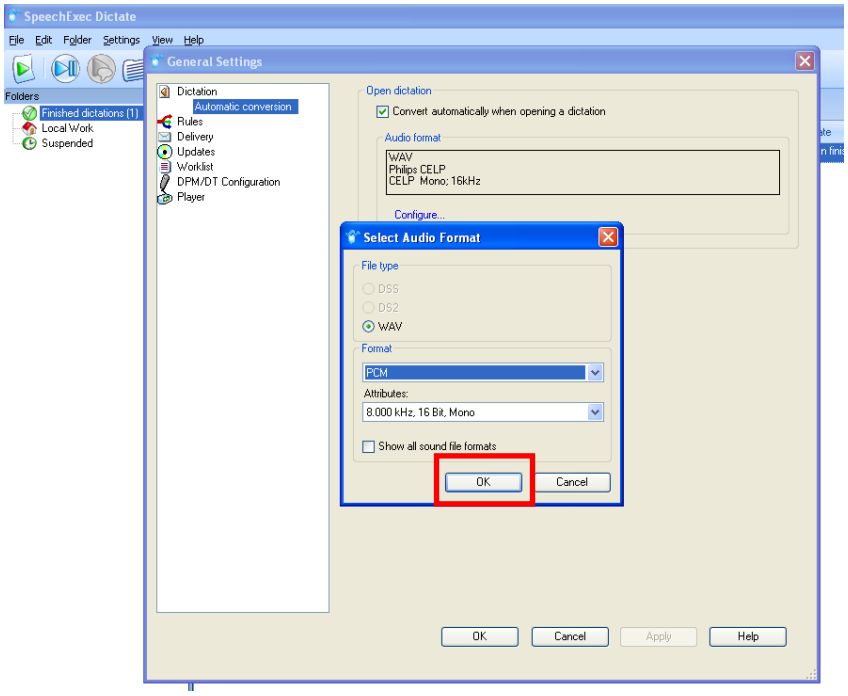
To change this click on “Configure”. A new box will appear as shown.



Click on the “format” dropdown menu and select PCM.



Then click OK



This will change the description of the Audio Format to read:

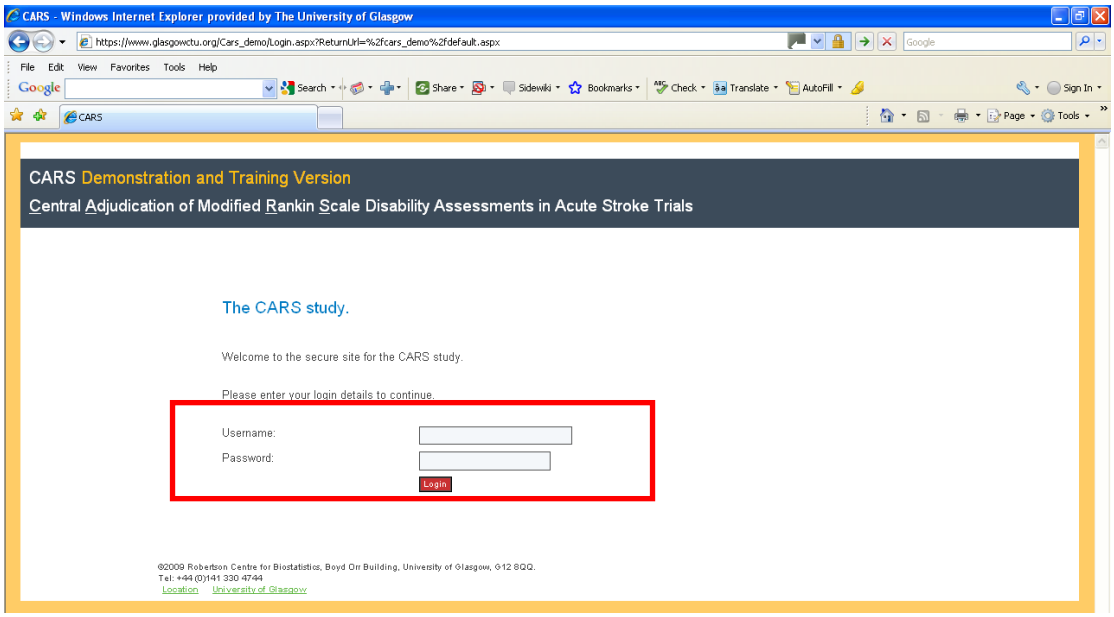
WAV PCM 8.000kHz, 16 bit, Mono

Ensure that the “Convert automatically when opening a dictation” is checked.
Click OK to complete the change in settings.

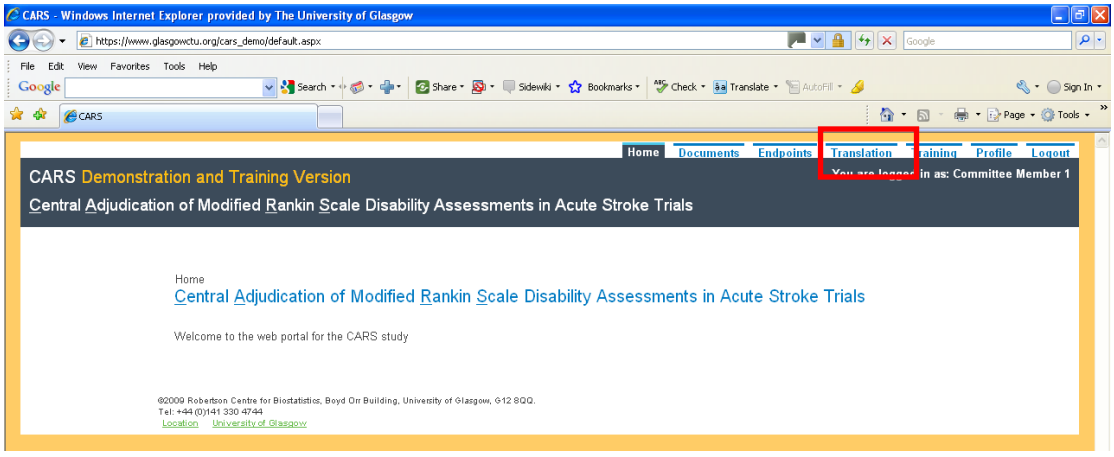
Using the CARS portal in the “Translator” role

When you are allocated a clip for translation you will receive an automated email to prompt you to enter the portal.

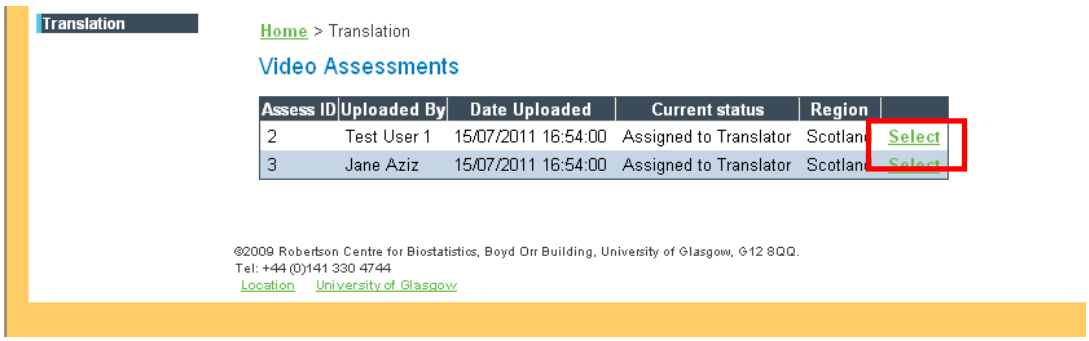
Login using your username and password to enter the portal



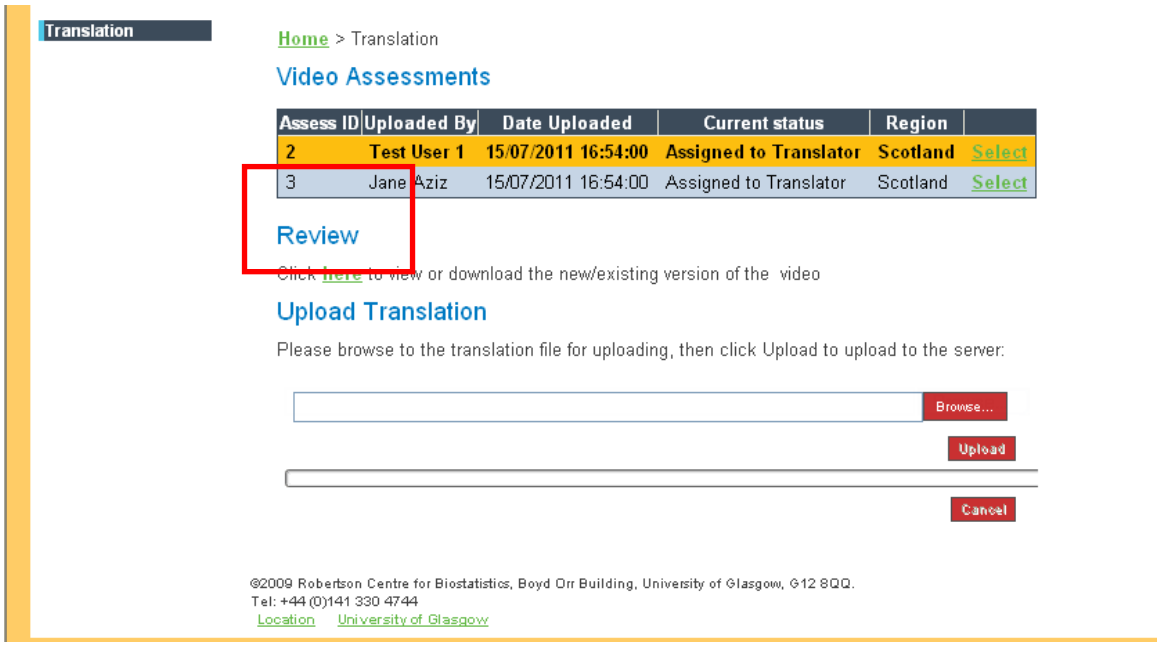
Access the translation section of the portal using the tab at the top right of the screen



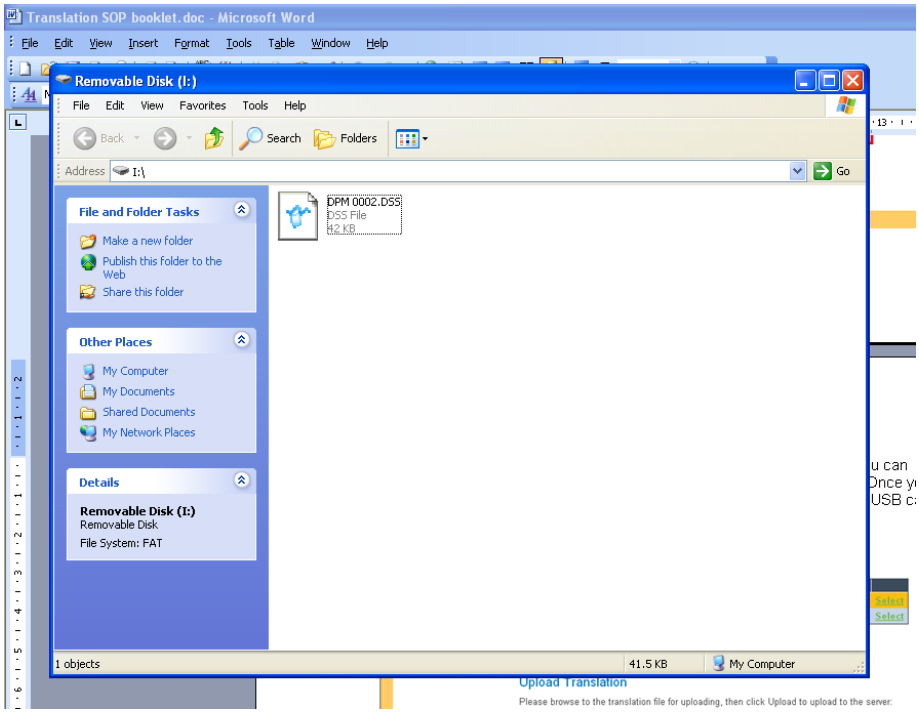
You will see a list of clips that are ready for translation. Each has a number to identify it (Assess ID) at the left of the list. Click on “select” to start with each clip.



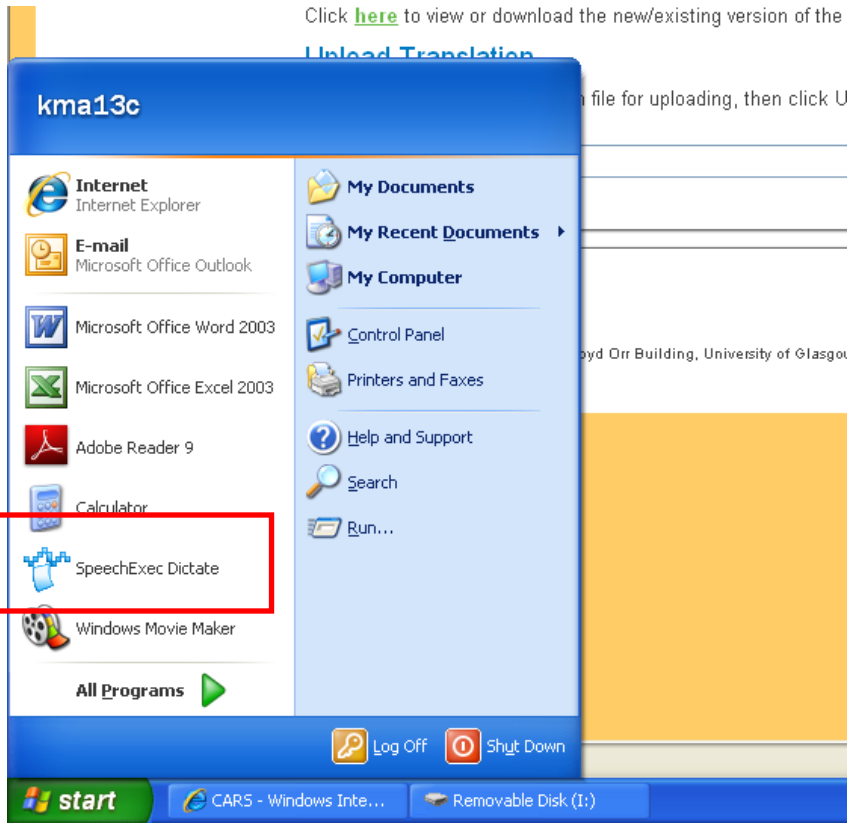
The selected clip will turn yellow in the list. Under the heading "review" you can download the video to watch it. Click "open" to open the video file in windows movie maker. Do not "save" the file to your computer due to data protection rules. While watching remember that you will need to enter a mRS score for the clip. While watching the clip use the dictaphone to provide a dictated translation file (see page 4). For details in the use of the Philips PocketMemo see page 6.



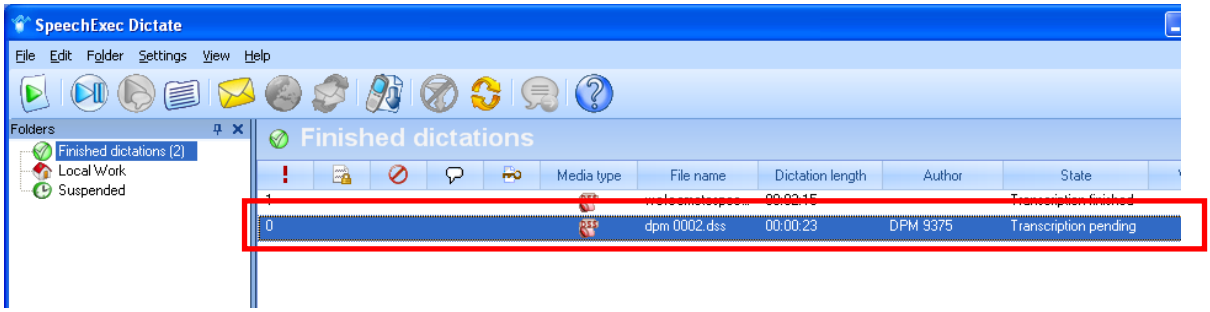
Once you have completed the dictation, connect the dictaphone to the PC using the USB cable as shown on page 11. Depending upon the set up of your PC this may or may not automatically be recognised by your PC as a removable disc (similar to a memory stick) and the folder may appear on the screen. It may not appear – this is not a problem as you will access it through SpeechExec instead.



Use the “start” menu to open the SpeechExec software. If it doesn’t appear in the initial list you will be able to locate it via the “All Programs” list.

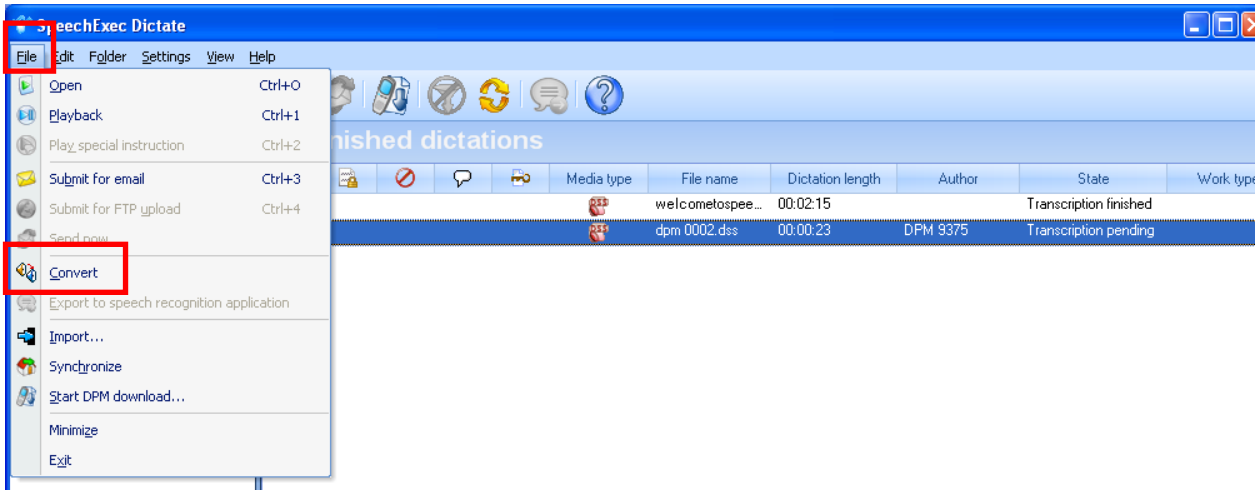


The new dictation file will automatically be downloaded into SpeechExec and will be seen on the list as a .dds file. This is seen below, highlighted in blue. Download to the computer will delete the file from the dictaphone so that there is no confusion when you start to dictate the next file.

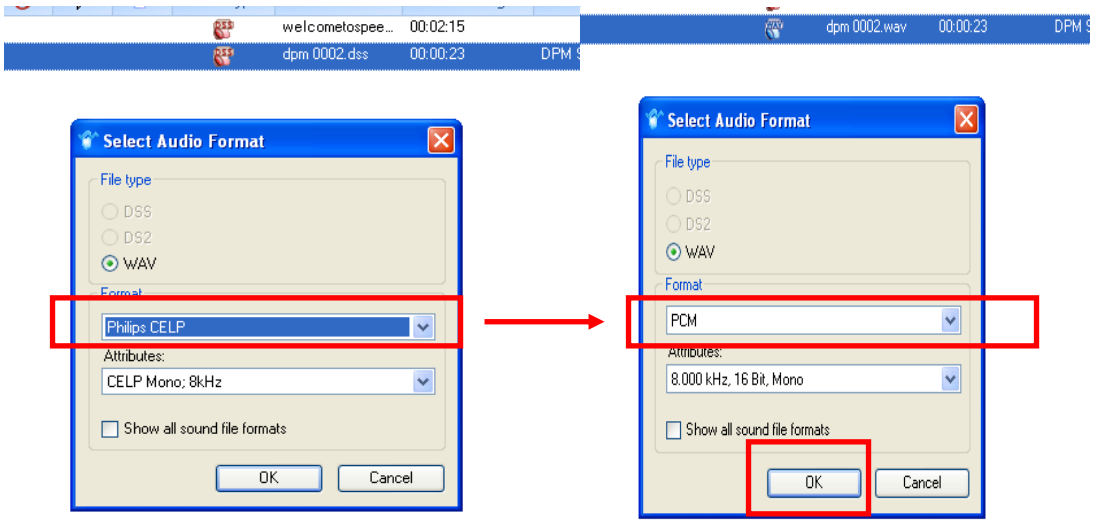


This .dds file needs to be converted to a .wav file before it can be uploaded to the portal. If you have set up the configuration of the software as shown on pages 23-25 then this will happen automatically when you double click on the highlighted dictation file to open it (the conversion will change the file ending from .dds to .wav and it changes colour.

If you haven't configured the software you can easily convert the file manually as follows. Open the File menu and select "convert" as shown below.



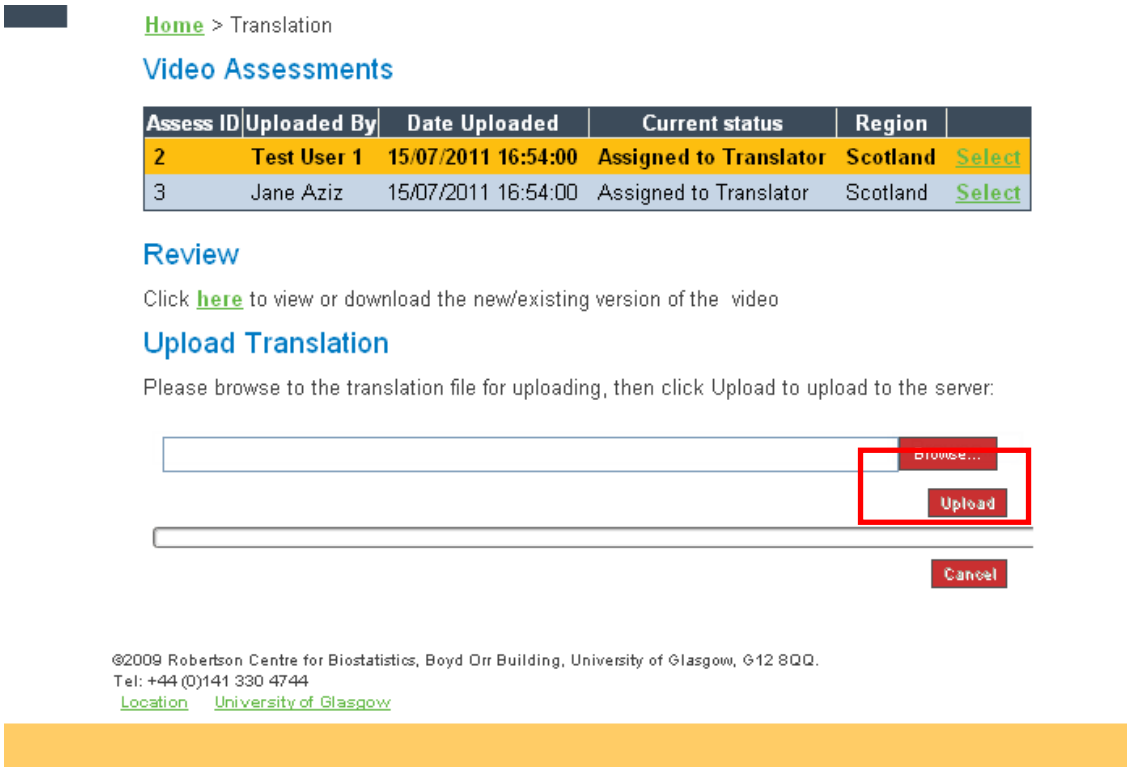
The options for audio format will be displayed, one change is required. The dropdown menu which displays "Philips CELP" should be changed to "PCM", then click "OK".



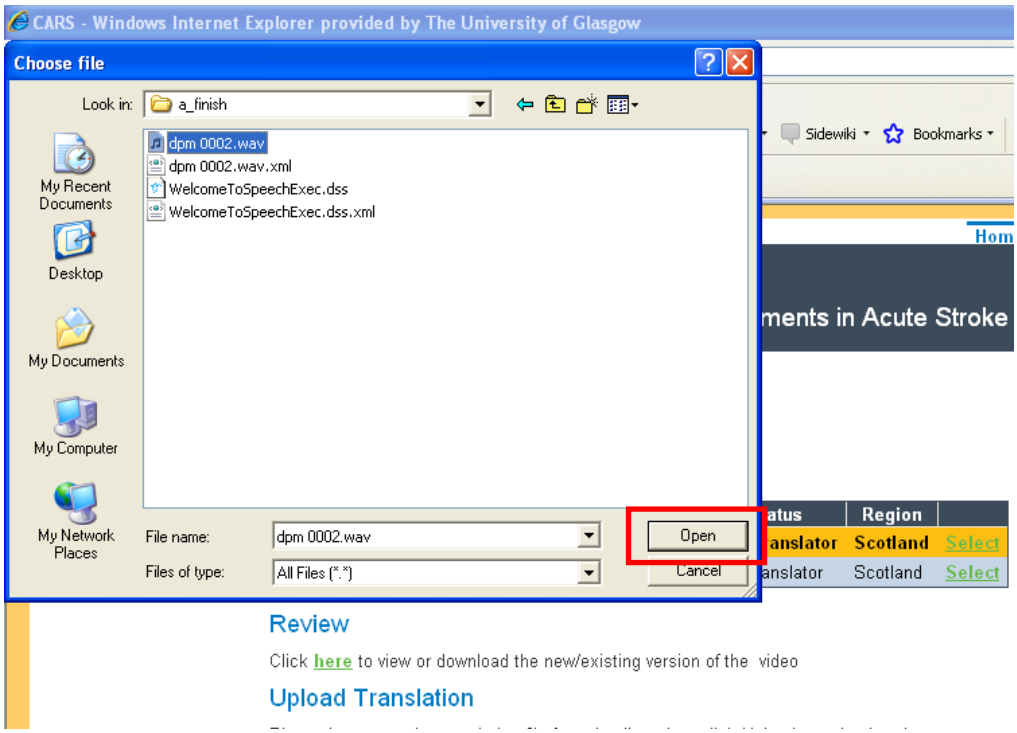
The conversion will change the file to .wav, ready for upload to the portal.

Next return to the CARS portal, where the clip you have dictated a translation for should remain highlighted. If the connection to the portal has timed out (this happens after 30 minutes of inactivity) – log back in and be sure to highlight the same Assess ID to ensure the correct translation goes with the correct clip, you may wish to watch a little bit of the clip again to ensure it is the right one.

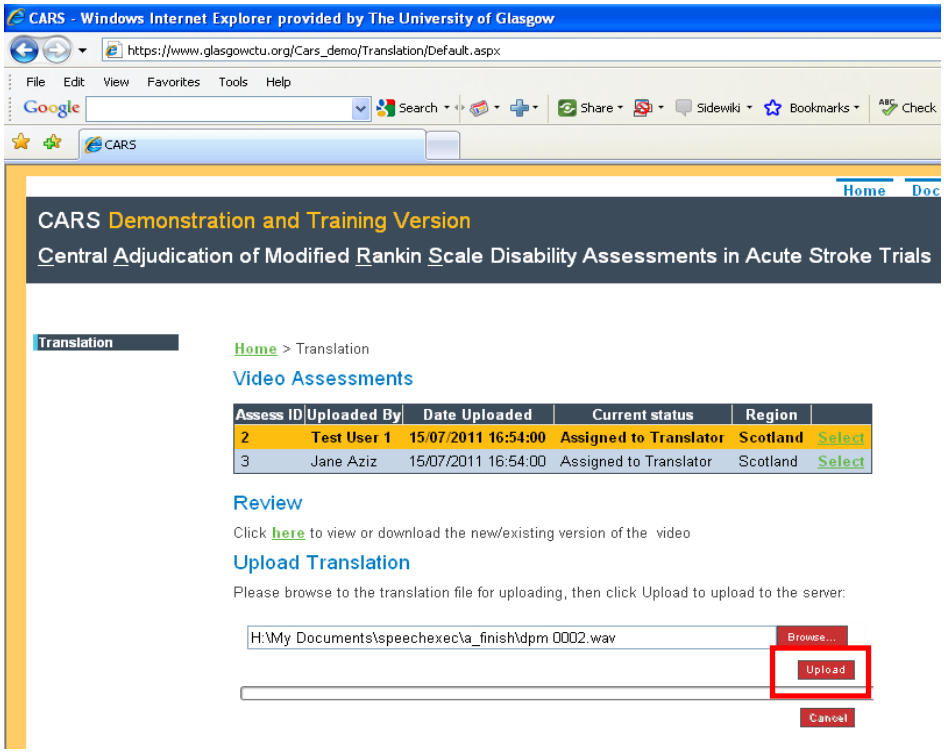
Once you have the clip highlighted, clip on the "Browse" button to locate the dictated file. This is exactly the same as locating an attachment for an email from your computer files.



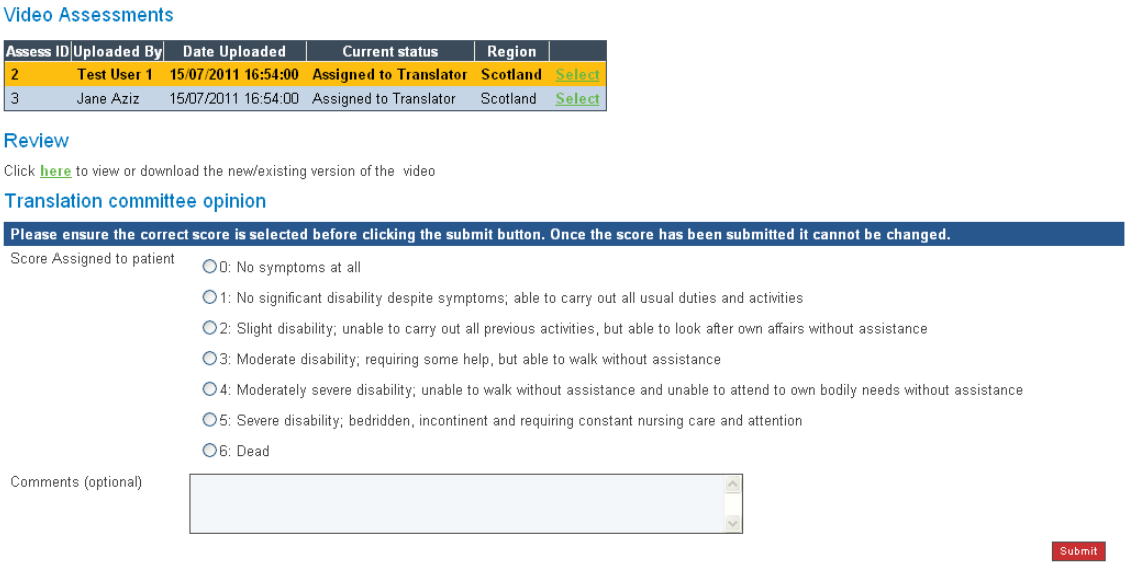
Locate the file in “My Documents” >> “speechexec” >> “a_finish” where your dictation file will be saved. There will be two versions of your file (ignore the Welcome to SpeechExec files). One version is the sound file .wav (identified by the musical note) and one is not for use .wav.xml (ignore this file). Highlight the file with the musical note and click “Open”



This will enter the file in the upload box, click “Upload” in the portal webpage and the file will be uploaded and automatically merged with the original video clip.



A green progress bar will demonstrate the upload / file merge progress. Do not click anything further until the mRS score sheet appears on the screen – this may take a few minutes. Enter an mRS score and any comments you might have. Click "Submit" to save your score.

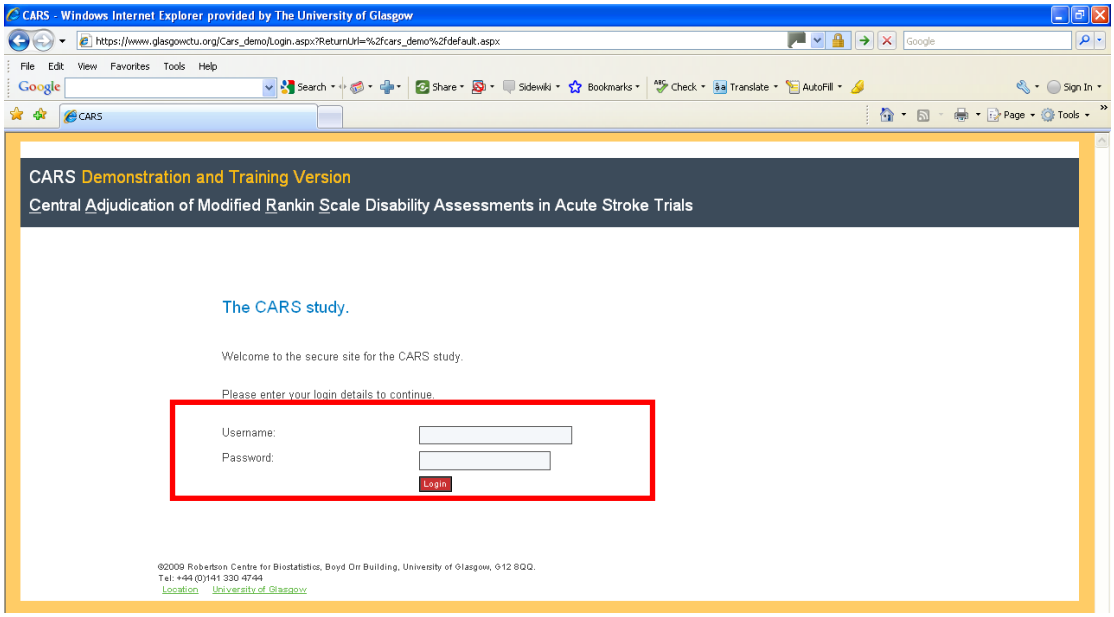


The clip will then be sent back to the co-ordinating centre for distribution amongst the other reviewers.

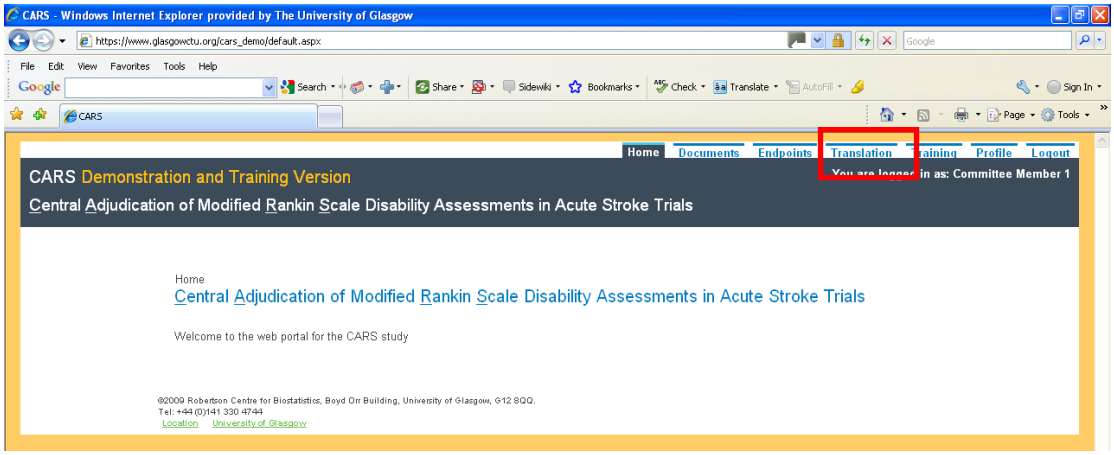
Using the CARS portal in the "Committee Assessor" role

When you are allocated a clip for review you will receive an automated email to prompt you to enter the portal.

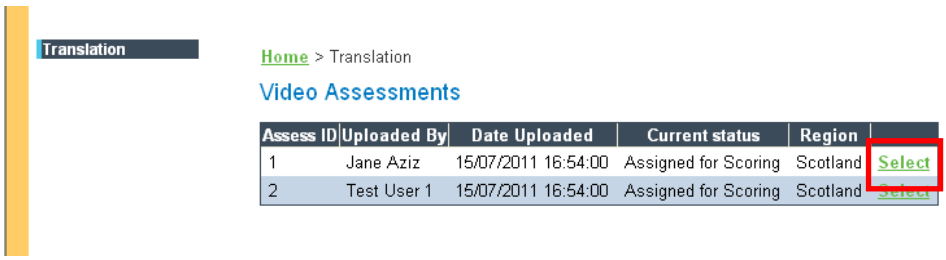
Login using your username and password to enter the portal



Access the translation section of the portal using the tab at the top right of the screen

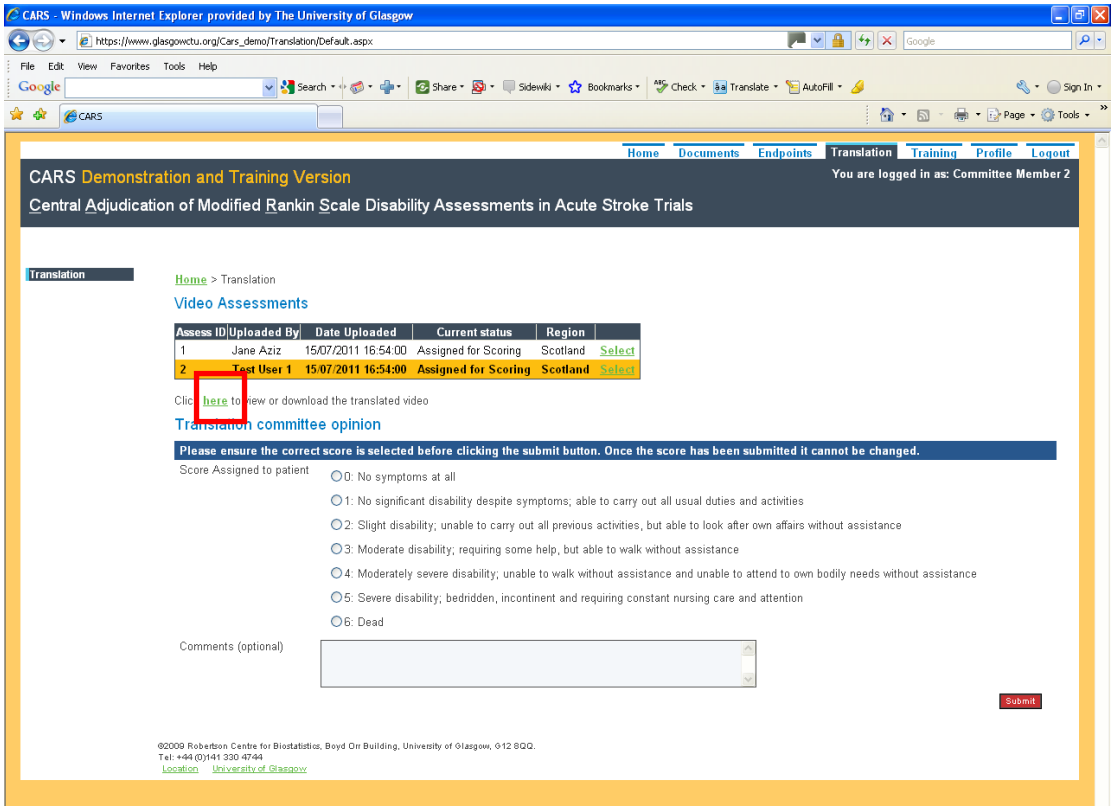


You will see a list of clips that are ready for review. Each has a number to identify it (Assess ID) at the left of the list. Click on "select" to start with each clip.



The selected clip will be highlighted in yellow. Click as indicated to download and view the translated video file. Click "open" to open the file in windows movie maker. Do not "save" the file to your computer due to data protection rules.

Enter your mRS score and any comments you might have in the scoring sheet below. Click "Submit" to save your score.



To review the next clip click "select" to highlight another clip and repeat the process.

Contacts:

Study co-ordinator kate.mcarthur@glasgow.ac.uk

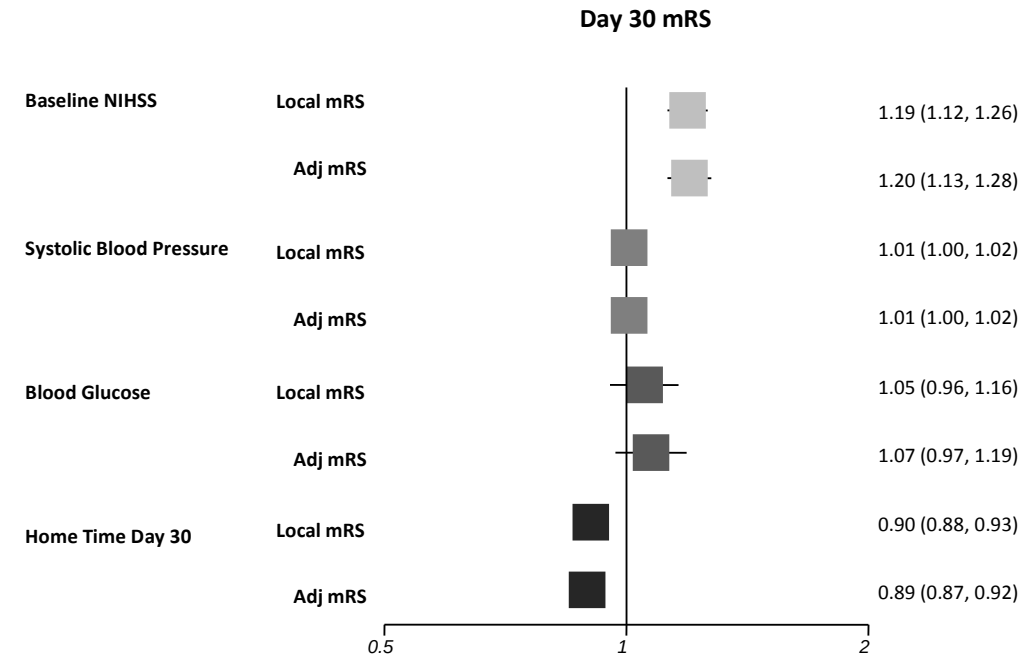
CARS helpdesk CARS@glasgowctu.org

Appendix C

Supplementary results of Validity Analyses

Table 41 - Adjusted proportional odds logistic regression of relationship between bNIHSS, SBP, blood glucose and home time at 30 days with each method of mRS assessment. Day 30 and 90 mRS

	Odds Ratio (OR)	95% CI for OR	p
Day 30 Local mRS			
Baseline NIHSS	1.191	1.124 – 1.263	<0.0001
Systolic Blood Pressure	1.008	1.000 – 1.016	0.046
Blood Glucose	1.053	0.955 – 1.161	0.030
Home Time Day 30	0.902	0.878 – 0.927	<0.0001
Day 30 Adjudicated mRS			
Baseline NIHSS	1.198	1.125 – 1.276	<0.0001
Systolic Blood Pressure	1.008	0.999 – 1.016	0.083
Blood Glucose	1.073	0.969 – 1.189	0.177
Home Time Day 30	0.894	0.868 – 0.920	<0.0001
Day 90 Local mRS			
Baseline NIHSS	1.151	1.058 – 1.220	<0.0001
Systolic Blood Pressure	1.011	1.002 – 1.019	0.013
Blood Glucose	1.082	0.979 – 1.196	0.123
Home Time Day 30	0.913	0.888 – 0.938	<0.0001
Day 90 Adjudicated mRS			
Baseline NIHSS	1.148	1.075 – 1.226	<0.0001
Systolic Blood Pressure	1.010	1.001 – 1.019	0.037
Blood Glucose	1.008	0.904 – 1.124	0.890
Home Time Day 30	0.890	0.862 – 0.918	<0.0001



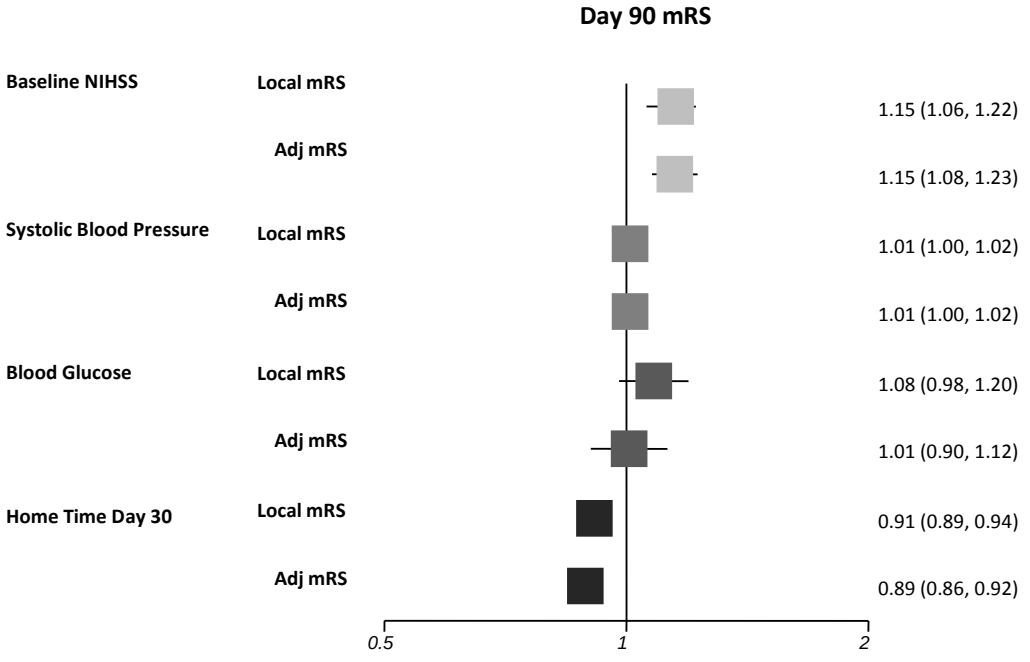


Figure 51 - Adjusted proportional odds logistic regression of relationship between bNIHSS, SBP, blood glucose and home time at 30 days with each method of mRS assessment. Day 30 and 90 mRS. Odds Ratio (95% CI)

Appendix D The CARS Investigators

Centre	PI	Co-Investigators	Research Nurse
Western Infirmary Glasgow G11 6NT	Professor Kennedy R Lees	Dr M Walters Dr J Dawson	Elizabeth Colquhoun Belinda Manak Lesley Campbell
Glasgow Royal Infirmary Castle Street Glasgow G4 0SF	Professor Peter Langhorne		Ruth Graham
Stobhill General Hospital 133 Balornock Road Glasgow G21 3UW	Dr Christine McAlpine		Ruth Graham
Aberdeen Royal Infirmary Foresterhill Aberdeen AB25 2ZN	Dr Mary J MacLeod	Dr Steve Wilkinson	Maggie Bruce A Joyson Michelle Kemp
Ninewells Hospital Dundee DD1 9SY	Dr R MacWalter		Angela Kelly Mairi Stirling

Southern General Hospital Glasgow	Professor Keith Muir		Angela Welsh Wilma Smith Sally Baird
Countess of Chester Foundation Trust Liverpool Road Chester Cheshire CH2 1UL	Dr K Chatterjee		Christine Kelly Helen Eccleston
Cumberland Infirmary Cumberland Infirmary Newtown Road Carlisle CA2 7HY	Dr Paul Davies		Lisa Armstrong Claire Hagon
Royal Glamorgan Hospital Ynysmaerdy Rhondda Cynon Taf CF72 8XR Swansea	Dr Richard Dewar	Dr Senthill Raghunathan	Margo Wigley Allison Cooper
Royal Devon & Exeter Hospital Barrack Road Exeter EX2 5DW	Dr Martin James		Nicola Wedge Leigh Barron Angela Bowring Julie Cageao Hayley Eastwood Sophie Thomas

York Hospitals NHS Trust York YO31 8HE	Dr John Coyle		Michael Keeling Christopher Rhymes Ina James
Mid Yorkshire Hospitals NHS Trust Wakefield (Pinderfields) Hospital	Dr Michael Carpenter	Dr Datta	Ann Needle Gavin Bateman Karen Mallinder
Harrogate District General Hospital Lancaster Park Road, Harrogate North Yorkshire HG2 7SX	Dr Sean Brotheridge		Jacqueline Strover
Leeds General Infirmary Great George Streer Leeds West Yorkshire LS1 3EX	Dr Peter Wanklyn		Ruth Bellfield Claire Coulson Linetty Mandizvida

Reference List

(1) Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. *Nat Rev Drug Discov* 2004 May;3(5):417-29.

(2) Writing Group, Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB, Bravata DM, Dai S, Ford ES, Fox CS, Fullerton HJ, Gillespie C, Hailpern SM, Heit JA, Howard VJ, Kissela BM, Kittner SJ, Lackland DT, Lichtman JH, Lisabeth LD, Makuc DM, Marcus GM, Marelli A, Matchar DB, Moy CS, Mozaffarian D, Mussolino ME, Nichol G, Paynter NP, Soliman EZ, Sorlie PD, Sotoodehnia N, Turan TN, Virani SS, Wong ND, Woo D, Turner MB. Heart Disease and Stroke Statistics 2012 Update: A Report From the American Heart Association. *Circulation* 2012 January 3;125(1):e2-e220.

(3) Mackay J, Mensah G. Atlas of Heart Disease and Stroke: Global Burden of Stroke / Deaths from Stroke. World Health Organisation; 2004.

(4) National Audit Office. Reducing Brain Damage: Faster access to better stroke care. 2005.

(5) Pendlebury ST, Rothwell PM, Algra A, Ariesen MJ, Bakac G, Czlonkowska A, Dachenhausen A, Krespi Y, Korv J, Krolikowski K, Kulesh S, Michel P, Thomassen L, Bogousslavsky J, Brainin M. Underfunding of Stroke Research: A Europe-Wide Problem. *Stroke* 2004 October 1;35(10):2368-71.

(6) Kidwell CS, Liebeskind DS, Starkman S, Saver JL. Trends in Acute Ischemic Stroke Trials Through the 20th Century. *Stroke* 2001 June 1;32(6):1349-59.

(7) Hong K-S, Lee M, Lee SJ, Hao Q, Liebeskind D, Saver J. Acute stroke trials in the 1st decade of the 21th century. *Stroke Conference: 2011 International Stroke Conference* 2011;42(3):01.

(8) Lees KR, Hankey GJ, Hacke W. Design of future acute-stroke treatment trials. *The Lancet Neurology* 2003 January;2(1):54-61.

(9) Hacke W. Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke: The European Cooperative Acute Stroke Study (ECASS). *JAMA* 1995 October 4;274(13):1017-25.

(10) Hacke W, Kaste M, Fieschi C, von KR, Davalos A, Meier D, Larrue V, Bluhmki E, Davis S, Donnan G, Schneider D, Diez-Tejedor E, Trouillas P. Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute

ischaemic stroke (ECASS II). Second European-Australasian Acute Stroke Study Investigators. *Lancet* 1998 October 17;352(9136):1245-51.

(11) Tissue plasminogen activator for acute ischemic stroke. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. *N Engl J Med* 1995 December 14;333(24):1581-7.

(12) Diener HC, Hacke W, Hennerici M, Radberg J, Hantson L, De KJ. Lubeluzole in acute ischemic stroke. A double-blind, placebo-controlled phase II trial. Lubeluzole International Study Group. *Stroke* 1996 January;27(1):76-81.

(13) Diener HC. Multinational randomised controlled trial of lubeluzole in acute ischaemic stroke. European and Australian Lubeluzole Ischaemic Stroke Study Group. *Cerebrovasc Dis* 1998 May;8(3):172-81.

(14) Gandolfo C, Sandercock P, Conti M. Lubeluzole for acute ischaemic stroke. *Cochrane Database of Systematic Reviews (1):CD001924, 2002* 2002;(1):CD001924.

(15) Wahlgren NG, Ranasinha KW, Rosolacci T, Franke CL, van Erven PMM, Ashwood T, Claesson L, for the CLASS StudyGroup. Clomethiazole Acute Stroke Study (CLASS): Results of a Randomized, Controlled Trial of Clomethiazole Versus Placebo in 1360 Acute Stroke Patients. *Stroke* 1999 January 1;30(1):21-8.

(16) Lyden P, Shuaib A, Ng K, Levin K, Atkinson RP, Rajput A, Wechsler L, Ashwood T, Claesson L, Odergren T, Salazar-Grueso E, on behalf of the CLASS-. Clomethiazole Acute Stroke Study in Ischemic Stroke (CLASS-I): Final Results. *Stroke* 2002 January 1;33(1):122-9.

(17) Lees KR, Asplund K, Carolei A, Davis SM, Diener HC, Kaste M, Orgogozo JM, Whitehead J. Glycine antagonist (gavestinel) in neuroprotection (GAIN International) in patients with acute stroke: a randomised controlled trial. GAIN International Investigators. *Lancet* 2000 June 3;355(9219):1949-54.

(18) Lees KR, Zivin JA, Ashwood T, Davalos A, Davis SM, Diener HC, Grotta J, Lyden P, Shuaib A, Hardemark HG, Wasiewski WW. NXY-059 for Acute Ischemic Stroke. *N Engl J Med* 2006 February 9;354(6):588-600.

(19) Grotta J. Why do all drugs work in animals but none in stroke patients? 2 Neuroprotective therapy. *Journal of Internal Medicine* 1995 January 1;237(1):89-94.

(20) Grotta J. Neuroprotection Is Unlikely to Be Effective in Humans Using Current Trial Designs. *Stroke* 2002 January 1;33(1):306-7.

(21) Stroke Therapy Academic Industry Roundtable II (STAIR-II). Recommendations for clinical trial evaluation of acute stroke therapies. *Stroke* 2001 July;32(7):1598-606.

(22) Samsa GP, Matchar DB, Goldstein L, Bonito A, Duncan PW, Lipscomb J, Enarson C, Witter D, Venus P, Paul JE, Weinberger M. Utilities for major stroke: results from a survey of preferences among persons at increased risk for stroke. *American Heart Journal* 1998 October;136(4:Pt 1):t-13.

(23) Gresham GE. Stroke outcome research. *Stroke* 1986 May 1;17(3):358-60.

(24) Gresham GEM. Past Achievements and New Directions in Stroke Outcome Research. *Stroke* 1990 September;21(9):II.

(25) Basmajian JVM. The Call for Action. *Stroke* 1990 September;21(9):II-3.

(26) Symposium Recommendations for Methodology in Stroke Outcome Research: Task Force on Stroke Impairment, Task Force on Stroke Disability, and Task Force on Stroke Handicap. *Stroke* 1990 September;21(9):II.

(27) Hewer RLP. Outcome Measures in Stroke: A British View. *Stroke* 1990 September;21(9):II.

(28) World Health Organisation. International Classification of Functioning, Disability and Health (ICF). Geneva, Switzerland: World Health Organisation; 2001.

(29) Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. *International Journal of Stroke* 2009;4(3):200-5.

(30) Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. *Annals of Internal Medicine* 1983 December;99(6):843-8.

(31) Feinstein AR. An Additional Basic Science for Clinical Medicine: I. The Constraining Fundamental Paradigms. *Annals of Internal Medicine* 1983 September;99(3):393-7.

(32) Feinstein AR. An Additional Basic Science for Clinical Medicine: III. The Challenges of Comparison and Measurement. *Annals of Internal Medicine* 1983 November;99(5):705-12.

(33) Wright JG, Feinsten AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *Journal of Clinical Epidemiology* 1992 November;45(11):1201-18.

(34) Feinstein AR. Clinimetric perspectives. *Journal of Chronic Diseases* 1987;40(6):635-40.

(35) Cote R, Battista RN, Wolfson CM, Hachinski V. Stroke assessment scales: guidelines for development, validation, and reliability assessment. *Canadian Journal of Neurological Sciences* 1988 August;15(3):261-5.

(36) Cyr L, Francis K. Measures of clinical agreement for nominal and categorical data: The kappa coefficient. *Computers in Biology and Medicine* 1992 July;22(4):239-46.

(37) Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968 October;70(4):213-20.

(38) Fleiss JL, Cohen J. The equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as measures of Reliability. *Educational and Psychological Measurement* 1973;33:613-9.

(39) Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society Series D (The Statistician)* 1983 September 1;32(3):307-17.

(40) Altman D, Bland JM. Statistical Methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;327:307-10.

(41) Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. *Journal of Chronic Diseases* 1986;39(11):897-906.

(42) Copay AG, Subach BR, Glassman SD, Polly DW, Jr., Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine Journal: Official Journal of the North American Spine Society* 2007 September;7(5):541-6.

(43) Cook CE. Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense. *Journal of Manual & Manipulative Therapy* 2008 October 1;16(4):82E-3E.

(44) Hacke W, Donnan G, Fieschi C, Kaste M, von KR, Broderick JP, Brott T, Frankel M, Grotta JC, Haley EC, Jr., Kwiatkowski T, Levine SR, Lewandowski C, Lu M, Lyden P, Marler JR, Patel S, Tilley BC, Albers G, Bluhmki E, Wilhelm M, Hamilton S, ATLANTIS T, I, ECASS T, I, NINDS rt-PA Study Group. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. *Lancet* 2004 March 6;363(9411):768-74.

(45) Brott T, Adams HP, Olinger CP, Marler JR, Barsan WG, Biller J, Spilker J, Holleran R, Eberle R, Hertzberg V. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke* 1989 July 1;20(7):864-70.

(46) Lyden P, Brott T, Tilley B, Welch KM, Mascha EJ, Levine S, Haley EC, Grotta J, Marler J. Improved reliability of the NIH Stroke Scale using video training. NINDS TPA Stroke Study Group. *Stroke* 1994 November;25(11):2220-6.

(47) Goldstein LB, Samsa GP. Reliability of the National Institutes of Health Stroke Scale: Extension to Non-Neurologists in the Context of a Clinical Trial. *Stroke* 1997 February 1;28(2):307-10.

(48) Lyden P, Raman R, Liu L, Grotta J, Broderick J, Olson S, Shaw S, Spilker J, Meyer B, Emr M, Warren M, Marler J. NIHSS training and certification using a new digital video disk is reliable. *Stroke* 2005 November;36(11):2446-9.

(49) Lyden PD, Lu M, Levine SR, Brott TG, Broderick J, and the NINDS rtPA Stroke Study Group. A Modified National Institutes of Health Stroke Scale for Use in Stroke Clinical Trials: Preliminary Reliability and Validity. *Stroke* 2001 June 1;32(6):1310-7.

(50) Meyer BC, Hemmen TM, Jackson CM, Lyden PD. Modified National Institutes of Health Stroke Scale for Use in Stroke Clinical Trials: Prospective Reliability and Validity. *Stroke* 2002 May 1;33(5):1261-6.

(51) Adams HP, Davis PH, Leira EC, Chang KC, Bendixen BH, Clarke WR, Woolson RF, Hansen MD. Baseline NIH Stroke Scale score strongly predicts outcome after stroke. *Neurology* 1999 July 1;53(1):126.

(52) Muir KW, Weir CJ, Murray GD, Povey C, Lees KR. Comparison of Neurological Scales and Scoring Systems for Acute Stroke Prognosis. *Stroke* 1996 October 1;27(10):1817-20.

(53) Woo D, Broderick JP, Kothari RU, Lu M, Brott T, Lyden PD, Marler JR, Grotta JC, for the NINDS. Does the National Institutes of Health Stroke Scale Favor Left Hemisphere Strokes? *Stroke* 1999 November 1;30(11):2355-9.

(54) Millis SR, Straube D, Iramaneerat C, Smith EV, Lyden P. Measurement Properties of the National Institutes of Health Stroke Scale for People With Right- and Left-Hemisphere Lesions: Further Analysis of the Clomethiazole for Acute Stroke Study–Ischemic (Class-I) Trial. *Arch Phys Med Rehabil* 2007 March 1;88(3):302-8.

(55) Mahoney FI, Barthel D. Functional evaluation: The barthel index. *Maryland State Medical Journal* 1965 February;14:61-5.

(56) Fortinsky RH, Granger CV, Seltzer GB. The use of functional assessment in understanding home care needs. *Medical Care* 1981 May;19(5):489-97.

(57) Novak S, Johnson J, Greenwood R. Barthel revisited: Making guidelines work. *Clinical Rehabilitation 10 (2) (pp 128-134), 1996 Date of Publication: 1996* 1996;(2):1996.

(58) Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. *International Disability Studies* 1988;10(2):61-3.

(59) Kirshner B, Guyatt G. A methodological framework for assessing health indices. *Journal of Chronic Diseases* 1985;38(1):27-36.

(60) Sinoff G, Ore L. The Barthel activities of daily living index: self-reporting versus actual performance in the old-old (> or = 75 years). *Journal of the American Geriatrics Society* 1997 July;45(7):832-6.

(61) Korner-Bitensky NP, Wood-Dauphinee SP. Barthel index information elicited over the telephone: Is it reliable? *American Journal of Physical Medicine & Rehabilitation* 1995 January;74(1):9-18.

(62) Gompertz P, Pound P, Ebrahim S. The reliability of stroke outcome measures. *Clinical Rehabilitation 7 (4) (pp 290-296), 1993 Date of Publication: 1993* 1993;(4):1993.

(63) Wade DT, Collin C. The Barthel ADL Index: a standard measure of physical disability?. *International Disability Studies* 1988;10(2):64-7.

(64) Loewen SC, Anderson BA. Reliability of the Modified Motor Assessment Scale and the Barthel Index. *Physical Therapy* 1988 July;68(7):1077-81.

(65) Wolfe CD, Taub NA, Woodrow EJ, Burney PG. Assessment of scales of disability and handicap for stroke patients. *Stroke* 1991 October 1;22(10):1242-4.

(66) Cincura C, Pontes-Neto OM, Neville IS, Mendes HF, Menezes DF, Mariano DC, Pereira IF, Teixeira LA, Jesus PA, de Queiroz DC, Pereira DF, Pinto E, Leite JP, Lopes AA, Oliveira-Filho J. Validation of the National Institutes of Health Stroke Scale, modified Rankin Scale and Barthel Index in Brazil: the role of cultural adaptation and structured interviewing. *Cerebrovasc Dis* 2009;27(2):119-22.

(67) Duffy L, Gajree S, Langhorne P, Stott DJ, Quinn TJ. Reliability (Inter-rater Agreement) of the Barthel Index for Assessment of Stroke Survivors: Systematic Review and Meta-analysis. *Stroke* 2013 February 1;44(2):462-8.

(68) Sainsbury A, Seebass G, Bansal A, Young JB. Reliability of the Barthel Index when used with older people. *Age & Ageing* 2005 May;34(3):228-32.

(69) Granger CV, Dewis LS, Peters NC, Sherwood CC, Barrett JE. Stroke rehabilitation: analysis of repeated Barthel index measures. *Archives of Physical Medicine & Rehabilitation* 1979 January;60(1):14-7.

(70) Granger CV, Hamilton BB, Gresham GE. The stroke rehabilitation outcome study--Part I: General description. *Archives of Physical Medicine & Rehabilitation* 1988 July;69(7):506-9.

(71) Granger CV, Hamilton BB, Gresham GE, Kramer AA. The stroke rehabilitation outcome study: Part II. Relative merits of the total Barthel index score and a four-item subscore in predicting patient outcomes. *Archives of Physical Medicine & Rehabilitation* 1989 February;70(2):100-3.

(72) Kwon S, Hartzema AG, Duncan PW, Min-Lai S. Disability Measures in Stroke: Relationship Among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale. *Stroke* 2004 April 1;35(4):918-23.

(73) Duncan PW, Samsa GP, Weinberger M, Goldstein LB, Bonito A, Witter DM, Enarson C, Matchar D. Health Status of Individuals With Mild Stroke. *Stroke* 1997 April 1;28(4):740-5.

(74) Quinn TJ, Langhorne P, Stott DJ. Barthel Index for Stroke Trials: Development, Properties, and Application. *Stroke* 2011 April 1;42(4):1146-51.

(75) Dromerick AW, Edwards DF, Diringer MN. Sensitivity to changes in disability after stroke: a comparison of four scales useful in clinical trials. *Journal of Rehabilitation Research & Development* 2003 January;40(1):1-8.

(76) Bamford JM, Sandercock PA, Warlow CP, Slattery J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 1989 June;20(6):828.

(77) Duncan PW, Wallace D, Lai SM, Johnson D, Embretson S, Laster LJ. The stroke impact scale version 2.0. Evaluation of reliability, validity, and sensitivity to change. *Stroke* 1999 October;30(10):2131-40.

(78) Duncan PW, Bode RK, Min LS, Perera S, Glycine Antagonist in Neuroprotection Americans Investigators. Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. *Archives of Physical Medicine & Rehabilitation* 2003 July;84(7):950-63.

(79) Kasner SE. Clinical interpretation and use of stroke scales. *Lancet Neurology* 2006 July;5(7):603-12.

(80) Tilley BC, Marler J, Geller NL, Lu M, Legler J, Brott T, Lyden P, Grotta J. 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* 1996 November 1;27(11):2136-42.

(81) Hacke W, Bluhmki E, Steiner T, Tatlisumak T, Mahagne MH, Sacchetti ML, Meier D. 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* 1998 October 1;29(10):2073-5.

(82) Saver JL. Novel End Point Analytic Techniques and Interpreting Shifts Across the Entire Range of Outcome Scales in Acute Stroke Trials. *Stroke* 2007 November 1;38(11):3055-62.

(83) Rankin J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scottish Medical Journal* 1957 May;2(5):200-15.

(84) Langhorne P, Dennis M, Hankey G, Weir C, Williams B. Organised inpatient (stroke unit) care for stroke. *Cochrane Database of Systematic Reviews (4) , 2007 Article Number: CD000197 Date of Publication: 2007* 2007;(4):CD000197.

(85) Farrell B, Godwin J, Richards S, Warlow C. The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: final results. *Journal of Neurology, Neurosurgery & Psychiatry* 1991 December;54(12):1044-54.

(86) van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 1988 May 1;19(5):604-7.

(87) Bloch RF. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 1988 November 1;19(11):1448.

(88) New PW, Buchbinder R. Critical appraisal and review of the Rankin scale and its derivatives. *Neuroepidemiology* 2006;26(1):4-15.

(89) de Haan R, Limburg M, Bossuyt P, van der Meulen J, Aaronson N. The Clinical Meaning of Rankin `Handicap' Grades After Stroke. *Stroke* 1995 November 1;26(11):2027-30.

(90) Samuelsson M, Soderfeldt B, Olsson GB. Functional Outcome in Patients With Lacunar Infarction. *Stroke* 1996 May 1;27(5):842-6.

(91) collaborative group. Effect of thrombolysis with alteplase within 6 h of acute ischaemic stroke on long-term outcomes (the third International Stroke Trial [IST-3]): 18-month follow-up of a randomised controlled trial. *Lancet Neurology* 2013 August;12(8):768-76.

(92) Huybrechts KF, Caro JJ, Xenakis JJ, Vemmos KN. The prognostic value of the modified Rankin Scale score for long-term survival after first-ever stroke. Results from the Athens Stroke Registry. *Cerebrovasc Dis* 2008;26(4):381-7.

(93) De HR, Horn J, Limburg M, van der Meulen J, Bossuyt P. A comparison of five stroke scales with measures of disability, handicap, and quality of life. *Stroke* 1993 August;24(8):1178-81.

(94) Dawson J, Lees JS, Chang TP, Walters MR, Ali M, Davis SM, Diener HC, Lees KR, for the GAIN and VISTA Investigators. Association Between Disability Measures and Healthcare Costs After Initial Treatment for Acute Stroke. *Stroke* 2007 June 1;38(6):1893-8.

(95) Lev MH, Segal AZ, Farkas J, Hossain ST, Putman C, Hunter GJ, Budzik R, Harris GJ, Buonanno FS, Ezzeddine MA, Chang Y, Koroshetz WJ, Gonzalez RG, Schwamm LH. Utility of perfusion-weighted CT imaging in acute middle cerebral artery stroke treated with intra-arterial thrombolysis: prediction of final infarct volume and clinical outcome. *Stroke* 2001 September;32(9):2021-8.

(96) Schiemanck SK, Post MW, Kwakkel G, Witkamp TD, Kappelle LJ, Prevo AJ. Ischemic lesion volume correlates with long-term functional outcome and quality of life of

middle cerebral artery stroke survivors. *Restorative Neurology & Neuroscience* 2005;23(3-4):257-63.

(97) Demchuk AM, Tanne D, Hill MD, Kasner SE, Hanson S, Grond M, Levine SR, mMulticentre tPA Stroke Survey Group. Predictors of good outcome after intravenous tPA for acute ischemic stroke. *Neurology* 2001 August 14;57(3):474-80.

(98) Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the Modified Rankin Scale. *Stroke* 2007 November 1;38(11):e144.

(99) Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the Modified Rankin Scale: A Systematic Review. *Stroke* 2009 October 1;40(10):3393-5.

(100) Wilson JTL, Hareendran A, Hendry A, Potter J, Bone I, Muir KW. Reliability of the Modified Rankin Scale Across Multiple Raters: Benefits of a Structured Interview. *Stroke* 2005 April 1;36(4):777-81.

(101) Newcommon NJ, Green TL, Haley E, Cooke T, Hill MD. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. *Stroke* 2003;34(2):377-8.

(102) Quinn TJ, Dawson J, Walters MR, Lees KR. Variability in Modified Rankin Scoring Across a Large Cohort of International Observers. *Stroke* 2008 November 1;39(11):2975-9.

(103) Banks JL, Marotta CA. Outcomes Validity and Reliability of the Modified Rankin Scale: Implications for Stroke Clinical Trials: A Literature Review and Synthesis. *Stroke* 2007 March 1;38(3):1091-6.

(104) Saver JL. Optimal End Points for Acute Stroke Therapy Trials: Best Ways to Measure Treatment Effects of Drugs and Devices. *Stroke* 2011 August 1;42(8):2356-62.

(105) Wilson JTL, Hareendran A, Grant M, Baird T, Schulz UGR, Muir KW, Bone I. Improving the Assessment of Outcomes in Stroke: Use of a Structured Interview to Assign Grades on the Modified Rankin Scale. *Stroke* 2002 September 1;33(9):2243-6.

(106) Edwards M, Feightner J, Goldsmith CH. Inter-rater reliability of assessments administered by individuals with and without a background in health care. *Occupational Therapy Journal of Research* 1995;15(2):103-10.

(107) Nichols-Larsen DS, Clark PC, Zeringue A, Greenspan A, Blanton S. Factors Influencing Stroke Survivors' Quality of Life During Subacute Recovery. *Stroke* 2005 July 1;36(7):1480-4.

(108) Duncan PW, Jorgensen HS, Wade DT. Outcome Measures in Acute Stroke Trials : A Systematic Review and Some Recommendations to Improve Practice. *Stroke* 2000 June 1;31(6):1429-38.

(109) Haacke C, Althaus A, Spottke A, Siebert U, Back T, Dodel R. Long-Term Outcome After Stroke: Evaluating Health-Related Quality of Life Using Utility Measurements. *Stroke* 2006 January 1;37(1):193-8.

(110) Rivero-Arias O, Ouellet M, Gray A, Wolstenholme J, Rothwell PM, Luengo-Fernandez R. Mapping the modified Rankin scale (mRS) measurement into the generic EuroQol (EQ-5D) health outcome. *Medical Decision Making* 2010 May;30(3):341-54.

(111) Hong KS, Saver JL. Quantifying the Value of Stroke Disability Outcomes: WHO Global Burden of Disease Project Disability Weights for Each Level of the Modified Rankin Scale. *Stroke* 2009 December 1;40(12):3828-33.

(112) Sentinel Stroke National Audit Programme: SSNAP core dataset. Royal College of Physicians; 2013 Feb 18. Report No.: 1.1.2.

(113) Fearon P, McArthur KS, Garrity K, Graham LJ, McGroarty G, Vincent S, Quinn TJ. Prestroke Modified Rankin Stroke Scale Has Moderate Interobserver Reliability and Validity in an Acute Stroke Setting. *Stroke* 2012 December 1;43(12):3184-8.

(114) Zhao H, Collier JM, Quah DM, Purvis T, Bernhardt J. The modified Rankin Scale in acute stroke has good inter-rater-reliability but questionable validity. *Cerebrovasc Dis* 2010 January;29(2):188-93.

(115) Dennis M, Mead G, Doubal F, Graham C. Determining the Modified Rankin Score After Stroke by Postal and Telephone Questionnaires. *Stroke* 2012 March 1;43(3):851-3.

(116) Wilson JT, Edwards P, Fiddes H, Stewart E, Teasdale GM. Reliability of postal questionnaires for the Glasgow Outcome Scale. *Journal of Neurotrauma* 2002 September;19(9):999-1005.

(117) Duncan PW, Reker DM, Horner RD, Samsa GP, Hoenig H, LaClair BJ, Dudley TK. Performance of a mail-administered version of a stroke-specific outcome measure, the Stroke Impact Scale. *Clinical Rehabilitation* 2002 August;16(5):493-505.

(118) Korner-Bitensky N, Wood-Dauphinee S, Siemiatycki J, Shapiro S, Becker R. Health-related information postdischarge: telephone versus face-to-face interviewing. *Archives of Physical Medicine & Rehabilitation* 1994 December;75(12):1287-96.

(119) Heuschmann PU, Kolominsky-Rabas PL, Nolte CH, Hunermund G, Ruf HU, Laumeier I, Meyrer R, Alberti T, Rahmann A, Kurth T, Berger K. The reliability of the german version of the barthel-index and the development of a postal and telephone version for the application on stroke patients. [German] Abstract Only. *Fortschritte der Neurologie-Psychiatrie* 2005 February;73(2):74-82.

(120) Candelise L, Pinardi G, Aritzu E, Musicco M. Telephone Interview for Stroke Outcome Assessment. *Cerebrovasc Dis* 1994;4(5):341-3.

(121) Merino JG, Lattimore SU, Warach S. Telephone assessment of stroke outcome is reliable. *Stroke* 2005 February;36(2):232-3.

(122) Janssen PM, Visser NA, Mees SMD, Klijn CJM, Algra A, Rinkel GJE. Comparison of Telephone and Face-to-Face Assessment of the Modified Rankin Scale. *Cerebrovasc Dis* 2010 January;29(2):137-9.

(123) Kasner SE, Chalela JA, Luciano JM, Cucchiara BL, Raps EC, McGarvey ML, Conroy MB, Localio AR. Reliability and Validity of Estimating the NIH Stroke Scale Score from Medical Records. *Stroke* 1999 August 1;30(8):1534-7.

(124) Quinn TJ, Ray G, Atula S, Walters MR, Dawson J, Lees KR. Deriving Modified Rankin Scores From Medical Case-Records. *Stroke* 2008 December 1;39(12):3421-3.

(125) McArthur KS, Beagan ML, Degnan A, Howarth RC, Mitchell KA, McQuaige FB, Shannon MA, Quinn TJ. Reliability and validity of proxy derived modified rankin scale assessment. *Stroke Conference: 2011 International Stroke Conference Los Angeles, CA United States* 2011;01.

(126) Knapp P, Hewison J. Disagreement in patient and carer assessment of functional abilities after stroke. *Stroke* 1999 May;30(5):934-8.

(127) Williams LS, Bakas T, Brizendine E, Plue L, Tu W, Hendrie H, Kroenke K. How Valid Are Family Proxy Assessments of Stroke Patients' Health-Related Quality of Life? *Stroke* 2006 August 1;37(8):2081-5.

(128) Wyller TB, Sveen U, Bautz-Holter E. The Barthel ADL index one year after stroke: comparison between relatives' and occupational therapist's scores. *Age & Ageing* 1995 September;24(5):398-401.

(129) Oczkowski C, O'Donnell M. Reliability of Proxy Respondents for Patients With Stroke: A Systematic Review. *Journal of Stroke and Cerebrovascular Diseases* 2009 September;19(5):410-6.

(130) Shinohara Y, Minematsu K, Amano T, Ohashi Y. Modified Rankin scale with expanded guidance scheme and interview questionnaire: interrater agreement and reproducibility of assessment. *Cerebrovasc Dis* 2006;21(4):271-8.

(131) Patel NB, Rao VAM, Heilman-Espinoza ERM, Lai RD, Quesada RAM, Flint ACM. Simple and Reliable Determination of the Modified Rankin Scale Score in Neurosurgical and Neurological Patients: The mRS-9Q. *Neurosurgery* 2012 November;71(5):971-5.

(132) Bruno A, Shah N, Lin C, Close B, Hess DC, Davis K, Baute V, Switzer JA, Waller JL, Nichols FT. Improving Modified Rankin Scale Assessment With a Simplified Questionnaire. *Stroke* 2010 May 1;41(5):1048-50.

(133) Saver JL, Filip B, Hamilton S, Yanes A, Craig S, Cho M, Conwit R, Starkman S, for the FAST-MAG Investigators and Coordinators. Improving the Reliability of Stroke Disability Grading in Clinical Trials and Clinical Practice: The Rankin Focused Assessment (RFA). *Stroke* 2010 May 1;41(5):992-5.

(134) Celani MG, Cantisani TA, Righetti E, Spizzichino L, Ricci S. Different Measures for Assessing Stroke Outcome: An Analysis From the International Stroke Trial in Italy. *Stroke* 2002 January 1;33(1):218-23.

(135) Bruno A, Close B, Switzer JA, Hess DC, Gross H, Nichols FT, Akinwuntan AE. Simplified modified Rankin Scale questionnaire correlates with stroke severity. *Clinical Rehabilitation* 2013 February 14.

(136) Bruno A, Shah N, Akinwuntan AE, Close B, Switzer JA. Stroke size correlates with functional outcome on the simplified modified rankin scale questionnaire. *Journal of Stroke and Cerebrovascular Diseases* 2013;22(6):August.

(137) Bruno A, Akinwuntan AE, Lin C, Close B, Davis K, Baute V, Aryal T, Brooks D, Hess DC, Switzer JA, Nichols FT. Simplified Modified Rankin Scale Questionnaire: Reproducibility Over the Telephone and Validation With Quality of Life. *Stroke* 2011 August 1;42(8):2276-9.

(138) Yuan JL, Bruno A, Li T, Li SJ, Zhang XD, Li HY, Jia K, Qin W, Chen AC, Hu WL. Replication and extension of the simplified modified rankin scale in 150 Chinese stroke patients. *European Neurology* 2012;67(4):206-10.

(139) Muir KW, Lees KR, Ford I, Davis S, Intravenous Magnesium Efficacy in Stroke (IMAGES) Study Investigators. Magnesium for acute stroke (Intravenous Magnesium Efficacy in Stroke trial): randomised controlled trial. *Lancet* 2004 February 7;363(9407):439-45.

(140) The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. International Stroke Trial Collaborative Group. *Lancet* 1997 May 31;349(9065):1569-81.

(141) Hofmeijer J, Kappelle LJ, Algra A, Amelink GJ, van Gijn J, van der Worp HB. Surgical decompression for space-occupying cerebral infarction (the Hemicraniectomy After Middle Cerebral Artery infarction with Life-threatening Edema Trial [HAMLET]): a multicentre, open, randomised trial. *The Lancet Neurology* 2009 April;8(4):326-33.

(142) Quinn TJ, Dawson J, Walters MR, Lees KR. Initial experience with video based modified Rankin assessment. *Cerebrovasc Dis* 2007;23:s115.

(143) Jaffar S, Leach A, Smith PG, Cutts F, Greenwood B. Effects of misclassification of causes of death on the power of a trial to assess the efficacy of a pneumococcal conjugate vaccine in The Gambia. *International Journal of Epidemiology* 2003 June;32(3):430-6.

(144) Choi SC, Clifton GL, Marmarou A, Miller ER. Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. *Journal of Neurotrauma* 2002 January;19(1):17-22.

(145) Lu J, Murray GD, Steyerberg EW, Butcher I, McHugh GS, Lingsma H, Mushkudiani N, Choi S, Maas AIR, Marmarou A. Effects of Glasgow Outcome Scale Misclassification on Traumatic Brain Injury Clinical Trials. *Journal of Neurotrauma* 2008 June 1;25(6):641-51.

(146) Lees KR. Training and Consistency in Stroke Assessments. *Stroke* 2009 July 1;40(7):2297.

(147) Albanese MA, Clarke WR, Adams HP, Jr., Woolson RF. Ensuring reliability of outcome measures in multicenter clinical trials of treatments for acute ischemic stroke. The program developed for the Trial of Org 10172 in Acute Stroke Treatment (TOAST). *Stroke* 1994 September 1;25(9):1746-51.

(148) Josephson SA, Hills NK, Johnston SC. NIH stroke scale reliability in ratings from a large sample of clinicians. *Cerebrovasc Dis* 2006 October;22(5-6):389-95.

(149) Lyden P, Raman R, Liu L, Emr M, Warren M, Marler J. National Institutes of Health Stroke Scale Certification Is Reliable Across Multiple Venues. *Stroke* 2009 July 1;40(7):2507-11.

(150) Pezzella FR, Picconi O, De LA, Lyden PD, Fiorelli M. Development of the Italian version of the National Institutes of Health Stroke Scale: It-NIHSS. *Stroke* 2009 July;40(7):2557-9.

(151) Quinn TJ, Lees KR, Hardemark HG, Dawson J, Walters MR. Initial Experience of a Digital Training Resource for Modified Rankin Scale Assessment in Clinical Trials. *Stroke* 2007 August 1;38(8):2257-61.

(152) Lai SM, Duncan PW. Stroke recovery profile and the Modified Rankin assessment. *Neuroepidemiology* 2001 February;20(1):26-30.

(153) Weisscher N, Vermeulen M, Roos YB, de Haan RJ. What should be defined as good outcome in stroke trials; a modified Rankin score of 0-1 or 0-2? *Journal of Neurology* 2008 June;255(6):867-74.

(154) Berge E, Barer D. Could stroke trials be missing important treatment effects?. *Cerebrovasc Dis* 2002;13(1):73-5.

(155) Sulter G, Steen C, Jacques DK. Use of the Barthel Index and Modified Rankin Scale in Acute Stroke Trials. *Stroke* 1999 August 1;30(8):1538-41.

(156) Bath PMW, Lees KR, Schellinger PD, Altman H, Bland M, Hogg C, Howard G, Saver JL. Statistical Analysis of the Primary Outcome in Acute Stroke Trials. *Stroke* 2012 April 1;43(4):1171-8.

(157) Murray GD, Barer D, Choi S, Fernandes H, Gregson B, Lees KR, Maas AI, Marmarou A, Mendelow AD, Steyerberg EW, Taylor GS, Teasdale GM, Weir CJ. Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. *Journal of Neurotrauma* 2005 May;22(5):511-7.

(158) Saver JL, Yafeh B. Confirmation of tPA Treatment Effect by Baseline Severity-Adjusted End Point Reanalysis of the NINDS-tPA Stroke Trials. *Stroke* 2007 February 1;38(2):414-6.

(159) Adams HP, Jr., Leclerc JR, Bluhmki E, Clarke W, Hansen MD, Hacke W. Measuring outcomes as a function of baseline severity of ischemic stroke. *Cerebrovasc Dis* 2004;18(2):124-9.

(160) Young FB, Lees KR, Weir CJ, for the GAIN International Trial Steering Committee and Investigators. Improving Trial Power Through Use of Prognosis-Adjusted End Points. *Stroke* 2005 March 1;36(3):597-601.

(161) Adams HP, Jr., Effron MB, Torner J, Davalos A, Frayne J, Teal P, Leclerc J, Oemar B, Padgett L, Barnathan ES, Hacke W, for the AbESTT-II Investigators. Emergency Administration of Abciximab for Treatment of Patients With Acute Ischemic Stroke: Results of an International Phase III Trial: Abciximab in Emergency Treatment of Stroke Trial (AbESTT-II). *Stroke* 2008 January 1;39(1):87-99.

(162) Mendelow AD, Gregson BA, Fernandes HM, Murray GD, Teasdale GM, Hope DT, Karimi A, Shaw MD, Barer DH, STICH i. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial. *Lancet* 2005 January 29;365(9457):387-97.

(163) Ehrenreich H, Hasselblatt M, Dembowski C, Cepek L, Lewczuk P, Stiefel M, Rustenbeck HH, Breiter N, Jacob S, Knerlich F, Bohn M, Poser W, Ruther E, Kochen M, Gefeller O, Gleiter C, Wessel TC, De RM, Itri L, Prange H, Cerami A, Brines M, Siren AL. Erythropoietin therapy for acute stroke is both safe and beneficial. *Molecular Medicine* 2002 August;8(8):495-505.

(164) Saver JL, Eckstein M, Stratton S, Pratt F, Hamilton S, Conwit R, Liebeskind D, Lyden P, Sanossian N, Sung G, Kramer I, Moreau G, Goldweber R, Starkman S. Abstract 214: The Field Administration of Stroke Therapy - Magnesium (FAST-MAG) Phase 3 Trial: Primary Results. *Stroke* 2014 February 1;45(Suppl 1):A214.

(165) The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international

stroke trial [IST-3]): a randomised controlled trial. *The Lancet* 2013;379(9834):2352-63.

(166) Saver JL, Gornbein J. Treatment effects for which shift or binary analyses are advantageous in acute stroke trials. *Neurology* 2009 April 14;72(15):1310-5.

(167) The Optimising Analysis of Stroke Trials (OAST) Collaboration. Can We Improve the Statistical Analysis of Stroke Trials?: Statistical Reanalysis of Functional Outcomes in Stroke Trials * OAST Supplemental Appendix I: Statistical Tests Compared (see Table I) * OAST Supplemental Appendix II: Supplementary Analyses * OAST Supplemental Appendix III: Trial Data (see Tables II and III) * OAST Supplemental Appendix IV: Results (see Table IV). *Stroke* 2007 June 1;38(6):1911-5.

(168) Savitz SI, Benatar M, Saver JL, Fisher M. Outcome analysis in clinical trial design for acute stroke: Physicians' attitudes and choices. *Cerebrovasc Dis* 2008 August;26(2):156-62.

(169) Saver JL. Number needed to treat estimates incorporating effects over the entire range of clinical outcomes: Novel derivation method and application to thrombolytic therapy for acute stroke. *Archives of Neurology* 2004 July;61(7):1066-70.

(170) Conover W. *Practical Non Parametric Statistics*. 3rd Edition ed. New York, USA: John Wiley; 1999.

(171) Moss FR, Pierson D.A., Orcogorman D.A. Stochastic Resonance: Tutorial and Update. *Int J Bifurcation Chaos* 1994 December 1;04(06):1383-97.

(172) Training Campus NIHSS English. 2013. 8-10-2013.
Ref Type: Online Source

(173) Training Campus mRS English. 2013.
Ref Type: Online Source

(174) Konig-Bitensky N, Ziegler A, Bluhmki E, Hacke W, Bath PMW, Sacco RL, Diener HC, Weimar C, on behalf of the Virtual International Stroke Trials Archive (VISTA) Investigators. Predicting Long-Term Outcome After Acute Ischemic Stroke: A Simple Index Works in Patients From Controlled Clinical Trials. *Stroke* 2008 June 1;39(6):1821-6.

(175) Saver JL, Altman H. Relationship Between Neurologic Deficit Severity and Final Functional Outcome Shifts and Strengthens During First Hours After Onset. *Stroke* 2012 June 1;43(6):1537-41.

(176) Willmot M, Leonardi-Bee J, Bath PMW. High Blood Pressure in Acute Stroke and Subsequent Outcome: A Systematic Review. *Hypertension* 2004 January 1;43(1):18-24.

(177) Sare GM, Ali M, Shuaib A, Bath PMW, for the VISTA Collaboration. Relationship Between Hyperacute Blood Pressure and Outcome After Ischemic Stroke: Data From the VISTA Collaboration. *Stroke* 2009 June 1;40(6):2098-103.

(178) Bruno A, Levine SR, Frankel MR, Brott TG, Lin Y, Tilley BC, Lyden PD, Broderick JP, Kwiatkowski TG, Fineberg SE, NINDS rt-PA Stroke Study Group. Admission glucose level and clinical outcomes in the NINDS rt-PA Stroke Trial. *Neurology* 2002 September 10;59(5):669-74.

(179) Hu GC, Hsieh SF, Chen YM, Hu YN, Kang CL, Chien KL. The prognostic roles of initial glucose level and functional outcomes in patients with ischemic stroke: difference between diabetic and nondiabetic patients. *Disability & Rehabilitation* 2012;34(1):34-9.

(180) Luitse MJ, Biessels GJ, Rutten GE, Kappelle LJ. Diabetes, hyperglycaemia, and acute ischaemic stroke. *Lancet Neurology* 2012 March;11(3):261-71.

(181) Quinn TJ, Dawson J, Lees JS, Chang TP, Walters MR, Lees KR, for the GAIN and VISTA Investigators. Time Spent at Home Poststroke: "Home-Time" a Meaningful and Robust Outcome Measure for Stroke Trials. *Stroke* 2008 January 1;39(1):231-3.

(182) Mishra NK, Shuaib A, Lyden P, Diener HC, Grotta J, Davis S, Davalos A, Ashwood T, Wasiewski W, Lees KR. Home Time Is Extended in Patients With Ischemic Stroke Who Receive Thrombolytic Therapy: A Validation Study of Home Time as an Outcome Measure. *Stroke* 2011 April 1;42(4):1046-50.

(183) Schulz KF, Grimes DA. Sample size slippages in randomised trials: Exclusions and the lost and wayward. *Lancet* 2002;359(9308):02.

(184) Hansson MG, Hakama M. Ulysses contracts for the doctor and for the patient. *Contemporary Clinical Trials* 2010 May;31(3):202-6.

(185) Edwards SJ. Assessing the remedy: the case for contracts in clinical trials. *American Journal of Bioethics* 2011 April;11(4):3-12.

(186) Kemmler G, Hummer M, Widschwendter C, Fleischhacker WW. Dropout rates in placebo-controlled and active-control clinical trials of antipsychotic drugs: a meta-analysis. *Archives of General Psychiatry* 2005 December;62(12):1305-12.

(187) Weissman J, Flint A, Meyers B, Ghosh S, Mulsant B, Rothschild A, Whyte E, STOP-PD Study Group. Factors associated with non-completion in a double-blind randomized controlled trial of olanzapine plus sertraline versus olanzapine plus placebo for psychotic depression. *Psychiatry Research* 2012 May 30;197(3):221-6.

(188) Hoste RR, Zaitsoff S, Hewell K, le GD. What can dropouts teach us about retention in eating disorder treatment studies? *International Journal of Eating Disorders* 2007 November;40(7):668-71.

(189) Kani C, Pehlivanidis A, Papanikolaou K, Papadopoulou DZ. Dropouts due to other reasons than adverse events during clinical trials in Alzheimer's disease. *European Neuropsychopharmacology Conference: 22 ECNP Congress Istanbul Turkey Conference Start: 20090912 Conference End: 20090916 Conference Publication: (var pagings)* 2009;19(pp S619-S620):2009.

(190) Coley N, Gardette V, Cantet C, Gillette-Guyonnet S, Nourhashemi F, Vellas B, Andrieu S. How should we deal with missing data in clinical trials involving Alzheimer's disease patients? *Current Alzheimer Research* 2011 June;8(4):421-33.

(191) Helliwell B, Aylesworth R, McDowell I, Baumgarten M, Sykes E. Correlates of nonparticipation in the Canadian Study of Health and Aging. *International Psychogeriatrics* 2001;13:Supp-56.

(192) Kitler ME. Clinical Trials in the Elderly: Pivotal Points. *Clinical Geriatric Medicine* 1990;6(2):235-55.

(193) Ferrucci L, Guralnik JM, Studenski S, Fried LP, Cutler GB, Jr., Walston JD, Interventions on Frailty Working Group. Designing randomized, controlled trials aimed at preventing or delaying functional decline and disability in frail, older persons: a consensus report. *Journal of the American Geriatrics Society* 2004 April;52(4):625-34.

(194) Llewellyn-Thomas HA, McGreal MJ, Thiel EC, Fine S, Erlichman C. Patients' willingness to enter clinical trials: Measuring the association with perceived benefit and preference for decision participation. *Social Science & Medicine* 1991;32(1):35-42.

(195) Lindstrom D, Sundberg-Petersson I, Adami J, Tonnesen H. Disappointment and drop-out rate after being allocated to control group in a smoking cessation trial. *Contemporary Clinical Trials* 2010 January;31(1):22-6.

(196) Thayabaranathan T, Cadilhac DA, Srikanth VK, Nelson MR, Kim J, Fitzgerald SM, Gerraty RP, Bladin CF, Phan TG, Thrift AG. Feasibility of a double-blind, cluster randomised-controlled trial of long-term risk factor management in survivors of stroke. *International Journal of Stroke Conference: STROKE 2012 Conference - Sydney, NSW Austr* 2012;7(pp 8):September.

(197) Shih WJ. Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine* 2002;3((1)):4.

(198) Spearman C. Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920* 1910;3(3):271-95.

(199) Quinn TJ, Dawson J, Walters MR, Lees KR. Exploring the Reliability of the Modified Rankin Scale. *Stroke* 2009 March 1;40(3):762-6.

(200) Kuritz SJ, Landis JR, Koch GG. A general overview of Mantel-Haenszel methods: Applications and recent developments. *Annual Review of Public Health* 1988;9(pp 123-160):1988.

(201) McCullagh P. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society Series B (Methodological)* 1980 January 1;42(2):109-42.

(202) Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. *Journal of Clinical Epidemiology* 1997 January;50(1):45-55.

(203) McHugh GS, Butcher I, Steyerberg EW, Marmarou A, Lu J, Lingsma HF, Weir J, Maas AIR, Murray GD. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clinical Trials* 2010 February 1;7(1):44-57.

(204) Quinn TJ, Dawson J, Walters MR, Lees KR. Predicting variability in modified Rankin Scale assessment. *Cerebrovasc Dis* 2009;27(Supplement 6 (1-241)):67.

(205) European Medicines Agency (EMEA). Reflection Paper on the Regulatory Guidance for the use of Health Related Quality of Life (HRQL) Measures in teh Evaluation of Medicinal Products. London: EMEA; 2006 Jan. Report No.: EMEA/CHMP/EWP/139391/2004.

(206) U.S.Department of Health and Human Services Food and Drug Administration. Draft Guidance for Industry. Patient Reported Outcomes Measures: use in Medical Product Development to Support Labeling Claims. Rockville, MD, USA: FDA; 2006 Feb. Report No.: Docket no. 2006D-0044.

(207) Wild D, Eremenco S, Mear I, Martin M, Houchin C, Gawlicki M, Hareendran A, Wiklund I, Chong LY, von MR, Cohen L, Molsen E. Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report. *Value in Health* 2009 June;12(4):430-40.

(208) Rungtusanatham M, Ng CH, Zhao X, Lee TS. Pooling Data Across Transparently Different Groups of Key Informants: Measurement Equivalence and Survey Research*. *Decision Sciences* 2008 February 1;39(1):115-45.

(209) Squires A. Language barriers and qualitative nursing research: methodological considerations. *International Nursing Review* 2008;55(3):265-73.

(210) Acquadro C, Conway K, Hareendran A, Aaronson N. Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value Health* 2008;11(3):509-21.

(211) Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, Erikson P. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health* 2005;8(2):94-104.

(212) Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 2000;25(24):3186-91.

(213) Koller M, Aaronson NK, Blazeby J, Bottomley A, Dewolf L, Fayers P, Johnson C, Ramage J, Scott N, West K. Translation procedures for standardised quality of life questionnaires: The European Organisation for Research and Treatment of Cancer (EORTC) approach. *European Journal of Cancer* 2007 August;43(12):1810-20.

(214) Twinn S. An exploratory study examining the influence of translation on the validity and reliability of qualitative data in nursing research. *Journal of Advanced Nursing* 1997;26(2):418-23.

(215) Kapborg I, Bertero C. Using an interpreter in qualitative interviews: does it threaten validity? *Nursing Inquiry* 2002;9(1):52-6.

(216) Berger K, Weltermann B, Kolominsky-Rabas P, Meves S, Heuschmann P, Bohner J, Neundorfer B, Hense HW, Buttner T. The reliability of stroke scales. The german version of NIHSS, ESS and Rankin scales. [German] Abstract Only. Fortschritte der Neurologie-Psychiatrie 67[2], 81-93. 1999.
Ref Type: Abstract

(217) Oveisgharan S, Shirani S, Ghorbani A, Soltanzade A, Baghaei A, Hosseini S, Sarrafzadegan N. Barthel index in a Middle-East country: translation, validity and reliability. *Cerebrovasc Dis* 2006;22(5-6):350-4.

(218) Petersen C, Morfeld M, Bullinger M. Testing and validation of the German version of the Stroke Impact Scale. [German] Abstract Only. Fortschritte der Neurologie-Psychiatrie 69[6], 284-290. 2001.
Ref Type: Abstract

(219) Geyh S, Cieza A, Stucki G. Evaluation of the german translation of the stroke impact scale using rasch analysis. *Clinical Neuropsychologist 23 (6) (pp 978-995), 2009 Date of Publication: August 2009* 2009;(6):August.

(220) Bohls C, Heise KF, Glogauer C, Scherfer E. Authorized German translation of the Motor Assessment Scale (MAS). [German] Abstract Only. Die Rehabilitation.47 (3) (pp 172-177), 2008.Date of Publication: Jun 2008. [3], Jun. 2008.
Ref Type: Abstract

(221) Kjendahl A, Jahnsen R, Aamodt G. Motor Assessment Scale in Norway: Translation and inter-rater reliability. *Advances in Physiotherapy 7 (1) (pp 7-12), 2005 Date of Publication: 2005* 2005;(1):2005.

(222) Nolte CH, Malzahn U, Rakow A, Grieve AP, Wolfe CD, Endres M, Heuschmann PU. The german version of the satisfaction with stroke care questionnaire (sasc) for stroke patients. [German] Abstract only. Fortschritte der Neurologie Psychiatrie.78 (6) (pp 355-359), 2010.Date of Publication: 2010. [6], 2010. 2010.
Ref Type: Abstract

(223) Lata-Caneda MC, Pineiro-Temprano M, Garcia-Fraga I, Garcia-Armesto I, Barrueco-Egido JR, Meijide-Failde R. Spanish adaptation of the Stroke and Aphasia Quality of Life Scale-39 (SAQOL-39). *European journal of physical & rehabilitation medicine* 2009 September;45(3):379-84.

(224) Pang MYC, Lau RWK, Yeung PKC, Liao LR, Chung RCK. Development and validation of the chinese version of the reintegration to normal living index for use with stroke patients. *International Journal of Stroke Conference: World Stroke Congress 2010 Seoul South Korea Conference Start: 20101013 Conference End: 20101016 Conference Publication: (var pagings) 5 (pp 296), 2010 Date of Publication: October 2010* 2010;(var.pagings):October.

(225) Brooks DN, Hosie J, Bond MR, Jennett B, Aughton M. Cognitive sequelae of severe head injury in relation to the Glasgow Outcome Scale. *Journal of Neurology, Neurosurgery & Psychiatry* 1986 May 1;49(5):549-53.

(226) Maas AIR, Braakman R, Schouten HJA, Minderhoud JM, van Zomeren AH. Agreement between physicians on assessment of outcome following severe head injury. *Journal of Neurosurgery* 1983 March 1;58(3):321-5.

(227) Lees KR, Bath PMW, Schellinger PD, Kerr DM, Fulton R, Hacke W, Matchar D, Sehra R, Toni D. Contemporary Outcome Measures in Acute Stroke Research: Choice of Primary Outcome Measure. *Stroke* 2012 April 1;43(4):1163-70.

(228) Ali M, Bath PMW, Curram J, Davis SM, Diener HC, Donnan GA, Fisher M, Gregson BA, Grotta J, Hacke W, Hennerici MG, Hommel M, Kaste M, Marler JR, Sacco RL, Teal P, Wahlgren NG, Warach S, Weir CJ, Lees KR. The Virtual International Stroke Trials Archive. *Stroke* 2007 June 1;38(6):1905-10.

(229) Burleigh E, Reeves I, McAlpine C, Davie J. Can doctors predict patients' abbreviated mental test scores. *Age and Ageing* 2002 July 1;31(4):303-6.