Improving outcome assessment for clinical trials in stroke

Dr Terence J Quinn MBChB,
BSc Med Sci. (hons), MRCP (UK)

Thesis Submitted in fulfilment of the requirements for the degree of MD

Cardiovascular and Medical Sciences
Faculty of Medicine
University of Glasgow

Submitted January 2010

Abstract

Clinical trials are at the centre of advances in our understanding of stroke and its optimal treatment. In this thesis the uses and properties of outcome assessment scales for stroke trials are described, with particular attention given to the modified Rankin Scale (mRS).

Through comprehensive literature review I will show that mRS is the most frequently used functional outcome scale in clinical trials but efficacy of the scale is potentially limited by inter-observer variability. Using a "mock" clinical trial design I demonstrate that inter-observer mRS variability in contemporary practice is moderate (k=0.57). Adding these data to systematic review of published data, confirms an overall moderate inter-observer variability across ten trials (k=0.46).

Differing strategies to improve mRS reliability will then be described. I will outline development of a bespoke training package, international training scores across 2942 raters again confirms suboptimal reliability (k=0.67). A pilot trial using endpoint committee review of video recorded interviews demonstrates feasibility of this approach. Attempts to improve reliability by deriving mRS from data recorded in patients' hospital records are not successful (k=0.34).

In the final chapters I present a novel methodology for describing stroke outcomes – "home-time". This measure shows good agreement with mRS, except at extremes of disability. Finally to put mRS in a historical context, the career of John Rankin and the development of his eponymous scale is recounted.

Summary of Thesis Chapters

In **chapter one** a brief overview of outcomes assessment is provided. Guided by the principles of clinimetrics (the study of properties and uses of scales in clinical practice in particular multidimensional assessment scales), uses, limitations and properties of several functional outcome measures are described.

It has been argued that the clinical application of mRS is limited by substantial inter-observer variability. In **chapter two** a systematic literature review of mRS reliability studies is described. Ten studies of mRS reliability are presented. There is heterogeneity in reliability reported, overall reliability of mRS is suboptimal (k=0.46). The studies of mRS properties that best mimic a contemporary clinical trial (multiple observers across different sites grading multiple patients) are shown to demonstrate particularly poor reliability of mRS.

A number of potential functional assessment scales are available to trialists. In **chapter three**, differing scales and methodologies for assessing functional outcomes stroke trials are described. Six journals were chosen to represent high impact publications in Stroke Medicine, Neurology and General (Internal) Medicine, all were hand searched for trials describing functional outcomes in stroke survivors. One hundred and twenty-six articles were reviewed. Of forty seven outcome measures used, the most prevalent tool was the mRS (81 trials; 64%). Trialists continue to utilise instruments that are poorly validated. The majority of papers did not provide details on how their functional outcome assessments were administered (93; 73%). This heterogeneity in the use and

description of functional outcome measures in stroke trials will potentially compromise comparison and meta-analysis across studies, explicit description of methodology should be mandatory for all trials and greater rigour is desirable.

In **Chapter four** a mock clinical trial design is used to explore inter and intra-observer variability of the scale. Consenting stroke patients had mRS performed by two independent assessors, with the second interviewer in the pair assessing the patient blinded to colleague's score. For each patient assessed, one rater was randomly assigned to video record their interview. After three months this interviewer reviewed and re-graded their original video assessment. Across 100 paired assessments, inter-observer agreement was found to be moderate (k=0.57); use of a structured interview (n=49) did not substantially improve reliability (k structured=0.50; k unstructured=0.64). Intra-observer variability was good, but less than would be expected from previous literature (k=0.72). These results suggest that there remains substantial inter-observer variability in mRS grades awarded even when administered by experienced researchers.

A previous criticism of mRS has been the lack of guidance on how to assess and grade patients. In **chapter five**, the development of a video based mRS teaching and certification resource is outlined. Formal assessment of training involved grading of real-life cases. After training, most trainees (90%) achieved certification in mRS assessment. The majority (85%) of investigators who did not reach an acceptable score on initial testing achieved certification after further exposure to the package. Mass training in mRS assessment for clinical trials is possible. Acceptability of the training has been demonstrated by its successful use in international stroke trials.

In **chapter six** the training and certification data from the DVD resource are used to explore reliability of mRS across a large cohort of researchers. In total, 2942 assessments from 30 countries were analysed. Overall reliability for mRS grading was good (k=0.67) with substantial heterogeneity across countries. Native English language had little effect on reliability. Within the United Kingdom, there was no significant variation by profession.

In **chapter seven** data from a study of deriving mRS from patient's hospital records are presented and discussed. Fifty sequential patients attending the cerebrovascular outpatient clinic were included. Two independent, blinded clinicians, trained in mRS, assessed case-records to derive mRS. They scored "certainty" of their grading on a 5-point Likert scale. Agreement between derived and traditional face to face mRS was calculated using attribute agreement analysis. Case-record appraisers were poor at deriving mRS (k=0.34). Accurate mRS cannot be derived from standard hospital records. Direct mRS interview is still required for trials.

In **chapter eight** a pilot trial of group assessment of recorded mRS interviews is reported. Remote assessment of endpoints by adjudication committee is commonplace in contemporary trials and has potential to improve data quality. Using patient videos from the study presented in chapter four, at three months after initial mRS assessment a panel experienced in use of mRS graded the videos assigning individual scores. This process was repeated again after a further three-month delay. Individual assessments were recorded and then the group discussed cases with final grading based on consensus. Inter-observer and intra-observer variability of remote assessment of video mRS was quantified

using attribute agreement analysis. Inter-observer variability of individual video mRS assessors was moderate (*k*=0.67); intra-observer variability was moderate also (*k*=0.64). There was significant agreement between consensus group mRS and standard mRS. Remote assessment of mRS by adjudication panel is feasible and has acceptable reliability and validity. Further studies using this video based approach are warranted.

In **chapter nine,** using data from the "GAIN" trial, relationships between duration of stay in the patient's own home or chosen residence post stroke – "Home-time" and other functional outcomes are explored. Baseline data were from 1717 of 1788 patients; functional outcomes included NIHSS; Barthel Index (BI) and mRS. Using analysis of variance with Bonferroni contrasts of adjacent categories, a significant association between increasing Home-time and improved mRS scores was found. The relationship held across all mRS grades except 4-5. Home-time offers a robust, useful and easily validated outcome measure for stroke, particularly across better recovery levels.

To allow discussion of mRS and other outcome assessments to be placed in a historical context, **chapter ten** outlines the development of the original Rankin scale and its creator Professor John Rankin. Using historical documents and publications, Rankin's pioneering work in the nascent speciality of Stroke Medicine is described as well as the genesis of his eponymous scale.

Table of Contents

Chapter two

Chapter three

Chapter four

Chapter five

Chapter six

Chapter seven

Chapter eight

Chapter nine

Chapter ten

List of Tables

List of Figures

List of Accompanying Materials

Appendix A: The modified Rankin Score.

Appendix B: The Barthel Index.

Appendix C: Manuscripts reviewed and excluded from systematic study of modified Rankin Scale reliability.

Appendix D: Complete list of functional outcome assessments used in contemporary stroke literature.

Appendix E: Pro-forma for assessment of video mRS.

Appendix F: Original Rankin Stroke Scale and derivations.

Appendix G: Application for funding to support multi-centre study of video based mRS.

Inside Cover: Digital video-disc of modified Rankin Scale training resource.

Acknowledgement

The work presented in this thesis represents the product of a number of successful collaborations and completion would not have been possible without the valued contribution of a number of colleagues.

I wish to thank my supervisors Dr Matthew Walters and Professor Kennedy Lees for the opportunity to take a lead on the mRS based projects and for their advice and support - which began during my time as a junior doctor in the Western Infirmary; was frequently needed during my time as a research fellow and continues to be appreciated.

I am grateful to the Western Infirmary stroke consultants; Professor G. McInnes and Dr P. Semple who provided clinical guidance, research tips and frequent anecdotes; and allowed me to pursue my research in their busy unit. I am particularly thankful for the mentoring and encouragement provided by my advisor Professor John Reid. I appreciate also the input of my colleague and friend Dr Jesse Dawson, with whom I shared office space; research ideas and many lattes.

Much of the work presented was based in the acute stroke unit of the Western Infirmary and I am indebted to the team who keep the clinical and research activity of the unit running so smoothly. In particular I am thankful to Mrs Pamela McKenzie who managed to keep the activity of the unit efficient and organised, in stark contrast to the chaos of my own research office. I am indebted to the input and good humour of the unit research sisters: Lesley

Campbell, Elizabeth Colquhoun and Belinda Manak; Mrs Karen Shields for assistance with imaging; Sister Sarah Dorward and the other ward staff of the unit. No research work would have been possible without the cooperation of the stroke unit patients and I thank them also.

Finally I acknowledge my wife Gina, who I love dearly and who has tolerated more modified Rankin Scale based conversations than most spouses would accept. I dedicate the thesis to her.

Author's Declaration

The work contained in this thesis was carried out during my two year tenure as research fellow in the University of Glasgow Department of Cardiovascular and Medical Sciences (August 2006 – August 2008).

All of the studies reported herein have either been published or submitted to journals for consideration of publication. A list of these papers and other published abstracts relating to the work reported is included. All of the work reported in this thesis was undertaken by me, with the assistance of a number of colleagues who are formally acknowledged below. All of the statistical analyses herein were performed by me and the manuscript was written solely by me.

Signed…….

Details of the collaborative contribution of colleagues

Chapter two: Systematic review of mRS reliability studies.

I was responsible for design of the study and literature search protocol. I was responsible for the original literature search, with Dr Jesse Dawson performing the second independent search. I was responsible for final trial selection, data extraction and description.

Chapter three: Functional outcome measures in contemporary stroke trials.

I was responsible for design of the study and literature search protocol. I was responsible for the original literature search, with Dr Jesse Dawson performing the second independent search. I was responsible for data extraction and description.

Chapter four: Exploring the reliability of the modified Rankin Scale.

Advice on optimal video recording hardware was helpfully provided by University of Glasgow Media Services. The study involved several raters who performed mRS assessments, these raters were our departmental research nurses: Lesley Campbell, Elizabeth Colquhoun and Belinda Manak; and medical staff of the Stroke Unit: Professor Kennedy R Lees, Dr Matthew R Walters and Dr Jesse Dawson. All these raters assisted with initial patient selection and recruitment.

Chapter five: Initial experience of a digital training resource for modified Rankin Scale assessment in clinical trials.

Colin Brierley, Nigel Hutchins, Barbara Farmer and their team at University of Glasgow Media Services provided technical assistance with the video recording of interviews. Sarah Dorward (stroke liaison sister) assisted in selection of suitable

patients for video interview. Professor Kennedy R Lees and Doctor Hans-Goran Hardemark assisted with initial mRS rating. The development of the video resource was partly supported by an educational grant from AstraZeneca. I was responsible for data collation and interpretation of these data.

Chapter six: Variability in modified Rankin Scale scoring across a large cohort of international observers. Mrs Pamela Mackenzie provided excellent assistance in collecting and inputting raw training data from various centres. I am grateful to all the trialists who contributed to the mRS training project, in particular the SAINT I and CHANT steering committee and investigators, Dr Algirdas Kakarieka and the AstraZeneca trial monitors who helped in administering many of the assessments and all other researchers who have completed the certification exercise and / or provided comments on the training resource.

Chapter seven: Deriving modified Rankin Scale grades from medical case records. Doctors Gautamanada Ray and Sari Atula performed the case record analysis.

Chapter eight: Pilot trial of remote adjudication for modified Rankin Scale Assessment in Clinical Stroke Trials. Again, the study involved several raters who performed mRS assessments - Departmental research nurses: Lesley Campbell, Elizabeth Colquhoun and Belinda Manak; and medical staff: Professor Kennedy R Lees, Dr Matthew R Walters and Jesse Dawson. All raters assisted with initial patient selection and recruitment.

Chapter nine: Time spent at home post stroke: "Home-time" a meaningful and robust outcome measure for stroke trials. Data from the GAIN International study were used for this analysis, GAIN was sponsored by GlaxoWellcome (now GlaxoSmithKline). Jennifer S Lees and Tau-Pin Chang assisted with initial data cleaning, data input and statistical analyses.

Chapter ten: Dr John Rankin; his life, legacy and the 50th anniversary of the Rankin stroke scale. I am grateful to David Null of University of Wisconsin Archives for his excellent help in sourcing relevant manuscripts and materials.

Publications and Presentations related to the thesis

All data presented in the thesis have been presented as oral or poster platform at scientific meetings including: American Stroke Association International Stroke Conference; European Stroke Conference; UK Stroke Forum and British Geriatric Society annual scientific meeting.

All studies included in the thesis have been submitted for publication in peer reviewed scientific journals.

Publications

Chapter one:

Components of the discussion of functional outcomes presented in chapter one have been published as:

a) Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the modified Rankin Scale. Stroke. 2008; 38: 144-5.

b) Quinn TJ. The uses and limitations of functional outcome assessment instruments in stroke trials B J Neuroscience Nursing 2008; 4: 60-66.

The data described in all other chapters have been published in whole or in part in peer reviewed journals.

Chapter two:

Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the Modified Rankin Scale. A Systematic Review. Stroke. 2009; 40: 3393-5.

Chapter three:

Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. International Journal of Stroke. 2009; 3: 200-5.

Chapter four:

Quinn TJ, Dawson J, Walters MR, Lees KR. Exploring the reliability of the modified Rankin scale. Stroke. 2009; 40: 762-6.

Chapter five:

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; 38: 2257-61.

Chapter six:

Quinn TJ, Dawson J, Walters MR, Lees KR. Variability in modified Rankin scoring across a large cohort of international observers. Stroke. 2008; 39: 2975-9.

Chapter seven:

Quinn TJ, Ray G, Atula S, Walters MR, Dawson J, Lees KR.

Deriving modified Rankin scores from medical case-records.

Stroke. 2008; 39: 3421-3.

These data were chosen by the International Standards Organisation, technical subcommittee on attribute agreement as a good example of statistical analysis of inter-observer agreement. Pending permission from "Stroke" the thesis chapter will be presented in their manuscript describing best practice and standards in attribute agreement analysis.

Chapter eight:

Quinn TJ, Dawson J, Walters MR. Lees KR. Initial experience with video based modified Rankin assessment. Cerebrovasc Dis 2007; 23,s115 (abstract)

Chapter nine:

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; 39: 231-3.

Chapter ten:

Quinn TJ, Dawson J, Walters M. Dr John Rankin; his life, legacy and the 50th anniversary of the Rankin Stroke Scale. Scottish Medical Journal. 2008; 53: 44-7.

List of Abbreviations

ANOVA.....Analysis of Variance

BI.....Barthel Index

CHANT.....Cerebral Haemorrhage Acute NXY Trial

CI.....Confidence Interval

CNS.....Canadian Neurological Scale

CONSORT.....Consolidated Standards for Reporting Trials

COSTAR.....Collaborative Stroke Audit and Research

CPMP.....Committee for proprietary medicinal products

CSS.....Canadian Stroke Scale

CT.....Computerised Tomography

DESTINY..... Decompressive Surgery for the Treatment of Malignant Infarction

DIAS.....Desmotoplase in Ischemic acute Stroke

DOH.....Department of Health

DVD.....Digital Video Disk

ECASS.....European Cooperative Acute Stroke Study

FAI.....Frenchay Activities Index

GAIN.....Glycine Antagonist (Gavestinel) In Neuroprotection

GP.....General Practitioner

GOS.....Glasgow Outcome Scale

ICC.....Intra-class Correlation Coefficient

ICH.....Intra-Cerebral Haemorrhage

IMP.....Investigational Medicinal Product

IQR.....Intra-Quartile Range

IST……International Stroke Trial

JAMA……Journal of the American Medical Association

k……Kappa coefficient

kw……Weighted kappa coefficient

LACS……Lacunar Stroke

LDL……Low Density Lipoprotein

LHS……London Handicap Scale

MCA……Middle Cerebral Artery

MeSH……Medical Subject Headings (National Library of Medicine)

MI……Myocardial Infarction

MOOSE……Meta-analysis and Observational Studies Epidemiology

MRI……Magnetic Resonance Imaging

mRS……Modified Rankin Scale

NEJM……New England Journal of Medicine

NIHSS……National Institutes of Health Stroke Scale

OAST……Optimising the analysis of stroke trials

OCSP……Oxford Community Stroke Project

OHS……Oxford Handicap Scale

PACS……Partial Anterior Circulation Stroke

PI……Principle Investigator

POCS……Posterior Circulation Stroke

PRISMA….Preferred Reporting Items for Systematic Review and Meta-Analyses

QOL……Quality of Life

RCT……Randomised Controlled Trial

RS……Rankin Scale

SAINT……..Stroke Acute Ischemic NXY Trial

SD……..….Standard Deviation

SI……..Structured Interview

SIS……..Stroke Impact Scale

SMJ……...Scottish Medical Journal

SSS……..Scandinavian Stroke Scale

STICH…….....Surgical Trial in Intra-cerebral Haemorrhage

TACS……Total Anterior Circulation Stroke

TIA…….……..Transient Ischemic Attack

UK……United Kingdom

VHS…….……Video Home System

WHO……..World Health Organisation

Chapter one

Assessment scales for stroke trials

Functional outcome assessment in stroke trials

Stroke represents a substantial and increasing global health problem. Cerebrovascular diseases are the third leading cause of death and the single greatest cause of disability in most Western countries.(1) Globally the burden of stroke is greater still, with the majority of incident cases in the developing world.(2) The economic burden of stroke is substantial, acute and chronic care of stroke is estimated to consume greater than 5% of many countries total healthcare budget.(3)

Prevention strategies and acute and longer term interventions for stroke patients have changed considerably in the last 20 years, in part driven by the increasing evidence base for both acute and rehabilitative strategies.(4) To inform and improve the practice of stroke medicine there has been an exponential increase in clinical trials.(5) A recent overview of randomised controlled trials in the field of acute stroke, reported an increase in the number of registered randomised controlled trials (RCT) per decade from 3 in the 1950's to 99 in the 1990's with corresponding increases in patient numbers per trial and improvements in overall quality of trial methodology.(6)

In any field of medicine, clinical trials are designed to compare efficacy of two or more interventions. To quantify the differences between treatment strategies requires some measure of effect. Treatment effect or outcome can be measured in several different ways, with each approach having advantages and disadvantages.

In many situations a direct measure may be appropriate. For example in a basic comparison of antihypertensive therapies the outcome measure of choice may be a direct measure of blood pressure. Such an approach is attractive in terms of immediacy and simplicity and lends itself to relatively straightforward comparative analysis. However, as experience of clinical trials has developed it has become increasingly apparent that direct measurement of "bio-markers" does not always correlate with clinical outcome.(7) This is especially true in the field of cardiovascular medicine, with multiple examples of well conducted clinical trials that reported significant benefits in terms of a relevant bio-marker, with either no corresponding clinical effect or even an unexpected deleterious clinical effect.(8) As example, the lipid lowering agent Ezetimibe has been shown to significantly lower mean serum levels of low density lipoprotein (LDL) but has shown no benefit in vascular risk reduction and may in fact be associated with increased risk of mortality.(9;10)

With this in mind, measurement of outcomes more directly relevant to patients becomes more attractive. As a primary aim of most medical interventions is to keep patients alive, the archetypal "hard" clinical outcome is mortality. Using clinical outcomes such as mortality or incidence of event provides unambiguous data that are easy to collate and analyse. Selection of optimal outcome(s) to use as clinical trial endpoint is more problematic. Ultimately, clinical trials are designed to test potential treatment benefits for patients. In cerebrovascular medicine, the physical, psychological and social cost of a stroke is only poorly represented by traditional trial endpoints.(11;12) Even "hard" outcomes such as mortality or event rate provide a poor measure of the global effect of a stroke.

In fact, it has been shown that most patients would rather be dead that suffer a disabling cerebrovascular event.(13) Thus an outcome measure where death is "negative" but disabling stroke is "positive" will give little meaningful data.

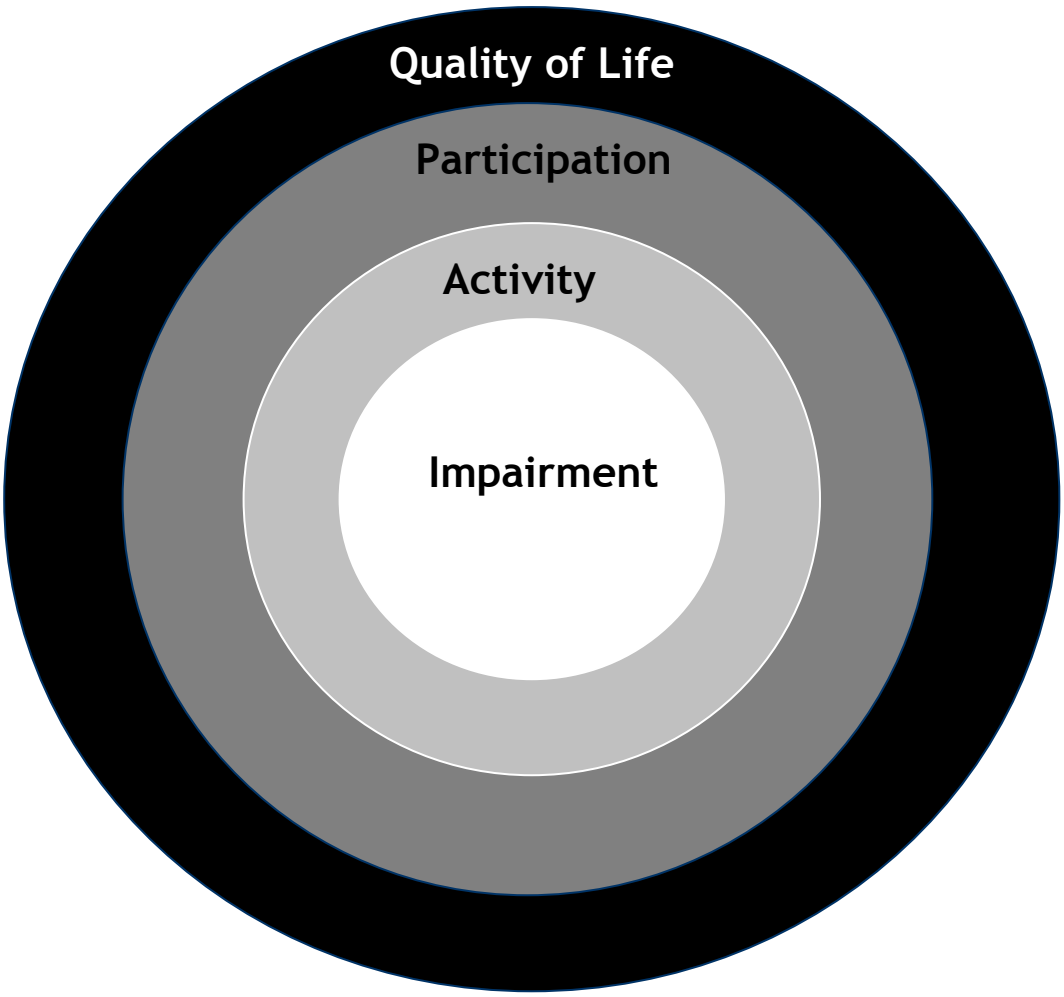
To better represent treatment effects, a number of outcome scales that make some measure of patient functioning have been developed and are now commonly used in clinical stroke trials. The importance of functional outcomes as the optimal measure of clinical effect in stroke trials has been recognised by regulatory authorities both in America(14) and Europe.(15) In this regard other cardiovascular disciplines could learn from stroke medicine – a recent analysis of clinical trials in diabetes mellitus suggested that majority of published trials continue to use bio-markers as endpoint and use of outcomes immediately relevant to patients was infrequent.(16)

Functional assessment scales in stroke

Assessment scales are designed to represent and measure quantities, qualities or categories. Responses are collected across a variable number of domains, standardised against pre-specified grades and can then be quantified and collated. These data can then be subject to statistical analysis. Functional outcome scales use this process to describe aspects of patient ability and wellbeing.

Post stroke functional recovery can be described in a number of domains and this is reflected in the large number of assessment scales available.(17) The WHO international classification of functioning, disability and health proposes a conceptual framework that can be used as an aid to classification of outcomes data.(18) Patient recovery can be described in terms of physical impairment, functional activity (formerly disability) or societal participation (formerly handicap). Assessment scales describing each of these domains are frequently used in contemporary stroke literature, examples from each domain include: NIHSS (National Institutes of Health Stroke Scale)(19) as a measure of impairment; mRS (modified Rankin Scale)(20) as a measure of functional activity and London Handicap Scale (21) as a measure of participation. A further domain that trialists have attempted to measure is that of quality of life (QOL).(22) QOL scales build on measures of societal participation and attempt to assess physical, mental, societal and spiritual aspects of a patient's condition.(12;23) (Figure 1)

Figure 1: Venn diagram illustrating levels of functioning.



Thus the need to robustly measure functional outcome for clinical stroke trials is evident, however the optimal methodology to describe these data remains a matter for debate. The confusion regarding functional outcome measurement is in part a reflection of the large number of assessment tools available. The University of Washington stroke research centre describe 20 different outcome assessment scales in common use in stroke trials, these include stroke specific tools and more general neuro-rehabilitative tools. These selected instruments represent only a small proportion of the total number of scales potentially available. Reference texts relating to outcome measures in neurological disease describe many hundreds of tools that have previously been used in clinical trial settings.(24;25)

Describing properties of assessment scales

The optimal functional outcome measure to be used will depend on the clinical application or research question to be answered. Important properties of an instrument intended for use in a busy outpatient service will differ from the desired features of a tool for use in detailed research. The ideal outcome measure would be easy to administer; would show consistency after repeated use and across multiple users; would capture information relevant to both the patient and the trialist and be able to detect small changes over time.(26) No perfect outcome measure exists (or is likely to ever exist), however understanding of the properties of outcomes measures has increased and we can use this to comprehensively examine existing and novel assessment instruments.

Clinimetrics is the methodological discipline that focuses on quality of clinical measurements.(26) Outcome scales are traditionally assessed in terms of validity, reliability and responsiveness and these will be described in turn. Other important properties of an outcome measurement include feasibility (in the desired setting); patient acceptability and cost benefit both in terms of economic and time resources. Assessment of these latter properties is by their nature more subjective and will vary according to the outcome measure and its proposed usage.

Validity

Validity is defined as the relationship between the concept to be measured and the scale used for assessment.(27) Validity can be defined using several inter-related and complimentary methods. Certain key measures of validity for example face validity are subjective and do not lend themselves to objective measurement.

Criterion validity – does the scale agree with a known "gold standard".

Convergent validity - does the scale of interest agree with other instruments that purport to measure the same or similar outcomes.

Construct validity – is there reasonable relationship between the scale and factors known to influence the property to be measured. In stroke outcome assessment, a functional assessment tool should show a correlation with factors such as size of infarct; pre-morbid disability etc.

Face validity – do the outcome data generated by an outcome scale make sense and agree with consensus opinion.

Responsiveness

Responsiveness is defined as the ability to detect meaningful change over time.(27) A stroke scale should be able to detect changes in patient ability as they progress through rehabilitation and recovery. The minimal degree of change that is felt to be clinically significant will vary according to the trial. However even clinically modest improvements in functioning can have substantial meaning to patients and be important at a population level.(28) Increasing responsiveness is by its nature often at the cost of increasing complexity. As example, with its six grades the Glasgow Outcome Scale (GOS)(29) is less responsive to change than the Barthel Index (BI)(30) with its hundred point scoring. However, any change in GOS is clearly of clinical relevance while this may not be true of a single point change in BI.

Reliability

Reliability is a measure of both internal consistency in multi-item scales and of the reproducibility of repeat scoring by the same observer (intra-observer reliability) or between scorers (inter-observer variability).(31) Measures of reliability should assess both reproducibility among the observer(s) and consistency across components of the scale.

The optimal statistical methods to quantify and analyse reliability data remain contentious. Inter and intra-observer reliability has traditionally been described in the medical literature using the kappa (*k*) statistic(32) – a measure of agreement across a number of observers for non-parametric scales. *k* statistics are calculated based on the observed proportion of agreement (Po) and the proportion of agreement expected by chance (Pe) where $k=(Po-Pe)/(1-Pe)$.(33)

Using this equation *k* can theoretically take any value from -1 to 1 where *k*=1 defines perfect agreement between assessors, while *k*=0 defines no agreement other than that expected by chance. Standard definitions of poor (*k* = 0-0.20); fair (*k* = 0.21 – 0.40); moderate (*k* = 0.41 – 0.60); good (*k* = 0.61 – 0.80) and very good (*k*= 0.81 – 1.00) agreement are now accepted.(34) For clinical use a reliability of *k*=0.61 or greater has been arbitrarily chosen as "acceptable".(32) In ordinal hierarchical scales an inter-observer difference of more than one grade in either direction implies a greater degree of variability than single unit change. For this reason some authors have used weighting of kappa statistics (*k*w) to better represent the size of disagreement between observers, most commonly weighting is "quadratic".(35)

Although kappa statistics have proven popular in the biomedical literature, solely using this method to describe variability has a number of limitations. Kappa statistics are dependent on the number of observers and categories within the scale. This makes for problematic comparative analysis of kappa statistics from different populations and studies.(32;34) It has also been argued that the basic assumptions underlying kappa statistics will not be met in a "real life" clinical trial setting. For example, kappa statistics assume complete observer independence, a situation that may not be met in a single centre trial, where observers are likely to work together and be aware of each others criteria for assessing recovery.(36)

The importance of optimal outcome assessment

An understanding of the clinimetric properties of a scale is of more than academic interest. Use of an inappropriate tool will jeopardise study quality and invalidate results, even if suitable rigour is exercised in all other elements of the trial. The need for adequate validity of an outcome tool is intuitive, a functional assessment that does not measures the clinical property it is designed for will clearly provide meaningless data. Similarly, the importance of scale responsiveness and the ability to detect meaningful change even if modest has already been discussed.

The deleterious effects of poor outcome reliability are less intuitive but potentially the most relevant for contemporary stroke trials. Poor reliability can substantially weaken the results of an otherwise well conducted trial. High levels of inter-observer variability in assessment of a trial endpoint are a signal that certain outcomes are being misclassified. Such misclassification will increase the likelihood of both type I (false positive) and type II error (false negative) and may ultimately decrease statistical power.(28) "Real life" instances of the detrimental effects of misclassification on otherwise well conducted trials can be found in recent high profile studies from numerous biomedical disciplines. As example, in a national trial of pneumococcal vaccine, modest misclassification of the cause of death (the trial's primary endpoint) reduced trial power by 40%.(37) In a recent neuro-trauma study, erroneous misclassification of outcome substantially reduced the observed treatment effect, reanalysis correcting for this misclassification revealed a significant benefit of intervention.(38)

Reducing variability is not only of significance when potential treatment effects are missed. Reduction of endpoint misclassification will favourably impact upon trial power, reducing total number of patients required and thus reducing overall running cost. In a heterogeneous clinical condition such as stroke, large numbers of patients are required to detect modest but meaningful treatment effects.(39) In the planning and statistical "powering" of many recent studies, treatment effects have been overestimated and trials have ultimately been underpowered.(40) These issues are compounded by the difficulties many stroke trialists have faced in recruiting to target.(40) It must also be remembered that ideally robust trial results should be achieved using the fewest trial participants as possible, thus reducing exposure to a risky intervention or not denying an effective treatment to a control group. Thus, there are clinical, economic and ethical reasons to design trials that are adequately powered with the fewest possible patients. Improving reliability of endpoint data should offer a method for achieving this aim.

To put this discussion of study endpoint quality in context, it is worth noting that several recent stroke trials (SAINT II(41); DIAS(42)) have been characterised by neutral results for compounds with good scientific and pre-clinical data. It is of course possible that interventions such as neuro-protectants and novel thrombolytic agents simply have no efficacy in man. However, the above discussion suggests a second possible explanation - that in some cases important beneficial treatment effects may have been lost through suboptimal endpoint data collection and analysis.

Prevalent functional outcome scales

For the purpose of this thesis, functional outcome scales are those assessment tools used as endpoint in trials and that purport to measure more than mortality or disease state. As discussed previously there are large numbers of functional assessment scales that have been used; continue to be used or have potential to be used in clinical stroke trials. Using the WHO international classification system, well recognised examples of scales that purport to measure impairment; activity and participation will be described paying particular attention to strengths and weaknesses of differing assessment approaches. The scales chosen for discussion represent those tools commonly used in acute stroke trials.

The National Institutes of Health Stroke Scale

The National Institutes of Health Stroke Scale (NIHSS)(19) is an example of a stroke specific impairment grading scale. The scale was originally developed by neurologists as an aid to non-specialists in the initial assessment of stroke severity. NIHSS uses clinical examination to measure limitation across a number of pre-specified physical domains. Scores for individual components of the assessment are summed to give a total score between 0 (no objective deficit) and 43.

NIHSS has grown in popularity and in many centres NIHSS assessment is a routine component of initial stroke “work up”.(43) Advantages of the NIHSS are ease of administration and favourable reliability and responsiveness.(44;45) Well validated training packages are available for NIHSS use and certification in competent use of the scale is a prerequisite for researchers in many acute stroke trials.(46)

There are certain well documented weaknesses in the NIHSS that limit clinical utility of the scale. The grading of the scale places an emphasis on left hemisphere damage and makes little assessment of cranial nerve function. Thus with standard NIHSS scoring a patient with a posterior stoke can achieve a minimal score despite significant impairment.(47) Many of the items incorporated in the NIHSS require the patient to be alert and so the scale may not differentiate impairment in the most severely affected strokes where patients may have fluctuant levels of consciousness.(17) As with any impairment scale, by restricting its focus to physical functioning, NIHSS is less appropriate for assessment in the longer term. At 90 days post stroke, many patients may have recovered motor function in their limbs but fewer will have returned to work or previous past-times.

The Barthel Index

The Barthel Index (BI)(48) is a general assessment scale used to quantify activities of daily living. The scale was originally developed for use in long stay hospitalised patients to assess care needs, but has subsequently been used in many other areas including functional assessment post stroke. The BI is scored according to ability and as such requires direct observation as well as historical data. Patients are scored across a number of areas including dressing; toileting and mobility. Scores are from 0 to 100 (or 0 to 20 if the modified BI is used).

The BI is one of the best known activity scales, it is straightforward to administer and has been shown to predict long term outcomes, patients with BI scores totalling less than 40 are unlikely to return to independent living. There

is a comprehensive literature on clinimetric properties of the BI. It is generally accepted that the scale has good reliability(49), although some authors have found high inter-observer variability in older patient groups (50) – a potential concern for stroke trialists as cerebrovascular diseases predominantly affect older cohorts.

The clinical utility of the BI as trial endpoint is limited by well documented "ceiling" and "floor" effects. (51) Patients can achieve maximal scores on the BI yet still have substantial residual impairment, while patients scored at the minimum of the scale can still show meaningful functioning. Thus the BI is less responsive to the range of clinical improvement expected in a clinical stroke trial setting. With its emphasis on activities of daily living, BI has also been criticised for disregarding other important aspects of recovery. (52) BI scores must be interpreted in the context of timing of administration. (53)The "floor" effects of the BI are particularly relevant to acute stroke trials as in the first hours post event most patients will be bed bound and requiring nursing care. While following discharge home the basic activities measured by the BI provide little useful data on extended activities of daily living required to fully function and integrate in community environments.

Stroke Impact Scale

The Stroke Impact Scale (SIS) is a novel multi-level outcome assessment tool that was designed to assess a broad range of domains including physical functioning; cognition and societal participation. (54) The development of the SIS was informed by feedback from patients, their carers and therapists as to their perception of the most important aspects of stroke recovery.

Unlike many traditional stroke assessment scales, the clinimetric properties of the SIS were robustly tested prior to dissemination and the scale continues to be modified to improve its use.(54) SIS purports to measure certain domains not well quantified by other existing scales and as such provides one of the more comprehensive assessments stroke recovery. However, despite the literature supporting the SIS it has been used infrequently in clinical trials and its generalisability as a functional outcome for stroke trials remains to be established.

The modified Rankin Scale

The modified Rankin Scale (mRS) is a stroke specific measure of functional recovery. The precise meaning of mRS data has been debated.(55) Certainly the scale offers a more comprehensive assessment than other activities of daily living scales such as the BI. Some have argued that as mRS includes constructs such as ability to return to work, it represents a measure of societal participation. However with an emphasis on perambulation it offers at best only a limited assessment of participation. Most researchers now describe mRS as a "global disability" measure.

The mRS grades outcome using an ordinal hierarchical scale that ranges from potential scores of 0 (no symptoms) to 5 (severe disability). For clinical trial use an extra category of 6 (death) is often employed. Each category of the mRS describes a broad range of global disability, as such responsiveness to change is less than with other scales.(52) This can in fact be an advantage in the clinical trial setting as a single point change is likely to be clinically important.

A growing literature describing clinimetric properties of the mRS is available.(31) Convergent validity has been demonstrated by health care economics studies, with incremental mRS associated with significant increases in length of admission and cost of episode of care.(56) There is a strong association between mRS and other functional outcome measures frequently used in clinical trials including NIHSS; BI; GOS and quality of life measures.(52) However mRS is not directly equivalent to these other activity measures, it has been shown that scores on a scale such as BI are not easily transformed into mRS grades.(57) In terms of construct validity mRS scores have been found to have a reasonable agreement with several measures known to influence outcome following stroke including final infarct volume on imaging(58) and recanalisation score following thrombolysis.(59) Importantly, mRS at 90 days has been shown to be an accurate predictor of longer term functional outcome; nursing home placement and mortality,(60) making 3 month mRS a useful marker of future outcome and thus a useful endpoint for clinical trials.

The principal limitation of the mRS is variability in grading. For a single level assessment scale such as mRS internal consistency is not an issue, however there is considerable potential for inter-observer variability in application and grading. A number of studies have attempted to quantify the reliability of mRS and will be discussed in the following chapter.

Administration of the mRS

There is no formal guidance on how best to administer the mRS. When first developed the scale was derived from a face to face interview with a research nurse.(20) However a number of methodologies for collecting mRS based outcome data have subsequently been used in trials.

Face to face interview

The mRS assesses disability through patient's historical descriptions of functional ability. Unlike other assessment scales it does not require patients to demonstrate evidence of their ability in physical or cognitive domains. Attempts to transform physical examination findings into relevant mRS grades have been described but this mixing of impairment and disability measures have not proven popular in the stroke literature.(61)

It makes intuitive sense that a scale so dependent on historical information is graded using direct patient interview and much of the literature concerning clinimetric properties of mRS is based on traditional face to face interview. However, differing methodologies for conducting the mRS interview have been described. In recognition of the variability with which mRS data may be elicited by patient interview and the further potential for variability in how these data are interpreted by the assessor – a structured approach to mRS interview has been proposed.(62) Various structured mRS interviews have been used, most using a "checklist" for data gathering. Such an approach should provide a more comprehensive and systematic assessment of disability, however practical results when structured mRS has been used in the field have been conflicting.(63;64) Further heterogeneity in face to face interview is seen in

choice of interviewer. Interviewers have been used from a variety of disciplines and backgrounds including physicians (from differing specialities); research nurses; medical students and non-clinical interviewers.

Telephone / postal mRS

As the mRS does not demand physical examination, remote assessment by telephone interview or postal questionnaire should be possible and may be preferable in terms of simplicity and convenience for both researchers and patients. No robust data could be found on the properties of postal based mRS. Clinimetric studies of postal versions of other stroke scales have been described(65;66), although we should be cautious in directly extrapolating from these other scales. In a comparison of postal and interviewer administered versions of the GOS (65) there was overall moderate to good reliability; however a study using a postal version of the Stroke Impact Scale found reasonable clinimetric properties but cost savings of the postal approach were offset by high rates of non-response from participants.(66)

Telephone assessment is commonplace in stroke and acceptable reliability of telephone based assessment has been demonstrated for scales measuring stroke free status(67) and cognitive impairment(68). Studies of telephone based disability assessment, particularly mRS have yielded conflicting results. In a German study of BI assessment, reliability of telephone and postal versions of the scale were equivalent and showed excellent agreement with traditional assessment.(69) For mRS, independent groups have demonstrated results ranging from poor reliability (k=0.30)(64) using trained telephone assessors and structured interview to good reliability (k=0.74), although results of this later

analysis are likely overestimated as single dichotomised mRS scoring was used.(70) Drawing firm conclusions from this heterogeneous literature is difficult, however it would seem that direct interview is preferable to indirect assessments of mRS.

Assessing mRS from interview of proxy

Simple face to face interview with stroke survivors is not always feasible in a clinical trial setting: cognitive impairment, speech disorder or inability of the patient to attend a research centre can complicate assessment. Use of proxies to determine mRS is often used in such situations. Again there are little data on the properties of such an approach and any understanding of the clinimetrics of proxy use must be extrapolated from studies of other stroke assessment scales. In a direct comparative study of patient and corresponding proxys' scores on BI and Frenchay Activities Index (FAI) there was moderate agreement.(71) In a comparison of proxy and patient responses to the SIS, there were significant differences in response for several key variables with proxies tending to over-score disability.(72) Agreement was best for observable physical domains. In other studies of disease specific neurological scales it has been suggested that use of proxys may bias results, specifically family and informal carers may overestimate ability while health care workers and formal carers underestimate ability.(73)

Statistical analysis of mRS outcomes data

Just as there is heterogeneity in methodology for collecting mRS outcomes data, so there is heterogeneity in the way such data are analysed. In choosing a primary endpoint and corresponding statistical analysis trialists must endeavour to analyse data in a fashion that is both statistically appropriate and that will allow for detection of treatment effects with the fewest patients recruited.

A popular method of analysing disability endpoints has been to dichotomise data into "favourable" and "non-favourable" outcomes.(74) Many early stroke trials used a BI of greater than 60 to define patients with "good" functional outcome.(75) With mRS a variety of cut offs have been used to define good 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, the baseline disability of the cohorts under study and the level of recovery felt to be important not to "miss". 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 thrombolytic therapy, as better functional outcomes are expected this is reflected in the use of mRS 0-1 (no significant disability) to define treatment success.(75)

Using data from completed clinical trials allows us to explore the potential effects of differing cut-off points on overall trial results. Such an analysis has been completed and has shown that using a BI of 60 is an inefficient disability cut off, reducing overall power of a study.(76) This makes sense as the distribution of post stroke disability represented by BI assumes a "U shaped

curve" with few patients scoring BI in the middle range. Shifting outcome cut offs to a higher BI allows for better determination of effect with smaller sample size, however use of the mRS is more efficient still. In fact comparing BI scores of greater than 60 and mRS 0-1, sample sizes using the latter can be up to one fifth of that required for BI based assessment.(28)

With any dichotomous endpoint a number of patients enrolled in the study will never contribute to the final outcome. For example with a cut off of mRS 0-1, patients admitted with severe stroke may improve considerably and yet not reach the predefined level of acceptable recovery. Exclusion criteria could be modified to only enrol patients likely to contribute to final result, however this will prolong recruitment. In an attempt to make more efficient use of a complete trial data set, two or more endpoint cut offs can be created. Stroke trialists have successfully created and used such trichotomised outcomes assessment.(77)

More complex statistical analyses that make use of the complete spread of disability represented by mRS have been proposed.(78) In the recent neuro-protective trials of NXY-059 (SAINT I and II)(79) distribution of disability across mRS were compared in the two trial arms. The precise statistical calculations employed for this analysis have been criticised,(80) (81)however the underlying premise of measuring change across the complete scale remains valid. This "sliding" outcome assessment makes use of a concept of prognosis adjusted endpoint analysis. More complex prognosis adjustment can be performed based on markers of initial stroke severity. The beneficial effects on trial power of

prognosis adjusted endpoints have been demonstrated using data from the neuro-protective study GAIN.(78)

To allow detection of meaningful clinical effects with realistic patient numbers, trialists now often combine a number of “hard” clinical events into one global endpoint.(82) Such an approach should give a more comprehensive analysis of recovery, allowing for measures of impairment, activity and perhaps even participation in a single scale. Choice of constituent components is crucial for a global endpoint as the statistical power will be limited by the least efficient scale included. Although popular in the stroke literature(83) and recommended in certain guidelines,(84) the use of global endpoints has been criticised for mixing conceptually distinct recovery descriptors.(82) Ultimately such an approach provides an abstract result that is less immediate than a single well defined outcome.

Conclusion and hypotheses of the thesis

The continuing advancement of acute and chronic stroke care is dependent on ongoing collection of robust clinical trials outcomes data. In this regard functional endpoints are preferable for assessment of a disabling conditioning such as stroke. Of the many scales available to measure functional outcome across the domains of impairment; activity; participation and quality of life, the modified Rankin Scale (a measure of global activity) is arguably the optimal assessment tool. Although not originally developed for use in stroke trials, Rankin's original stroke scale and its subsequent modification have been used in a number of pivotal stroke trials. A literature on the clinimetric properties of the mRS exists and suggests good validity and responsiveness. A potential limitation of mRS is its poor reliability. Inter-observer variation and misclassification can ultimately impact on the power of a trial to detect a treatment effect. Various methods of statistical analysis can improve the strength of mRS as a trial endpoint. However, combining appropriate statistics with improved reliability of raw outcomes data would be more powerful still.

In this thesis, use of mRS in the contemporary stroke literature will be described along with summary of the available literature on mRS reliability. Reliability of mRS in a mock clinical trial setting will be explored along with the properties of mRS derived from patient case sheets. Finally the potential effects of methods to improve reliability of mRS will be described including use of video based training and offline assessment of disability.

Chapter two

Reliability of the modified Rankin Scale

– a systematic review

Introduction

As discussed in chapter one, a perceived weakness of mRS grading has been the potential for inter-observer variability in assigned grades. Problems of reliability may be inherent in the scale - mRS has been criticised for its relative lack of structure. Rankin grades encompass a broad range of potential outcomes and boundaries between grades are poorly defined relative to other outcome assessment instruments. Poor reliability is not a serious issue when the scale is used clinically, as one observer will chart functional change. However, mRS is principally used as a tool for clinical trials and in a large scale study involving many raters reliability becomes more important.

Attempts to quantify the reliability of mRS have been reported by several international groups, with conflicting results.(20;63) Clinimetric studies of outcome scales are often small in comparison to the clinical trials in which these outcome scales are used.(81) Thus single trials aiming to describe the properties of an assessment tool may not be adequately "powered" to answer the question of interest. In any field where the evidence is based on several small scale studies, systematic review and meta-analysis can provide useful summary data.

As previously discussed many methods of mRS administration exist and the optimal methodology for mRS assessment remains to be established. Traditional mRS interview is conducted "face to face" with study personnel.(20) Differing groups have attempted to improve reliability of stroke outcome scales using techniques such as standardising the interview process(62); providing training in use of the scale(46) or using video based technologies to allow remote

assessment of mRS.(85) To date the potential for such interventions to impact upon inter-observer variability of mRS is poorly described.

The recent years have seen an increase in the number of studies exploring clinical properties of mRS and other scales.(17) Previous reviews of stroke scales have concisely summarised important English language studies of mRS reliability.(31;86) However, clinical trial use of the mRS is international.(87;88) Similarly, mRS is often employed by research nurses and professions allied to medicine.(89) A contemporary, systematic review of the international literature including allied health care journals would compliment the ongoing work in this area.

I sought to systematically review the literature concerning mRS reliability, collate relevant studies and perform meta-analysis to better understand the reliability of mRS as a stroke outcome tool.

Methods

Two clinical researchers with a background in stroke (TJ Quinn, J Dawson) independently reviewed the literature. To date there are no specific guidelines on systematic review and meta-analysis of studies reporting clinimetric properties of scales. Throughout the process I adhered to the PRISMA(90) and MOOSE guidelines for conduct of systematic review and meta-analysis in clinical trials and observational studies.(91) In brief, these guidelines contain specifications for reporting analyses of observational studies in epidemiology, including search strategy, assessment of study quality and structuring discussion.

Eligibility Criteria and Study Selection

Participants: Study populations had to include human stroke survivors only. I used no restrictions for the mRS assessor and specifically did not exclude studies on the basis of background or training of observers.

Study methodology: All studies purporting to measure mRS reliability through patient interview (inter or intra-observer variability of mRS scoring) were reviewed with no specific restrictions on the basis of study design, intervention or language.

Outcomes: No restrictions on the basis of mRS assessment methodology were applied. Studies using mRS and derivatives: Rankin Scale (RS)(92) and Oxford Handicap Scale (OHS)(93) were included for review. However, only studies of mRS were included in the final analysis.

Search Strategy

A comprehensive battery of cross-discipline electronic databases were interrogated: AMED 1985 – 2008; British Nursing Index 1985 – 2008; CINAHL1981 - 2008; Embase 1980 – 2008; Health and Psychosocial Instruments 1985 - 2008; Internurse.com 1995 – 2008; Medline 1950 – 2008; PsychINFO 1967 – 2008.

Keywords were formulated using MeSH headings and study specific terms and were designed to be as inclusive as possible. : Stroke*; Cerebrovasc*; Modified Rankin*; Rankin*; Oxford Handicap*; Observer variation*

In addition to the electronic database search, contemporary reviews and key reference works were hand searched.(17;31) To identify studies not yet in print, proceedings of scientific meetings for the period Jan 2006 – Nov 2008 were hand searched (American Stroke Association - International Stroke Conference; European Stroke Conference; World Stroke Organisation - World Stroke Congress; British Geriatric Society – annual scientific meeting). Bibliographies of all retrieved articles were searched for further references and the process was repeated until no new articles were found.

Abstracts were reviewed for appropriateness to the study question. I retrieved the full text of any article that either reviewer believed may be relevant, data were extracted according to pre-specified criteria. Appropriateness of studies to be included was decided by consensus.

Where potentially relevant data were not available in the published manuscript, electronic or postal contact with the authors was attempted. For those studies not published in English, professional translation services were used.

Statistics

As discussed in Chapter one, reliability is both a measure of consistency in multi-item scales and a measure of reproducibility of results across differing test subjects or observers. As a single item scale, internal consistency of mRS can be assumed. Quantifying the reproducibility of repeat scoring between graders gives inter-observer variability.

Reliability of mRS van be quantified using a number of statistical techniques. Inter-observer variability is traditionally described using either kappa (*k*) statistics or simple percentage agreement between observers. Some studies have used quadratic "weighting" of kappa statistics (*k*w) to quantify degree of disagreement across the ordinal scale. To allow for comparison and where available data permitted, both *k, kw* and percentage agreement were derived from the included studies.

Based on previous work suggesting a beneficial effect of a structured interview approach(62) I planned the analysis to compare "structured" and "traditional" mRS. A one group descriptive study using average absolute difference with a fixed effects model was performed using MIX software version 1.7 (www.mix-for-meta-anlaysis.info). [last accessed January 2010]

Results

The review profile is detailed in Figure 2. From 3461 original titles, 312 abstracts were eligible for review, 31 (20;62-64;70;72;85;87;93-114) studies were initially considered for inclusion and 10 studies involving 587 patients were included in the final analysis. (20;62-64;87;94;97-99;114) (Table 1)

Two reports required translation (*German and Portuguese*)(87;114). For one report(114) the authors provided additional data not available in the published manuscript and these have been included in the final analysis. Other authors did not reply or were unable to provide additional information and for this reason certain data are missing from results tables.

Of the reports considered, reasons for exclusion included but were not limited to: use of mRS in a non-stroke population (n=1)(102); use of outcomes other than mRS (n=5)(93;100;103;106;113;115) and use of dichotomised (favourable / non-favourable) mRS outcome with no corresponding non-dichotomised data (n=2).(70;105) In 5 reports, mRS data were reported with no patient interview, for example mRS derived from patient case-record or from a video based training exercise.(85;95;96;101;104) A full description of all complete manuscripts considered and reasons for exclusion is provided in the appendix.

For the purposes of presenting a comprehensive analysis of mRS reliability, I have included my own mRS studies in the systematic review. A full description of the methodology and results of these studies will be presented in chapter four. To avoid confusion, in the thesis my departmental study of mRS reliability (chapter four) will be referred to as the "study" and the data presented in this chapter will be referred to as the "meta-anlaysis". For reference the corresponding figures for mRS reliability if the departmental study is not included would be:

Inter-observer reliability mRS (traditional approach) *k*=0.43 (0.39 – 0.50).

Inter-observer reliability mRS (structured approach): *k*=0.65 (0.58 - 0.73).

Intra-observer reliability mRS: *k*=0.91 (0.83 - 0.99).

Inter-observer variability of mRS described in the included studies varied from "near perfect" (*kw*=0.95) to "poor" (*k*=0.25). (Table 2a/2b)

In the included studies, multiple methodologies were used to administer mRS and study its properties. (Table 1) Previous reports have suggested that factors such as: timing of assessment (acute or post discharge from hospital)(116;117); background and training of observers(118); native language(96); use of a proxy(119) and use of structured interview(62) can impact on stroke outcome scales or specifically influence reliability of mRS. These data were extracted for the included studies if available.(Table 1) Three studies purported to measure intra-observer variability of mRS.(63;94;97) (Table 3)

Figure 2: Review profile for mRS reliability literature search.

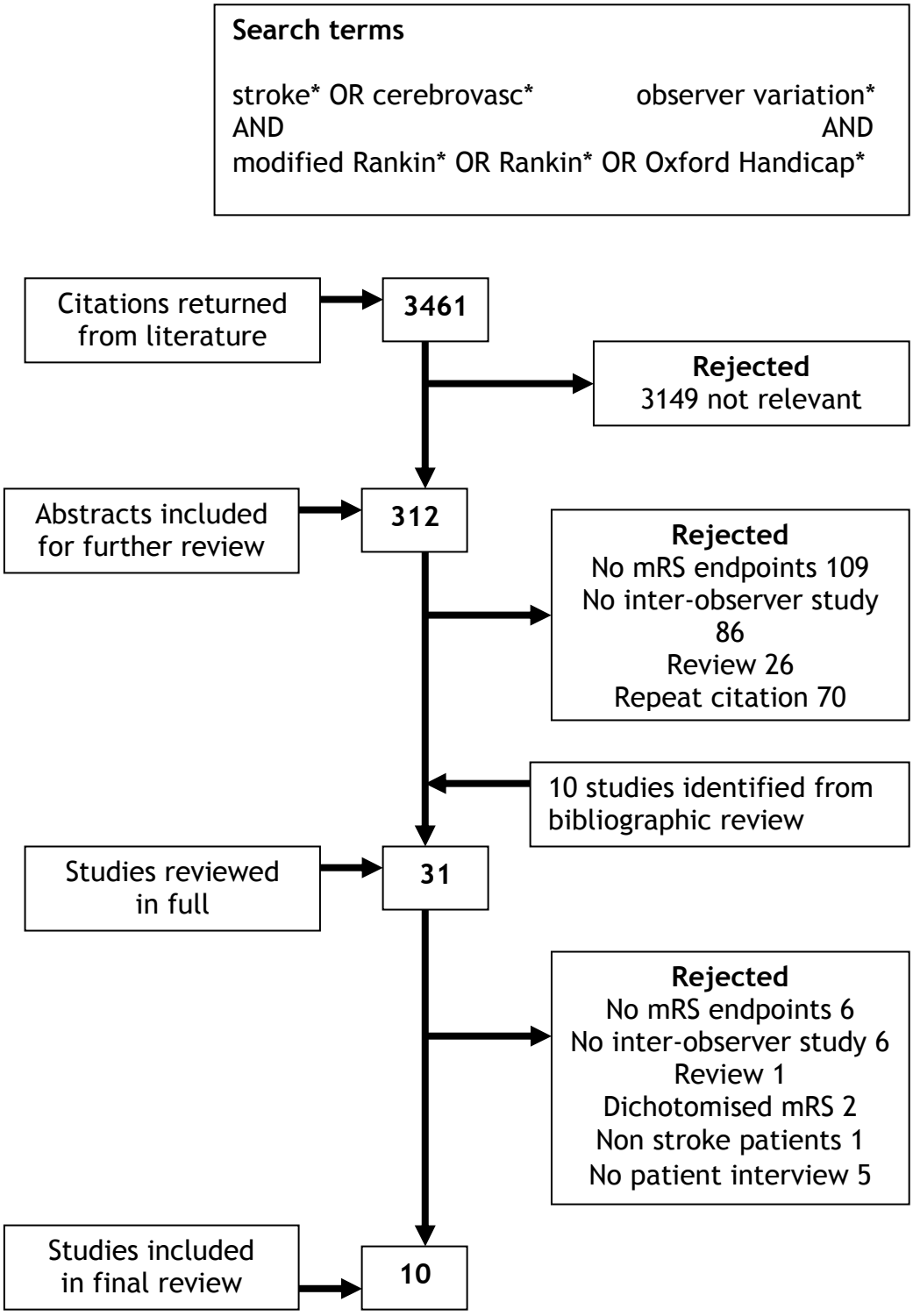


Table 1: Studies of mRS reliability, with reference to study methodology.

*"**Medical assessors**" are stroke physicians or other clinicians;*

*"**S.I**" Structured Interview*

*"**Timing**": "Acute" any assessment performed within first seven days and while still hospital inpatient; "Chronic" represents all other assessments*

N/A information not available

Study	Cases (n)	Assessor (n)	Medical assessor	mRS Language	Methodology	Proxy	S.I	Single site	Timing
van Swieten	100	35	Yes	English	Face to face	Yes	No	Yes	Acute
Wolfe	36	3	No	English	Face to face	Yes	No	Yes	Chronic
Berger	43	2	Yes	Non-English	Face to face	N/A	No	Yes	Chronic
Wilson 2002	63	2	Yes	English	Face to face	No	Yes	Yes	Chronic
Newcommon	34	4	No	English	Telephone	No	Yes	Yes	Chronic
Wilson 2005	117	15	No	English	Face to face	N/A	Yes	No	Chronic
Gur	18	2	Yes	N/A	Face to face	No	No	Yes	Acute
de Canada	51	2	Yes	Non-English	Face to face	No	No	Yes	Acute
Meyer	25	2	Yes	English	Telemedicine	No	No	Yes	Acute
Quinn	99	7	No	English	Face to face	Yes	Yes	Yes	Chronic

Table 2a: Reliability of traditional mRS as measured by kappa statistics and percentage agreement between observers.

Kappa statistics are presented as standard kappa (k) and with quadratic weighting (kw). Corresponding 95% confidence intervals are also given.

"N/A" – data not available

Study	Kappa (*k*)	Weighted (*k*w)	Agreement (%)
van Swieten 1988	0.56 (0.45 - 0.68)	0.91 (0.71 - 1.00)	65%
Wolfe 1991	N/A	0.87 (0.84 - 0.97)	80%
Berger 1999	0.56 (0.41 - 0.71)	0.88 (0.58 - 1.00)	N/A
Wilson 2002	0.44 (0.29 - 0.62)	0.78 (0.53 – 1.00)	57%
Newcommon 2003	0.72 (0.55 - 0.89)	N/A	N/A
Wilson 2005	0.25 (0.16 - 0.35)	0.71 (0.53 - 0.88)	43%
Gur 2006	N/A	0.95 (0.89 - 1.00)	N/A
de Canada 2006	0.45 (0.31 - 0.60)	0.70 (0.58 - 0.82)	58%
Meyer 2008	N/A	0.90 (0.59 - 1.00)	N/A
Quinn 2009	0.64 (0.48 - 0.79)	0.91 (0.65 - 1.00)	72%
TOTALS	0.46 (0.41 – 0.51)	0.91 (0.86 – 0.93)	60%

Table 2b: Reliability of mRS using a structured interview approach as measured by kappa statistics and percentage agreement between observers.

Kappa statistics are presented as standard kappa (k) and with quadratic weighting (kw). Corresponding 95% confidence intervals are also given.

"N/A" – data not available

Study	Kappa (*k*)	Weighted (*k*w)	Agreement (%)
Wilson 2002	0.70 (0.56 - 0.85)	0.93 (0.67 - 1.00)	78%
Newcommon 2003	0.34 (0.17 - 0.55)	N/A	50%
Wilson 2005	0.74 (0.64 - 0.84)	0.91 (0.73 - 1.00)	81%
Quinn 2009	0.50 (0.34 - 0.68)	0.74 (0.455 - 1.00)	63%
TOTALS	0.62 (0.56 – 0.69)	0.87 (0.75 – 1.00)	73%

Table 3: Studies of intra-observer variability in mRS as measured by kappa statistics and percentage agreement between observers.

Kappa statistics are presented as standard kappa (k) and with quadratic weighting (kw). Corresponding 95% confidence intervals are also given.

"N/A" – data not available

† = average across structured and standard interview approaches.

Study	Kappa (*k*)	Weighted (*kw*)	Agreement (%)
Wolfe 1991	N/A	0.95 (0.88 - 1.00)	86%
Wilson 2005	0.83 (0.66 - 1.00)	0.96 (0.68 - 1.00)	91%
Quinn 2009	0.72 (0.61 - 0.82)	0.93 (0.81 - 1.00)	77%†
TOTALS		0.94 (0.88 -1.00)	84%

Variability in mRS was described using *k* statistics (n=7)(20;62-64;87;94;114) *k*w (n=8)(20;62;63;87;97-99;114); intra-class correlation coefficient (ICC) (n=2)(87;114) and percentage agreement (%). (n=5).(20;62;63;94;97) For all included studies there were insufficient data presented in the original reports to allow "back" derivation of other measures of reliability.

Inter-observer variability of mRS varied from "near perfect" (*kw*=0.95) to "poor" (*k*=0.25) in the original descriptions. (Table 2a/2b) Use of the structured interview was not consistently associated with improved reliability, with weighted kappa similar for the two approaches (structured mRS:*kw*=0.87; traditional mRS:*kw*=0.90). (Table 2a/b)

In the included studies, diverse methodologies were used to administer mRS and study its properties. (Table 1) No study met the "minimum" criteria to allow adequate assessment of quality: no description of patient selection (n=5); no data on blinding between assessors (n=5); inadequate description of mRS methodology (n=2); no description of location / timing of mRS (n=2). As a result I decided to include all relevant studies regardless of poor methodological quality or potential bias.

Three studies purported to measure intra-observer variability of mRS(62;94;97) with 162 patients included. Overall intra-observer reliability was very good *kw*=0.94. (Table 3)

Discussion

The potential for variability to impact on the utility of mRS as a stroke outcome scale has been appreciated since its inception.(20) Previous studies of mRS variability have described differing results.(63;97) This review suggests that overall reliability of standard mRS is moderate but there remains potential for improvement. There was some suggestion in the combined data that use of a structured interview may improve reliability, however the apparent benefits were lost when "weighted" kappas were applied. The non-parametric nature of kappa and its derivatives does not allow for comparative meta-analysis and so the safest conclusion is that structuring mRS may partly improve mRS reliability but effects have not been consistent across studies.

It is interesting that those studies with larger numbers of patients and observers reported poorer reliability. The importance of maintaining inter-observer reliability in a contemporary clinical trial becomes readily apparent when the number of potential endpoint assessors is considered. For a modest sized phase III clinical trial, several hundred patients may have to be enrolled. Outcome assessment for such a trial will require a number of assessors at numerous sites. In the recent SAINT trials of the putative neuro-protectant NXY-059, over one thousand assessors from twenty-five countries were trained in outcome assessment for the study.(120) In comparison numbers included in reliability studies are small, in this meta-analysis median number of patients was less than fifty with median two observers from a single site. If we are to better understand the impact of mRS reliability on contemporary clinical trials, the ideal study methodology would involve a series of trained observers of differing

backgrounds and from differing international centres, assessing mRS on pre-selected patients at a fixed time-point following discharge. Only one study in my meta-analysis approaches this "ideal" and it reports a concerningly low reliability for standard mRS.(63)

Factors internal and external to the mRS interview may impact upon reliability. As well as structuring the interview, previous reports have proven or suggested that: timing of assessment (acute or post discharge from hospital)(121); background and training of observers(118;122); native language of assessor and patient(123) and use of a proxy(73;119;124) can impact on reliability. There was considerable heterogeneity between studies in all these areas and this may have impacted upon results. As example, only my own departmental study made use of a recognised mRS training resource. This study reported no beneficial effect of structured interview, suggesting that the structured approach may be unnecessary if assessors are adequately trained. Collation of studies with fundamental differences in methodology potentially weakens the meta-analysis, but is perhaps necessary for mRS work. As will be described in the next chapter, clinical trials report substantial heterogeneity in mRS assessment with no consensus as to optimal approach.

The quality of included studies also varied. For all studies certain data were incomplete for important contributors to trial quality such as blinding and patient selection. This lack of detail is unfortunate but perhaps not surprising. Standardised criteria such as the CONSORT (http://www.consort-statement.org/) [last accessed January 2010] statement have improved

reporting of clinical trials. No equivalent, universally accepted criteria for reporting of reliability / clinimetric studies are currently available.

Heterogeneity was further evident in the statistical methods employed in reliability studies. Several techniques have been used to describe reliability.(36) Four methods of describing reliability were used: kappa; weighted kappa; intra-class correlation coefficient and percentage agreement. Although all methods are appropriate, the resultant data are not readily interchangeable. The ideal would have been to access individual patient data from each of the trialists, however these data were not available. It is interesting that more authors chose to present data as "weighted" kappa. For a rating scale where differences between observers are unlikely to be of more than 2 grades, a quadratic weighting system can "inflate" the final kappa and this was demonstrated in my data. My own use of statistics demands some discussion. No universally accepted method for analysis of multiple kappa statistics from differing populations has been described. Recognising this limitation I used a group analysis technique that made the fewest assumptions of the underlying data.

Accepting these limitations my meta-analysis does have certain strengths. My literature searching strategy was as comprehensive and systematic as possible. The spread of reliability estimates obtained suggest no overt publication bias, a formal analysis of publication bias such as funnel-plot was not performed. I considered a number of reports from non-English and "non-medical" sources and certain of these were included in the final analysis. I was inclusive in my approach to reports, although excluded studies that would not help describe the variability of mRS in a clinical trial setting. Studies using the original Rankin or OHS were excluded as these scales are no longer used in contemporary stroke

research (for reference in comparable inter-observer reliability studies RS gave *kw* 0.79(113); OHS *kw* 0.72(93)). Similarly, I excluded those studies where mRS was derived with no patient contact (from case notes; postal questionnaire).

I also present studies of intra-observer variability of mRS, suggesting excellent reliability. Here again problems in study methodology limit the strength of conclusions that can be drawn. Two studies measured mRS at distinct periods in the patient's recovery and as such are prone to recall bias and the potential for functional ability to change between assessments.(63;97) My own study used a novel based approach that will be described in detail in chapter four. Although theoretically interesting, intra-observer variability of mRS assessment may be of less relevance to clinical trials, where primary outcome assessment is usually performed once only.

There remains uncertainty regarding the reliability of mRS as an outcome measure. Available reliability studies are likely underpowered and have design flaws that limit their generalisation. Those studies closest in their design to large scale contemporary clinical trials demonstrate potentially significant inter-observer reliability. These data suggest that researchers should conduct further studies using methodologies that "mimic" large scale clinical trials. While we await definitive data on mRS reliability we must acknowledge that a degree of inter-observer variability is inherent in standard mRS grading and further work on methods to reduce variability will be equally important.

Chapter three

Functional outcome measures in contemporary stroke trials – a systematic review

Introduction

Accurate and meaningful assessment of patient outcomes is essential for observational studies and interventional trials.(125) As stroke represents the leading cause of adult disability(56), an important consideration for any stroke trial is valid quantification of functional outcomes. Some discussion of the clinimetric properties of functional outcome scales and summary of the debate regarding the relative strengths and limitations of diverse assessment instruments has already been presented. At present there is no consensus on optimal outcome measure(s) for use as clinical trial endpoint.

Post stroke recovery can be described in a number of domains. Use of the WHO international classification of functioning, disability and health (15) (http://www.who.int/classifications/icf/site/icftemplate.cfm.) [last accessed January 2010] as a framework for describing post stroke recovery states was discussed in chapter one. Multiple assessment scales exist to describe impairment; activity (formerly disability) and societal participation (formerly handicap). For stroke trialists, stroke specific; bespoke and generic outcome assessment scales exist.(107) Thus, there is potential for heterogeneity both in the domain measured and within that domain.

In addition to heterogeneity in choice of outcome measure, there is further potential for heterogeneity in the methodology used to collect and describe functional data. For many of the popular outcome measures, diverse approaches to data collection and statistical analysis have been employed, with little formal guidance on best practice. Methods to improve reliability and

validity of outcome assessment, such as observer training, are increasingly available.(46) However, the frequency of their use in clinical trials is not well described.

Some authors have suggested that heterogeneity in outcomes assessment is a particular problem in the field of stroke rehabilitation literature, although at the time of writing this has never been quantified.(126;127) In fact, there are limited data on the extent of outcomes heterogeneity in all areas of stroke research.

Previous review of functional outcome assessment in acute stroke trials (1995 – 1998) reported that the BI was the most frequently used end-point.(57) Given the last decade's exponential increase in stroke related research, an updated review of outcome measures in the stroke literature was required.

I sought to describe the frequency of use, and methodology of application employed for functional outcome measurement in contemporary stroke literature.

Methods

I performed a literature review of stroke trials contained in a selection of high profile international journals, targeted at general medical; neurology and stroke specific readerships. Choice of publication was based on impact factor; target audience and frequency of publication of stroke related literature. The aim was to describe outcomes particularly relevant to acute stroke trials and I did not include rehabilitation specific rehabilitation journals. A similar analysis looking at the rehabilitation literature has recently been completed.(128)

Following informal review of a number of titles I chose to restrict analysis to the following publications: "Stroke" (*Lippincott, Williams and Wilkins for the American Heart Association*); "Neurology" (*Lippincott, Williams and Wilkins for the American Academy of Neurology*); "Lancet" and "Lancet Neurology" (*Elsevier*); "New England Journal Medicine" (*NEJM - Massachusetts Medical Society*) and "Journal of the American Medical Association" (*JAMA - American Medical Association*). The chosen publications were hand searched and titles were screened. Abstracts of potentially relevant papers were independently reviewed. To ensure no potential manuscripts were missed I reviewed all journal content, including letters and short reports. In addition, an independent Medline search was performed across each title using the key terms "stroke" and "cerebrovascular accident" and limited to "human studies"; years "2001 – 2006".

A “functional outcome” was defined as a quantified measure across any of the domains of impairment, activity or participation. For this analysis, within the domain of participation I included those scales that purport to measure quality of life or related outcomes. The functional measure did not have to be the primary endpoint of the trial. A “stroke trial” was defined as any active intervention in stroke patients (ischaemic or haemorrhagic) or primary prevention. I made no assessment of the quality of the trial’s aims, design or conclusions and included any manuscript that purported to assess an active intervention. I limited the literature search to articles published in the period Jan 2001 to December 2006 inclusive, including only those trials that involved human subjects.

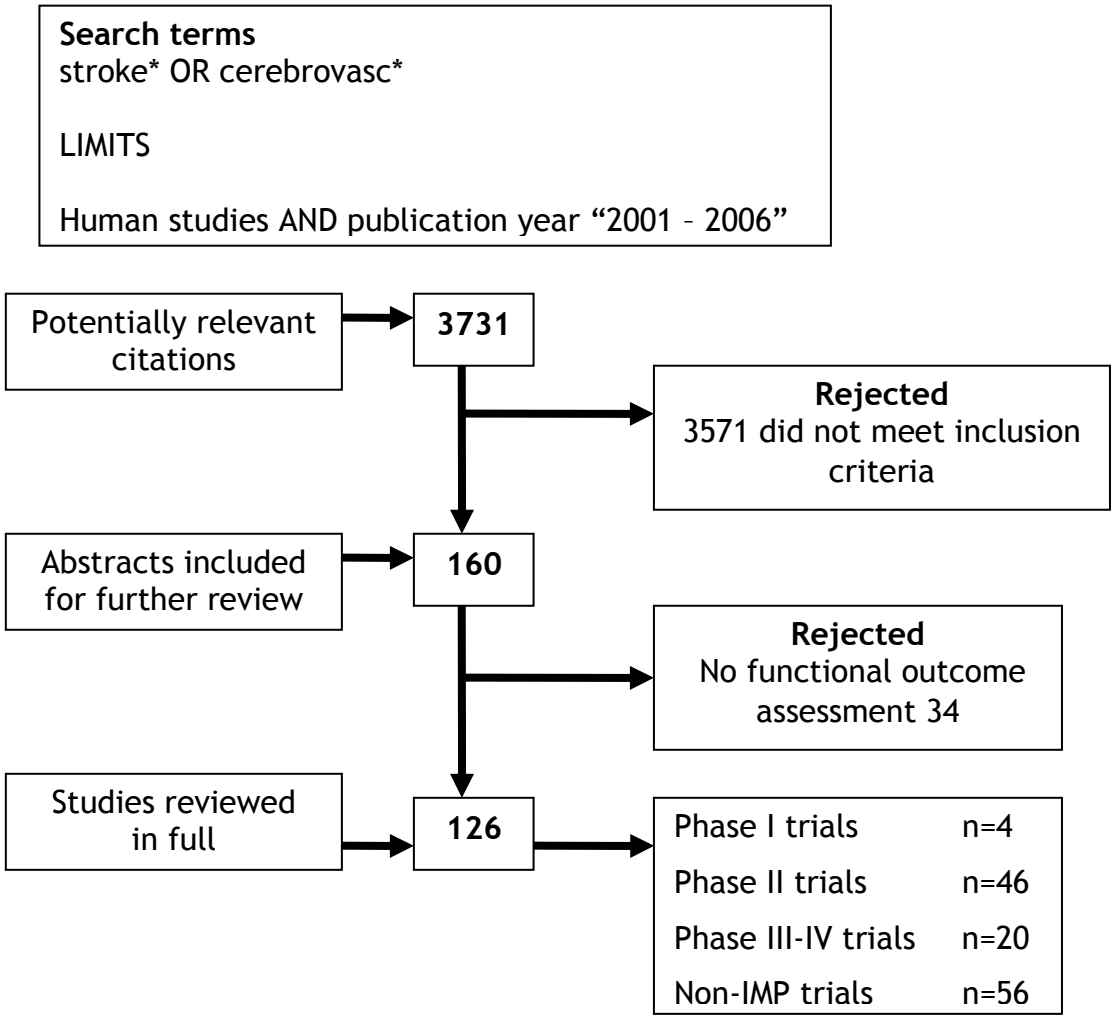
Two independent researchers (TJ Quinn, J Dawson) reviewed all articles potentially meeting inclusion criteria. Final choice of included manuscripts was by consensus. Separate papers using the same trial dataset were only included if the functional outcomes described differed. Complete manuscripts were reviewed and relevant data extracted on to a standard form. Where additional methodology was described in on-line or paper supplement this was also accessed. As the purpose of this review was to document outcome assessment tools as described in the published literature, no attempt was made to contact authors of manuscripts where description of methodology was unclear. (Figure 3)

The following data were collated: intervention; year of publication; journal; size of study population; primary outcome measure; functional outcome(s) assessed; functional domain(s) assessed and for each functional assessment used I recorded: timing of assessment and method of assessment including details of any training offered. Where the nature of an assessment instrument was not clear I sought the original description of the scale or referred to reference works.(129)

I performed simple statistical analyses to compare functional outcomes assessment in general medical and stroke/neurology specific journals; comparing trials utilising an investigational medicinal product (IMP) and non-IMP trials. Proportions were compared using chi-square testing. All analyses were performed using Minitab software (version 14.0, Minitab Inc, PA, USA).

Figure 3: Review profile for functional outcomes literature search.

Search included only pre-specified journals:
Journal of the American Medical Association
Lancet, Lancet Neurology
Neurology, New England Journal of Medicine,
Stroke



Results

From a total of 3731 screened articles, 160 were reviewed in full and 126 were considered suitable for inclusion in the analysis. Included studies were a mix of pre-clinical, clinical and non-IMP interventions (phase I 4; phase II 46; phase III/IV 20; and non-IMP studies 56). Study interventions included (but were not limited to) anti-thrombotic drugs 12; thrombolytic treatment 32; putative neuro-protectants 19; rehabilitation strategies 29; stenting or mechanical intervention 8 and stroke unit care 4. There were three primary prevention studies with stroke related functional outcomes. Median study size was 100 patients (range 9 to 7121; IQR 367). The numbers of trials from the chosen titles were: JAMA 5 trials; Lancet 12; Lancet Neurology 3; Neurology 33; NEJM 6; Stroke 67.

Forty-seven outcome measures were described in the included studies (full list available in appendix D). In 100 studies, an assessment of functional outcome was used as the trial's primary endpoint. The median number of functional outcomes recorded per trial was 2 (range 1-9; IQR 2). The most frequently used outcome measure was mRS, followed by BI and NIHSS (Table 4). A composite or global scale that incorporated a number of outcome measures was used in 10 papers; in 3 papers a bespoke scale created by the authors was employed. Seventy-seven studies purported to measure impairment; 103 activity and 11 participation. Six trials described recovery across the three domains. The most frequently used impairment scale was the NIHSS; the most frequently used activity scale was the mRS; the most frequent participation measure was SIS.

Table 4: Frequency of use of functional outcomes assessment scales in contemporary stroke trials.

"1° endpoint" is number of papers where outcome measure is used as the studies primary endpoint, does not include studies where measure is part of a combined "global" endpoint.

Outcome Measure	Number of Trials Instrument Used	1° Endpoint
Modified Rankin Scale	81 (64.3%)	33 (26.2%)
Barthel Index	51 (40.5%)	10 (7.9%)
Nat. Institutes of Health Stroke Scale	35 (27.8%)	15 (11.9%)
Scandinavian Stroke Scale	11 (8.7%)	2 (1.6%)
Glasgow Outcomes Scale	8 (6.3%)	2 (1.6%)
Frenchay Activities Index	6 (4.7%)	1 (0.8%)
Timed Walk/ 6 Minute Walk	6 (4.7%)	3 (2.4%)
EuroQOL	4 (3.1%)	0
Fugl-Meyer Motor	4 (3.1%)	4 (3.1%)
Wolf Motor Functional Test	4 (3.1%)	1 (0.8%)
Rivermead Mobility Index	3 (2.4%)	1 (0.8%)
Short Form 36	3 (2.4%)	1 (0.8%)
Stroke Impact Scale	3 (2.4%)	0
Berg Balance Scale	2 (1.6%)	1 (0.8%)
Canadian Stroke Scale	2 (1.6%)	1 (0.8%)
Tinetti Balance Assessment Tool	2 (1.6%)	1 (0.8%)

Data on timing of outcome assessment were available for 113 trials. Median time from ictus to assessment was 90 days (range 2 days – 5 years; IQR 150). Only 4 trials described use of training in outcome assessment. Thirty-four papers described the methodology used to collect functional data (Table 5). Published descriptions of functional outcome measurement methodology were infrequent and did not increase from 2001 – 2006 (*P*=0.889).

Comparing IMP and non-IMP studies, a greater total number of outcome measures were used in the latter (non-IMP 47 measures across 56 trials; IMP 15 measures across 70 trials). A greater number of functional measures were described per trial for non-IMP studies (median=3 versus median=2 *P*=0.029). There were no significant differences in number of papers describing recovery using impairment; activity or participation scales (IMP trials: impairment 41; activity 63; participation 3. Non-IMP trials: impairment 36; activity 40; participation 3 *P*=0.529). There were no differences in number of papers describing trial methodology (non-IMP 33.9%; IMP 32.9% *P*=0.256).

For the most popular outcome measure (mRS) I collated information on data handling and statistical analysis. The majority of studies (n=55) collected mRS data at 90 days post event (range 5 days to two years). In 46 trials outcomes were dichotomised (mRS 0-1 – 16; mRS 0-2 – 18; mRS 0-3 – 9; mRS 0-4 – 3); 11 trials used trichotomised outcomes. MRS was the most popular outcome measure for each year studied, frequency of use did not significantly change across 2001 – 2006 (Table 6, *P*=0.426). Papers presented in general medical journals were more likely to use mRS as clinical endpoint (Medical journals 18/23; neurology journals 61/103 *P*=0.035).

Table 5: Methodologies for assessment of stroke functional outcomes.

Method for collecting functional outcomes data	Number of trials
Case-sheet review	4 (3.1%)
Face to face interview	13 (10.2%)
Postal survey	2 (1.6%)
Questionnaire	3 (2.4%)
Structured interview	2 (1.6%)
Telephone interview	17 (13.4%)
No description of method	93 (73.2%)

Table 6: Numbers of published stroke trials

a) Providing comprehensive description of outcome assessment methodology employed.

b) Using modified Rankin Scale as outcome measure.

Year	(a) Methodology described	(b) mRS used as outcome measure
2001 n=18	5 (27.8%)	10 (55.6%)
2002 n=14	6 (42.9%)	6 (42.9%)
2003 n=16	5 (31.3%)	10 (62.5%)
2004 n=15	4 (26.7%)	11 (73.3%)
2005 n=32	11 (34.4%)	23 (72.0%)
2006 n=31	8 (25.9%)	21 (68.0%)

Discussion

There is an increasing variety of assessment scales available to describe post stroke outcomes. I found significant heterogeneity in functional outcome assessments across a number of high-impact medical journals. This heterogeneity was evident in choice of outcome, method of application and analysis of data. For many studies, description of methodologies used to collect outcomes data was incomplete or absent.

My analysis confirms that mRS is now the most frequently used functional outcome measure in stroke trials. Where details of methodology were given, the majority of papers administered mRS at 90 days post event, used telephone-based assessment and analysed data using dichotomisation.

The University of Washington stroke centre describes 20 different outcome assessment scales in common use in the stroke trials. (http://www.strokecenter.org/trials/scales/scales-overview.htm.) [last accessed January 2010] My review of the recent stroke literature found 47 outcome assessment instruments in current use. These included a number of tools that are poorly validated or are recognised to have clinimetric weaknesses. Despite the variety of valid tools available, some authors continue to use their own bespoke assessment scales. This substantial heterogeneity in trial end-points makes meaningful comparisons between trials challenging and can preclude formal meta-analysis.

The majority of trials describe functional outcomes in terms of activity. This is not surprising, as efficacious acute stroke treatments will show the greatest change on a suitably responsive impairment measure, this allows for adequate powering with smaller numbers.(39) Surprisingly few studies attempted to describe participation, although arguably this is the most meaningful measure for the patient. This may represent the relative lack of established, robust instruments to quantify this domain. It may also reflect a fear that beneficial treatment effects may be swamped by variation in opportunities for participation. Domains of recovery are not interchangeable. In fact, relationships between impairment, activity and participation measures are poorly understood.(130) An argument can be made for describing recovery in more than one domain. A small number of trials have taken this approach. Six trials described outcomes across all three domains, either separately or combined into a single global outcome measure.

For individual outcome measures there was heterogeneity in the methodology used to capture data. This is perhaps unsurprising as for many of the outcome measures used there is little formal guidance on how to administer the tool. I collated data on timing of assessment(131); use of training(85); background of assessor(108;132) and use of a standardised structured interview(62) - as each of these factors is known, or suspected, to impact on validity of outcome data. Comprehensive descriptions of methodology were infrequent, no trial was described in terms of all of the listed factors. Where an attempt was made to describe methodology there was again marked heterogeneity.

It could be argued that full description of methodology is unnecessary as for many scales there is an agreed, albeit informal, approach to data collection among researchers. While this may be true for certain, simple impairment scales, it is not true for the most popular tools. Taking mRS as example, the scale has variously been applied using direct interview(20), telephone interview(64); video recorded interview(85); structured questionnaire(62) or through case-sheet derivation.(133) In the majority of papers using mRS, no description of methodology was apparent. It is likely that most of these studies used "traditional" face to face interview, however given the variety of assessment techniques available it is unacceptable to omit such details from manuscripts. The optimal method for performing mRS grading is still debated, however it is recognised that mRS methodology can impact on the validity and reliability of the data collected.(31) Thus, lack of clarity on grading techniques precludes critical assessment of trial quality.

The lack of information on use of investigator training is concerning. It is well recognised that training in use of an outcomes assessment tool improves reliability.(46) A validated training program has been available for NIHSS scoring for many years(115); a specific mRS training package will be described in a later chapter. It is disappointing then that only four trials specifically described training of assessors. It is possible that training was utilised but not described in the methods section of the published report. Word limits utilised in paper journals may force authors to remove sections of methodology they consider extraneous. However, omission of a technique likely to impact on validity of data is clearly not acceptable and in the absence of any formal description the reader must assume training was not used. The need for explicit description of

training methods used could be incorporated into future CONSORT guidelines for presentation of clinical trials.

The difficulty in defining clinically meaningful outcomes in trials of rehabilitation is well described.(134;135) I found that non-IMP studies made use of a greater range of outcome instruments, and used more functional outcomes per trial. However, there was no difference in quality of outcomes reporting. As rehabilitation is traditionally considered a more "holistic" specialty one would expect increased use of measures of participation. My data do not support this assumption. In interpreting these data, I recognise that no "power" calculation were performed and the modest numbers included may mask a true difference, I recognise also that the journals chosen for review did not include any rehabilitation specific titles.

Previous authors have highlighted the lack of uniformity in numerical data handling procedures published in acute stroke trials.(136) My review confirms heterogeneity in statistical analysis of functional outcomes data. Taking mRS as example, the majority of papers dichotomised mRS to describe outcome - that is they transformed the ordinal categories into a binary outcome of good "or" poor outcome. A range of mRS grades were used as cut-off point for defining good outcome, from mRS 1 to mRS 4.(137) The rationale for using differing definitions of favourable outcomes is clear: expected "good" outcome from a condition such as malignant MCA infarct(138) will not correspond to expected "good" outcome from a transient ischaemic attack (TIA).(139) However, dichotomising with no consistency between trials complicates comparison and meta-analysis. More sophisticated methods of data handling have been proposed

including use of global statistics, responder analysis and shift analysis and future trials will likely make more use of these techniques.(78;140)

There are limitations in the data I have collected. The large and increasing stroke literature precluded a comprehensive review of stroke outcomes across all published stroke studies. I note that previous groups' attempts to characterise acute stroke trials comprehensively have been unsuccessful(141) due in part to the large numbers of published and unpublished studies in the field.(142) Therefore I chose to limit my analysis to journals with a large international readership across the disciplines of neurology; stroke medicine and general internal medicine. Thus my data are open to publication bias; however, the intention was to describe outcome assessment use in popular medical journals rather than across the complete stroke literature. Although not comprehensive my search strategy was systematic. I defined search terms of "functional outcome" and "stroke trial" in a manner that was robust but inclusive. Such an approach is not without precedent and accepting these limits, my literature review still provided a mix of large-scale multi-centre trials; smaller hypothesis generating exercises and non-IMP interventions. My intention was purely to describe current outcomes assessment methodology; I have not attempted to compare strengths and weaknesses of different instruments, discussion of this topic is presented in chapter one and contemporary published reviews are available.(17;86;143)

Given the heterogeneity in outcome assessment scales that I have described, it could be argued that cross disciplinary stroke researchers should agree on a "core" set of outcome tools. In the research setting a "common language" of

instruments used would facilitate better use of the clinical information derived from these tools. In a non-research clinical setting the use of a limited number of scales with which all members of the team are familiar would enhance communication between different professions and disciplines.

The British Society of Rehabilitation medicine has recognised the benefits of such an approach and has developed a "basket of measures" – a series of expert selected assessment aids that they would recommend for routine use. (http://www.bsrm.co.uk/ClinicalGuidance/OutcomeMeasuresB3.pdf) [last accessed January 2010]. These measures were not stroke specific and may not applicable to acute stroke trials.

An attempt to standardise stroke trial outcomes and in doing so promote collaboration can be seen in the recent Department of Health supported COSTAR project (Collaborative Stroke Audit and Research). COSTAR are working towards agreed methodological standards including recommended approaches to consent, randomisation, blinding and assessment. (http://www.dh.gov.uk/en/Researchanddevelopment/A-Z/Cardiovasculardiseaseandstroke/DH_4002086) [last accessed January 2010]. Recognising that complex interventions require multi-centre collaboration, pre-planned collaboration between autonomous studies has been suggested, so called "epi-analysis". Such an approach requires standardisation of outcomes.

As there is little specific guidance on outcome assessment, the lack of uniformity I have described is perhaps unsurprising. For future trials, optimal outcome assessment should be based on established, clinimetric sound procedures, ideally with training available for assessors. Regardless of the outcome scale chosen, trialists are urged to consider and describe fully the methodologies employed in assessment of stroke recovery.

Chapter four

Exploring the reliability of the modified Rankin Scale

Introduction

As demonstrated in the previous chapter, mRS is the most prevalent functional outcome measure in contemporary stroke literature. Prior to any clinical or trial use, assessment scales should be proven "fit for purpose". Clinimetrics is the methodological discipline that describes clinical measurement quality. Outcome scales are traditionally assessed in terms of responsiveness, validity and reliability.(26;144) The original Rankin scale was not designed for clinical trial use and like many other stroke assessment tools, mRS became established as a study endpoint prior to any formal clinimetric assessment.(92) Recent studies have quantified responsiveness of mRS(52) and proven excellent construct and convergent validity of the scale.(31;145)

In the systematic review presented in chapter two, I have shown that a potential problem inherent in mRS grading may be inter-observer variability. Poor reliability is a concern for trialists, as arguably for an instrument that will be used by many hundreds of raters in large-scale multi-centre clinical trials, reliability is the most important property of the scale. My previous conclusion of heterogeneity in the degree of mRS variability, may be partly explained by differing study methodologies. As trialists are principally concerned with the variability present in clinical studies, the most informative analysis of mRS would be conducted using current researchers working in a clinical trial setting and interviewing real stroke survivors. Those few studies that have attempted this design, used only limited numbers of assessors and / or patients.(63)

Thus, there are several unanswered questions regarding reliability of mRS assessment. Using a mock clinical trial design I set out to study the inter and intra-observer variability when mRS is applied by experienced and trained study personnel. I further describe the effect of a structured interview format on properties of the mRS. Finally, recognising that initial clinical judgement often influences final scoring in assessment scales, I described the ability of researchers to estimate disability from limited review, prior to formal mRS assessment.

Methods

Patients and assessors

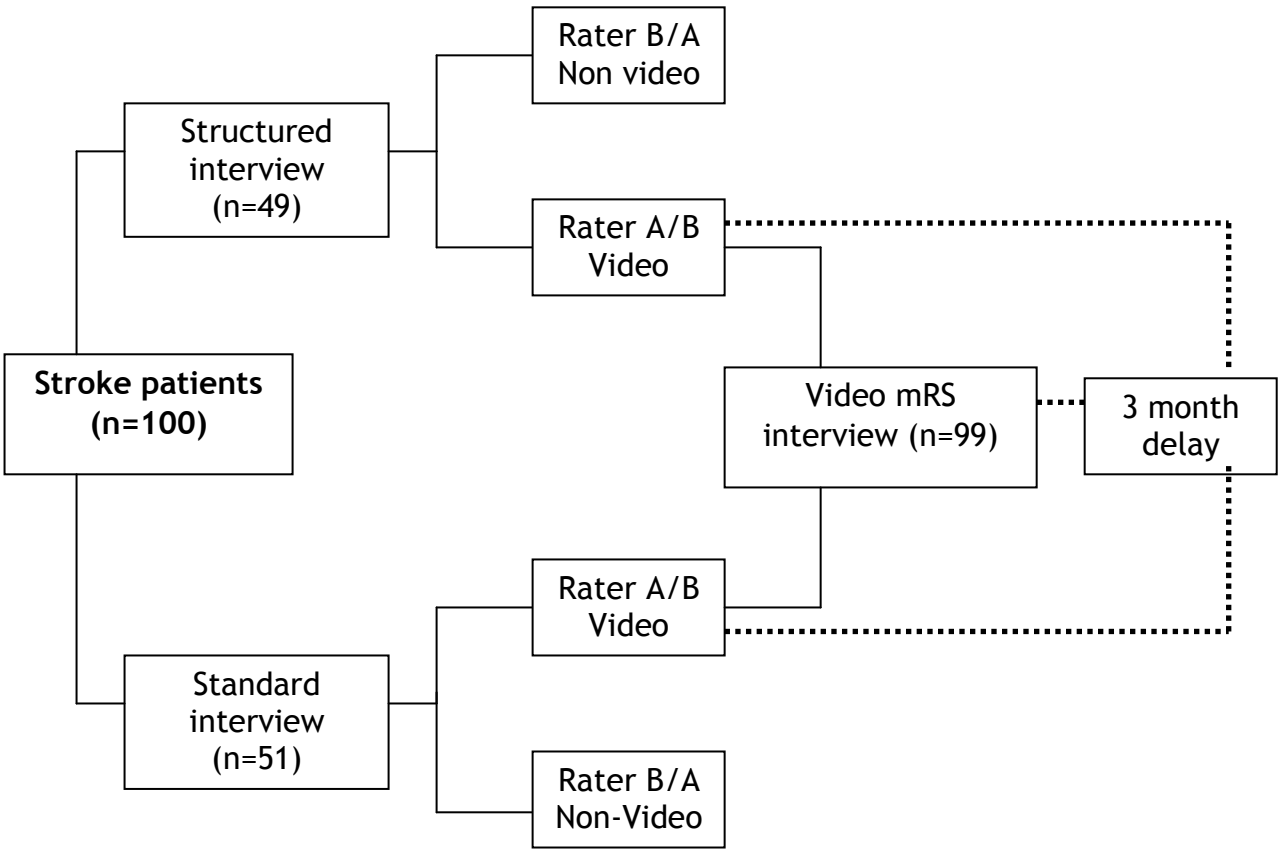
I approached sequential patients attending the Western Infirmary Glasgow, University Hospital cerebrovascular clinic for their routine post stroke assessment. Clinic patients have usually been inpatient in the local acute stroke unit and typically attend for review at 90 days post stroke; however I did not set fixed time related or geographic exclusion criteria. All patients were considered for inclusion. If cognitive impairment or language problems precluded satisfactory mRS interview, a proxy (family member or carer) was used. Informed consent was given by all participants or designated proxy prior to recruitment and reconfirmed following the assessment. The local ethics committee approved the study protocol.

To allow assessment of variability across a representative group I involved seven assessors: 4 stroke physicians and 3 research nurses. All assessors had been trained and certified in mRS assessment using the bespoke training package (as outlined in chapter five) and all have considerable experience of mRS application in clinical trials.

I used a stratified assessment technique to test the related hypotheses (Figure 4). The selection of mRS assessors; interview methodology used and selection for interview recording were all pre-specified using an online randomisation program (www.random.org/integers) [last accessed January 2010] and allocation was concealed from interviewers and patients using an opaque envelope system.

Figure 4: Schematic of evaluation process for mRS reliability assessment.

A, B represent trained mRS assessors, selected from a pool of 7 research nurses / stroke physicians. Order of video interview (first or second in pair) and designation to A or B were determined by randomisation. At three months, video recorded assessments were reviewed by the original interviewer and graded again.



Statistical analysis

Reliability was described with kappa (*k*) statistics - kappa statistics were chosen for primary analysis as clinicians are familiar with the test and as previous published studies of mRS reliability have used similar statistical techniques.

Formal comparisons between kappa statistics are problematic, particularly if numbers in each group are not comparable.(34) Therefore to allow for basic comparative analysis I also calculated the number of interviews where rater pairs agreed exactly on mRS, (expressed as percentage agreement) and compared values using chi-square testing. Specific analyses performed for each hypothesis will be detailed in the relevant subsections. All statistics were performed using Minitab software (version 14.0, Minitab Inc, PA, USA).

Inter-observer variability for traditional mRS

For each patient enrolled, two assessors allocated from the pool of seven performed mRS grading. Interviews were performed using a standard mRS approach or a structured interview, with choice of methodology randomly allocated. Thus patients had two independent assessments in succession, each using the same interview methodology (structured or standard mRS) and blinded to colleagues' grading. I used the previously validated, questionnaire style interview for the structured assessments as originally described by Wilson et al(62), with roughly half the assessments conducted using this structured interview approach.

This initial series of face to face paired interviews will be further referred to as "traditional mRS". I measured inter-observer agreement between the paired mRS assessments, first for all interviews and then with sub-analysis to compare structured interview against standard mRS. I further evaluated duration of interview for the structured and standard mRS interview using paired "Student's t" testing.

Intra-observer variability

One researcher from each interview pair was randomly selected for video recording. Following advice from Media Services Department, University of Glasgow, audio-visual recording was captured using a portable digital camera (HDVR-HC1E 1080i digital HD video camera recorder - Sony, Japan) and stored on digital video disc using readily available image processing software (Windows Movie Maker - Microsoft, Washington USA).

At a later date, the interviewer who performed the original mRS assessment viewed this recording and rescored mRS. I left a minimum three month delay between interview and assessment of recording to reduce recall bias. Repeat scoring was performed independently and raters had no access to their previous scores. Assessment of intra-observer variability was made comparing all raters' original, traditional mRS score to their subsequent video review score.

Estimating mRS

To gauge the added value of formal interview, raters were asked to grade disability prior to beginning their formal mRS assessment. This meant assigning a preliminary mRS using only such information as would be available in the first few seconds of patient interaction, for example assessment of patient mobility in the consulting room; or interaction with nurses or carers and initial conversation. This score was recorded and sealed in an opaque envelope. Raters then conducted and scored the formal mRS assessment. Properties of the preliminary mRS score were described by comparing these estimates to final mRS and by describing variability within the estimated scores

Results

Of 104 patients approached, 102 consented to mRS interview and video recording. Of these, 100 video recordings were of sufficient technical quality to allow repeat grading and were included in the final analysis. Patients reflected a heterogeneous group of stroke subtypes typical of 3-month survivors (total anterior circulatory stroke [TACS]16; partial anterior circulatory stroke [PACS] 30; lacunar stroke [LACS] 43; posterior circulation stroke [POCS] 11; unclassified 2). Mean age was 69.8 (SD 12.9) years; mean NIHSS score at baseline was 5.5 (SD 5.2) and median time since event was 12 weeks (IQR: 6 - 21). Five patients had problems with communication such that assessment involved a proxy.

Inter-observer variability for traditional mRS

Variability in traditional mRS grading was moderate (*k*=0.57) for the group of 100-paired interviews, with least variability at extremes of mRS (Table 7). Exact agreement in mRS was 67%. I compared reliability from my study to independent data from other studies using the literature described in the previous meta-analysis, there was no significant difference between my data and other published studies (P=0.073).

Of the traditional mRS assessments, 49 used a structured interview approach. There was no difference in spread of disability as graded on mRS between the two groups (p=0.699 on chi-square testing). Use of the structured interview did not decrease variability (P=0.295; Tables 7). Mean duration of mRS assessment was 4.9 (SD 2.4) minutes. There was a significant difference between duration of structured (5.6 SD 2.5 minutes) and unstructured (4.2 SD 2.1 minutes) interviews (p=0.003).

Table 7: Group reliability of traditional mRS assessment.

To quantify spread of disability the average number of patients scoring at each mRS grade is presented, these numbers were derived from total number of scores at a particular grade divided by the total number of assessors. Variability is described as un-weighted kappa (k).

Modified Rankin Score	Average no. of cases at this mRS	Variability - *k* Traditional mRS
0	16	0.70
1	16	0.43
2	33.5	0.51
3	16.5	0.57
4	15.5	0.66
5	2.5	0.79
Overall *k*	N/A	**0.57**
Structured interview *k*	N/A	**0.5**
Standard interview *k*	N/A	**0.64**

Intra-observer variability for mRS

One patient withdrew consent for video assessment following recording, leaving ninety-nine video assessments that could be reviewed and scored by the original mRS assessors. Intra-observer reliability was good for the group; *k*=0.72; 77% complete agreement, this differs significantly from other published studies of mRS intra-observer variability ($p<0.0001$). Intra-observer variability for individual raters were calculated, percentage agreement revealed no statistically significant difference between raters, although small numbers preclude definitive comparative statements and a trend towards differing reliability is seen (rater 1: 86%; rater 2: 89%; rater 3: 75%; rater 4: 40%; rater 5: 63%; rater 6: 100%; rater 7: 91%).

Estimating mRS

A convenience (non-randomised) sample of preliminary mRS interviews was included. As estimation of mRS is dependent on confidence in basic mRS administration, I included only the latter 40 mRS interviews in this analysis to eliminate any potential training effect. Agreement between estimated and final mRS was 38%, reliability was poor *k*=0.16. The median estimated mRS was 1.0 (IQR 0-4), median final mRS was 2.0 (IQR 0-4). Comparing estimated scores between the paired assessors, there was again poor agreement 30% and significant variability *k*=0.38.

Discussion

Using a mock clinical trial design I assessed reliability of mRS across a large number of patients. I have demonstrated substantial inter-observer and intra-observer variability in mRS assessment. Furthermore I have found that a structured interview approach does not significantly improve reliability and confirmed that researchers are poor at estimating mRS if they do not conduct a formal interview.

Despite considerable experience in clinical use of mRS, the local team of clinicians and nurses show only moderate reliability in mRS grading. This inter-observer variability is in keeping with previous published estimates presented in chapter two. With increasing use of mRS as trial end-point and ready availability of specific training resources some improvement in mRS reliability was expected. Diverse study methodologies preclude any more definitive comment on these differences, suffice to say that problems with reliability represent an ongoing limitation of standard mRS as a trial endpoint. In the absence of a pre-training "control", my findings do not allow us to comment on utility of the training resource or on any training effects associated with increasing experience of real life mRS administration.

Variability was most apparent for mRS grades 1-4. This is of particular importance for clinical trial endpoint analysis, where mRS outcomes are often dichotomised around these middle grades.(74) Misclassification of endpoints increases the likelihood of type I and type II statistical error and decreases

statistical power. The potential impact of mRS variability on clinical trial results has yet to be modelled, but we must assume that poor reliability will influence final results. Real life examples of trials compromised by variability in end-point classification are well recognised(37) and may be particularly relevant in the field of acute stroke.(28)

Quantification of intra-observer reliability for clinical scales is challenging and if methodology is poor there is potential for bias. Measuring test-retest variability over a short time period will be biased by observer recall of previous grading; delaying the second grading can allow for patient improvement or disease progression. Previous published studies have not accounted for these sources of bias in their design and as such the negligible inter-observer variability they report for mRS should be questioned.(63;97) The use of videos provides a more rigorous assessment of intra-observer reliability and may explain the significantly higher variability demonstrated. As trialists are unlikely to be performing serial mRS over short time periods, it could be argued that proving intra-observer variability of mRS is of little clinical relevance. However, I describe these findings as further evidence of the imperfections of standard mRS as an endpoint assessment tool.

Use of a structured interview approach to mRS assessment did not reduce inter-observer variability in this cohort. The authors of one questionnaire style structured interview previously reported significant improvements in reliability(63); however other groups have failed to replicate these findings(64) and at present the structured interview is infrequently utilised by stroke trialists. My results show that for experienced raters, fully trained in mRS

administration, use of a structured approach may have little to add. The difference in interview duration between the traditional and structured approach with no improvements in reliability, suggests that certain components of the structured interview may be redundant.

My final analysis described efficacy of initial limited disability assessment as a predictor of final mRS grading. Such an approach is not without precedent. It is recognised that for many scales, raters may not perform a comprehensive assessment; rather they will estimate final grading based on initial basic review and "clinical intuition".(146) For disability scales, including mRS, full assessment has been reduced to a limited number of key questions while preserving clinimetric properties.(147) The mRS is heavily weighted towards locomotor independence and so I hypothesised that distinction between higher and lower grades may be possible simply by observing the patient entering the clinic. I have shown that experienced raters are poor at predicting final mRS from initial assessment and that a formal interview is still required to grade disability.

A particular strength of this study was the mock clinical trial design simulating those situations where mRS is likely to be used. I adopted an inclusive policy, studying a large representative cohort of stroke survivors. I deliberately selected a panel of assessors from different clinical backgrounds as previous work has suggested that profession and training may impact on reliability of outcomes assessment.(108;132) Limited numbers of patients and use of assessors from similar backgrounds have compromised previous studies of mRS reliability. The use of video recording to assess intra-observer variability was

successful with minimal expenditure in terms of money and training. Other centres have also demonstrated efficacy of remote video based mRS assessment.(85) These results suggest feasibility of remote video based mRS assessment as a further aid to improve reliability.

Although the number of patients included in the analysis is greater than in many previous studies of mRS, there were still relatively few assessors and all were from the same department. Ideally I would have involved multiple centres in the analysis. In particular, to better represent the range of assessors seen in a contemporary trial, I should have included therapists and other health care workers as well as physicians and dedicated research nurses. In this regard this chapter's study is complemented by the work presented in chapter six describing moderate to good overall reliability on a five patient mRS assessment exercise across a large cohort of international trialists.

I deliberately chose to test a number of related hypotheses using a pre-defined structured design, thus deriving substantial data from a single clinical encounter. However, I pre-specified several hypotheses to limit the risk of drawing false conclusions as a result of multiplicity. My results do not negate the potential benefit of training and I would encourage trialists to continue to use specific mRS training resources.

Future trials designed to improve mRS assessment are planned, pending these results, these current data encourage caution in use and interpretation of standard mRS. As measures to date have not substantially improved mRS inter-observer variability a possible option for future trials is to limit the number of observers. Remote adjudication panel assessment of laboratory and imaging endpoints is commonplace in contemporary multi-centre trials and perhaps should now become routine for assessment of functional outcomes. I will explore this concept and a possible methodology for use in clinical trials in Chapter 8.

In conclusion, I have shown that despite increasing familiarity with mRS and availability of specific training packages there remains substantial variability in mRS that could compromise clinical trial results. Further measures to improve mRS reliability are urgently required. Possible strategies to improve reliability of functional outcome assessment including mass training of observers; group assessment and novel outcome measures will be described in the following chapters.

Chapter five

Initial experiences with a Digital Training resource for modified Rankin Scale assessment in clinical trials

Introduction

Potential problems of reliability have been demonstrated for standard mRS application, using both my own mock clinical trial and through review of other available literature. Various strategies have been proposed to improve reliability, including video recording(85) and use of a structured interview(63). The importance of reducing variability has already been discussed, but it is worth re-emphasising that reliability of the mRS is of more than clinimetric interest. Inter-observer variability will increase the risk of endpoint misclassification which can introduce bias, impact on type I or type II error rates(14) and ultimately estimated effect size will be reduced. It has been argued that statistical under-powering has contributed to the lack of significant treatment effects seen in recent acute stroke trials.(148)

Until recently mRS users had little guidance on rating. This lack of guidance likely contributes to the unfavourable inter-observer variability present in traditional mRS grading. Use of a training resource to improve consistency in the application of the mRS makes intuitive sense.(149) In the original work describing inter-observer variation in mRS scoring, the authors commented that improvements could be achieved if observers were afforded the chance to practise use of the scale but that "…such training is hardly realistic in the context of a multi-centre trial."(20) Recent improvements in audio-visual technology mean mass training across a number of centres is now feasible and economical.

A variety of potential formats are available for training. A stand alone package would allow better standardisation and dissemination than a lecture based series while a programme incorporating “live” assessment of real patients is preferable to purely text based instruction. The former approach is not without precedent. A well validated video based training programme and certification procedure exists for the NIHSS and it is now a requirement that clinical trial investigators complete this training and undergo certification.(46) The strengths of the NIHSS programme - in particular its mix of didactic teaching, video explanation and assessment procedure – could be easily applied to the mRS.

As a department, we developed a training digital video disk (DVD) and accompanying explanatory booklet, which include recordings of real Rankin assessments, certification cases and further recertification cases. This has been used successfully in two large scale clinical stroke trials.(79;150) Although brief reference to certification scoring has been made in a previous review(28), to date there has been no detailed description of the training package, its development and the initial experience of its use. I present this here along with the initial results of the certification programme. A more detailed analysis of training scores will be described in the proceeding chapter.

Methods

Development of the mRS training - audio visual Issues

A criticism of the early NIHSS video training was poor image quality, an important consideration as subtle clinical signs may have a substantial effect on final scoring.(46) In an interview based assessment such as the mRS, high quality sound recording is of equal importance. To ensure optimal audio fidelity expert technical help was enlisted (Media Services Department, University of Glasgow). A recording studio was used for the majority of interviews, where patient disability made travel to the studio inappropriate, the Media Services mobile crew filmed in the hospital.

A DVD based format was chosen for the recorded materials contained in the training resource. This reflects the gradual replacement of conventional VHS recordings with DVD and the easy availability of economical DVD players. It also recognised the improved clarity of sound and vision afforded by this format. Digital recording ensures optimal quality even after mass reproduction. Finally, this allows for ease of transfer to a web based server. Internet based NIHSS training has facilitated cost-effective global dissemination of the training package.

Patient Selection

The stroke liaison team based in the Western Infirmary Glasgow selected suitable patients from recent admissions to the acute stroke unit or local inpatient rehabilitation facility. The stroke unit accepts all patients presenting within 72 hours of onset of suspected stroke irrespective of age or severity of the neurological deficit, thus a cohort of patients with varying degrees of disability and background co-morbidity were available. The intention was to include at least one patient from each potential mRS category. Thirteen patients were selected of whom four were used as training cases, five were used for initial assessment and four were used for recertification. To reflect clinical practice, in one case disability was such that answers were provided by a carer. Although the final selection of patients was a sample of convenience, they were felt to be representative of a "real life" cross section of post stroke outcomes and to be suitable for inclusion in the training package. Consent for videotaping and use for training and research purposes was obtained from all patients in line with national and local protocols.

For the training component of the package, two patients with easily categorised disability were chosen as initial "introductory" cases. The remaining two training cases were chosen to highlight perceived problem areas in the application of the scale. Each of the training cases was followed by an explanatory discussion of the correct mRS score and the rationale for this grading. An accompanying booklet gave background information on the general principles of mRS scoring, including detailed definitions of the categories and discussion of what is considered to be best practice in the application of the scale. In formulating this advice, reference was made to the original description of the modified

Rankin and to recent work using a structured interview.(63;92) To minimise potential language problems a transcript of the text was provided with translation into local language. Assessment for certification was performed in a variety of settings including individual viewing of the cases, group viewing within a centre and supervised group viewing sessions at formal training meetings.

Recording and Scoring of the Assessments:

Recordings of the interviews were analysed and scored independently by two observers who were both experienced in mRS grading (Professor KR Lees [KRL] and Doctor HG Hardemark [HGH]). No attempt was made to script the mRS assessments and there was little post interview editing. It was immediately identified during the piloting of the study that some of the answers given by patients were ambiguous and in at least two of the cases debate arose as to the most appropriate category. However, a decision was made to keep the complete interviews as recorded - it was felt that scenarios artificially scripted to fit a mRS grade neatly would not have adequately prepared assessors for the difficulties inherent in grading real patients.

Scoring for the assessment component of the package took account of those patients who did not unequivocally fit a single grade. A final decision on correct grading was made by KRL and HGH, supplemented by the analysis of results from an international pilot involving 100 participants. A correct grade was defined as one assigned by both trainers and by >50% of trainees. A grade of “acceptable” response was defined arbitrarily as one which was deemed by the trainers to be incorrect but that followed the basic scoring guidelines and had been assigned by a substantial minority (10-49%) of assessors in the pilot. Any grade offered by

<10% of assessors or that clearly did not follow the accepted scoring was defined as unacceptable. Using these scoring rules any candidate who graded all cases correctly, including "acceptable" answers, was awarded certification (Table 8). This scoring system was developed so that certification was only awarded to assessors who demonstrated a good knowledge of mRS application, but recognises that some merit should be given for "acceptable" but incorrect answers. To minimise variability, assessors were instructed to choose the more severe mRS grade when hesitating between two scores.

Certificates of completion were awarded, along with separate confidential feedback on actual score achieved to all who achieved the target. Scores for individual patients were not released. Assessors who did not achieve certification on first attempt were encouraged to review the training material and resubmit an amended set of grades for the full set of certification cases. No specific feedback on errors made was provided. Candidates could attempt certification as often as required. Successful completion of the training materials and examination was a prerequisite for trialists involved in the NXY series of neuroprotection trials.(41;120;150)

Table 8: Scoring system for certification using the mRS training resource.

Explanation of the scoring criteria used to define "correct" and "acceptable" gradings are described in the main body of the text.

Grading	Certification
5/5 cases correct	Pass (Qualified 5)
4/5 cases correct, 1 acceptable	Pass (Qualified 4)
3/5 cases correct, 2 acceptable	Pass (Qualified 3)
Any other combination	Fail (Not Qualified)

Results

The majority of respondents were part of the investigating team from countries involved in the "Stroke – Acute Ischemic – NXY-059 Treatment" (SAINT-I)(120) trial only or in both the SAINT-I and the "Cerebral Haemorrhage and NXY-059 Treatment" (CHANT) study.(150) Assessors comprised a mixed group of principal and co-investigators, study nurses and research assistants (Table 9).

The correct mRS scores for the certification cases, along with the proportion of scores assigned by observers are shown in Table 10. To allow continued use of the training resource, the "correct" scores for assessment and recertification were not made public and are shown in a different sequence to the cases on the DVD.

There was a spread of opinion on all of the cases, with submitted answers spanning 3 – 5 Rankin grades for each answer. For three of the cases the majority of respondents opted for the correct grading; in two of the cases opinion was split with a substantial proportion of assessors (39.2% and 40.4%) choosing a lower mRS grading than the correct answer. This variation in opinion was accounted for in the final scoring, and the lower grading was defined as an "acceptable" answer. Twenty three assessors gave two scores, despite explicit instructions to choose the best single score. If the scores given were the correct and an acceptable score then a pass was allowed otherwise the grading was considered invalid.

Percentages of respondents achieving acceptable scores for the certification assessment are given (Table 9). The majority of assessors 1464, (81.3%), achieved a "pass" on the certification exercise. However, only 38% of these individuals graded all of the five cases correctly. The remainder of the group comprised those assessors who wrongly assessed one or both of the previously described equivocal cases, but whose assessment was still defined as acceptable. Of the 336 who did not achieve certification on first attempt, 85% scored a "pass" on second attempt.

Table 9: mRS certification scores by background training.

For explanation of Qualified 5, 4, 3 see Table 8.

Position	Not Qualified	Qualified 3	Qualified 4	Qualified 5
Co-investigator (n=747)	125 (16.7%)	96 (12.95)	224 (30.0%)	302 (40.4%)
Principle investigator(n=159)	28 (17.8%)	18 (11.3%)	48 (30.2%)	65 (40.9%)
Study nurse (n=212)	53 (25.0%)	29 (13.7%)	64 (30.2%)	66 (31.1.%)
Other (n=682)	130 (19.0%)	71 (10.4%)	228 (33.4%)	253 (37.2%)
Total	336	214	564	686

Demographic data on assessors submitting for certification were collated. Results of the certification assessment are presented by training (Table 9). A more detailed analysis of mRS training scores by country, centre participating in trial and profession will be presented in the next chapter.

It was intended that the recertification process would be undertaken at one year post initial training, as such full data on recertification are presently limited to only 370 results. Of these only 6.5% assessors failed to achieve a satisfactory score (Table 10).

Table 10: mRS grades submitted for certification and recertification (%).

*To allow for future use of the training resource, mRS grades are **not** presented in the order they appear during the assessments.*

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
mRS 0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1
mRS 1	0.0	0.9	26.5	2.7	0.0	0.0	0.0	0.0	2.2
mRS 2	0.7	55.0	72.6	39.2	0.0	1.5	2.3	0.1	40.4
mRS 3	6.3	44.1	0.7	58.1	0.0	3.7	90.4	1.7	56.2
mRS 4	92.1	0.0	0.0	0.0	14.0	93.8	7.3	97.4	1.0
mRS 5	0.9	0.0	0.0	0.0	86.0	0.9	0.0	0.6	0.0
Correct mRS	**4**	**2**	**2**	**3**	**5**	**4**	**3**	**4**	**3**

Discussion

Consistency in grading of post stroke disability is crucial both in daily clinical work and in the context of a clinical trial, although with potentially hundreds of assessors grading endpoints, for contemporary trial purposes, consistency is more important than accuracy and consistency in trials is more important than in routine clinical work where a single clinician assesses the patient. Potential for significant variation in application of the modified Rankin scale has been demonstrated in previous chapters. The digital training resource for mRS grading was developed in an attempt to improve the situation. My results show that mass training of observers in use of the mRS is achievable in the context of a clinical trial via the use of a novel DVD training package.

Several issues arose during development of the DVD which deserve comment. No specific criteria were used to select patients for the training or certification components of the package, although patients used for assessment were judged to be suitably taxing to allow a valid assessment of ability. A clustering of grades around the mid-range was noted in the cases selected to be used for the certification process. It has been shown that clinicians are comfortable to assign grades at the extremes of the Rankin Scale, possibly because these grades are well defined or because deviation can be in one direction only. This pattern was also demonstrated in my own study of mRS reliability, described in chapter four. Thus, it was decided to proceed to use this relatively biased sample. Furthermore, it is in this mid-range of scores that the reliability of the mRS and discrepancies between raters assumes the greatest importance; mRS based

outcomes are frequently dichotomized, with mRS ≥ 4 defining poor outcome (DESTINY trial)(138) or scores ≤ 2 defining good outcome (ECASS II).(151)

It is difficult to measure adequately the "success" of a multidimensional intervention such as an educational resource. Analysing only pass rates for the assessment is a relatively crude measure, as it is likely that even those viewers who failed the initial certification will have gained improved knowledge of the mRS scores and assessments. Despite this, the high rate of satisfactory scoring on the certification exercise is reassuring.

Assessors were given confidential feedback on their total score. This allowed all users to review the cases and perhaps to correct any grading errors. Data are not available for all the second attempts at certification, which raises the possibility of sample bias but the high pass rate seen provides further support for utility of the training package. Given that the training package is the only educational resource available for training it is safe to assume that this improvement was achieved with no extra tuition other than repeat viewing of the package and knowledge of previous scoring.

Purists will argue that my data do not conclusively prove the benefit of the training package. The primary purpose of the package was to improve reliability of mRS grading in large clinical trials so a "control" arm of assessors not exposed to training was not factored into this original study. Extrapolating evidence from the success of the NIHSS and other video certification schemes(46) suggests this was a logical decision. Assuming that the training will not worsen

reliability, it was felt to be unethical to deprive trialists of the training in an effort to prove efficacy of the resource.

It is widely accepted that there is too little formal guidance on application of the mRS(17) and few would argue against an attempt at formalising its use. Anecdotally, feedback from participants has been uniformly positive and there is little to suggest that exposure to the digital training worsens mRS grading or introduces systematic bias. Even if the training were to influence grading systematically, it could be argued that if all assessors were taught to grade in the same fashion, and in a manner which reflects the mRS categories, that this could only improve outcome assessments for trial purposes.

Pragmatic evidence of utility of the training package comes from its application in the SAINT-1(120) and CHANT studies.(150) During conduct of these studies over 1500 investigators were trained. It cannot be proven that this did improve the quality of endpoint assessment but demonstrates the feasibility of mass training and given the inherent problems with use of the Rankin scale is unlikely not to have helped. I believe that formal mRS training should be routine for all acute stroke trials. It is recognised that some questions as to optimal delivery of the package remain unanswered, such as how best to address the issue of repeated failure to achieve certification and whether training should be performed alone or in a group setting with the opportunity to discuss the cases and content. Work is ongoing to answer these questions(46) and already there is scope for further improvements, for instance making the training available on the internet could ease dissemination to the target audience. An online mRS training resource that makes use of the original training materials described in

this chapter has now been launched by "Training Campus" (http://trials-rankin.trainingcampus.net) [last accessed January 2010] and collation and monitoring of training data continue.

Although accepted in the stroke literature from inception, the modified Rankin was not clinimetrically tested prior to its use.(92) These data being collated from the Rankin certification process provide a powerful tool for better definition of the properties of the scale. Analysis of the inter-observer and intra-observer variability of the scale with further sub-analysis of individual components of the scale and relationships to country of origin and level of training will be presented in the next chapter.

We live in a digital age, and utilising the available technology to deliver educational resources makes scientific and economic sense. Strategies to improve the reliability of the mRS are needed, and it is likely that electronic dissemination of teaching material to participants in multi-centre clinical trials will be widely used in future. I have demonstrated that digital training in post stroke assessment is feasible and accepted by most potential assessors. Further work to quantify the potential impact of such training on the quality of future stroke trials is required.

Chapter six

Variability in modified Rankin Scale scoring across a large cohort of international observers

Introduction

Through my own study and review of the available literature I have attempted to describe the variability seen when mRS is used in a clinical trial setting. Individual studies of mRS properties provide useful data; however, by limiting themselves to a few experienced raters from a single centre or city they likely underestimate the true variability of mRS assessment manifest in a large scale, international trial.

To improve quality of multi-centre trials a standardised approach to outcome measurement is required. One potential method of improving consistency is to offer training and examination in endpoint classification. Case-based training in application of the NIHSS is well established, and as discussed in the previous chapter, successful completion of the NIHSS training is a pre-requisite for several stroke trials.(45;46) As a department, we have developed a comprehensive training package for mRS assessment, consisting of: written educational materials; video based tutorials and mRS cases for grading. A full description of the package is available in the previous chapter. In brief, investigators study the training resources and then attempt an assessment exercise that comprises a series of real time mRS interviews for grading. Examination attempts are externally graded and success leads to certification in mRS training.

The mRS training resource has been available for almost four years and has been widely used. Thousands of assessors from various countries and backgrounds have attempted the certification exam. These data have been collated centrally and offer a powerful resource for analysis of mRS variability and its

determinants. Using these training data I describe variability in mRS assessment across a large cohort of international researchers. To explore reasons for mRS variability I further described mRS reliability by country, native language and background speciality of assessor.

Methods

MRS training data

The video based mRS certification exercise comprises 5 non-scripted interviews with stroke survivors. There is a further recertification exam comprising 4 cases. Interviews were originally recorded in English. Fully translated training packages, with native speakers overdubbing the interview, have been made available for Finnish, French, German, Italian, Portuguese, Spanish and Swedish researchers; a subtitled Chinese version has also been produced.

Certification data are collated and scored centrally. Investigators can submit responses individually or through a local study co-ordinator. Standardised paper or electronic score sheets are used and mRS grades are then transcribed directly onto an electronic spreadsheet. Score sheets that are incorrectly completed or poorly legible are returned to the assessor for resubmission. An assessor who fails to achieve certification is invited to review the training materials and resubmit on the assessment exercise, no guidance is given on errors in original grading. There is no limit on the number of attempts an assessor can make. Initial certification is valid for one year after which time raters are invited to take a further assessment.

"Correct" mRS grades for each video interview were derived by two experts in mRS grading and were informed by the results of previous pilot data. MRS certification is graded using a "pass" / "fail" system. Embedded formulae within a dedicated electronic database automatically calculate the investigator's final grade. Again, complete description of the process used to score mRS grading is available in chapter five.

Certification data were anonymised prior to analysis of reliability; to allow for pre-specified analyses: data on participating centre; location (country) of participating centre and profession of the rater were maintained, any other identifying information was removed. I did not seek formal written consent to use these anonymous data in my study of reliability. Consent was assumed as some degree of data collection and analysis is evidently necessary for scoring and quality control. As with any registry, participants can ask for details to be removed. Throughout the period of data collection and since my publication of the mRS training system methodology I have received no such requests.

Statistical Analyses

To assess for significant differences in certification performance between countries; professions and centres I used chi-square analyses, comparing proportions achieving the following certification results: fail; pass 3/5 correct; pass 4/5 correct; pass 5/5 correct. As numbers in the group "pass 3/5" were small, to allow for meaningful statistical analysis, the group "pass 3/5" were combined with the "fail" group.

To maintain consistency with all previous analyses, variability was described using kappa statistics (k), where k=1 defines perfect agreement between assessors, while k=0 defines no agreement other than that expected by chance. Agreement was described across the cohort of observers and against the "standard" of pre-defined correct answers. Using an equivalent approach I also described variability at each potential mRS grading. I performed two principal analyses: first describing reliability for assessors' initial attempt at the mRS exercise; and a second analysis limited to those assessors who successfully completed the exercise. Further sub-analysis was performed to describe agreement by country and by native English and non-native English language countries. Native English speakers were arbitrarily defined as any assessors from centres in Australia, Canada, New Zealand, South Africa, United Kingdom and United States of America.

Variability by background profession was also described. Background profession was classified as Neurology; Geriatrics/Care of the Elderly; specialist stroke physicians not specifically trained in neurology or geriatric medicine were classified as "General Medicine". Non-physician, clinical research assistants

were classified as “Research Nurses”. Background was extracted from submitted assessment documentation. Where background was not available these data were collected from the host institution. To eliminate the effect of language and country I limited the background analysis to UK based assessors. To assess for potential bias in grading, median and Inter-quartile range (IQR) were calculated for each grading, although these ordinal data should be described with non-parametric statistics, to further describe any patterns in grading I also calculated mean mRS and standard deviation. An equivalent analysis describing variability by institution was performed for UK assessors.

All available certification attempts have been included in the analyses. Kappa statistics were described using attribute agreement functions. Statistics were performed using Minitab software (version 13.1, Minitab Inc, PA, USA).

Results

Certification assessments have been collated since March 2003. The total number of assessments to date is 2,942 (2636 1st certifications [5 cases]; 306 re-certifications [4 cases]). Thus, data on 14,404 mRS assessments were available. Certification cases spanned a range of potential mRS scores: for mRS grade 2 I included 1 video case for assessment; mRS 3 = 4 cases; mRS 4 = 3 cases; mRS 5 = 1 case.

Country of origin was available for 2349 certification attempts. Assessors were based in a variety of international centres (n=30 countries). The majority of assessors (1958, 75%) achieved certification at first attempt; 20 assessors required a 3rd attempt at certification; 4 assessors required 4 or more attempts.

Proportions of countries achieving certification grades of: fail; pass 3/5 correct; pass 4/5 correct and pass 5/5 correct are presented graphically for all countries submitting more than 50 assessments. Presented data represent performance according to assessor's first attempt (Figure 5a) and performance limited to those assessors who achieved certification (Figure 5b). There was a significant difference in the performance of countries for both analyses ($P<0.0001$).

Figure 5a: Performance on the mRS certification exercise, first attempt.

Data are for all countries submitting more than 50 assessments.

Key:

nq=not qualified

Q3=qualified 3/5 answers correct;

Q4=qualified 4/5 answers correct;

Q5=qualified 5/5 answers correct.

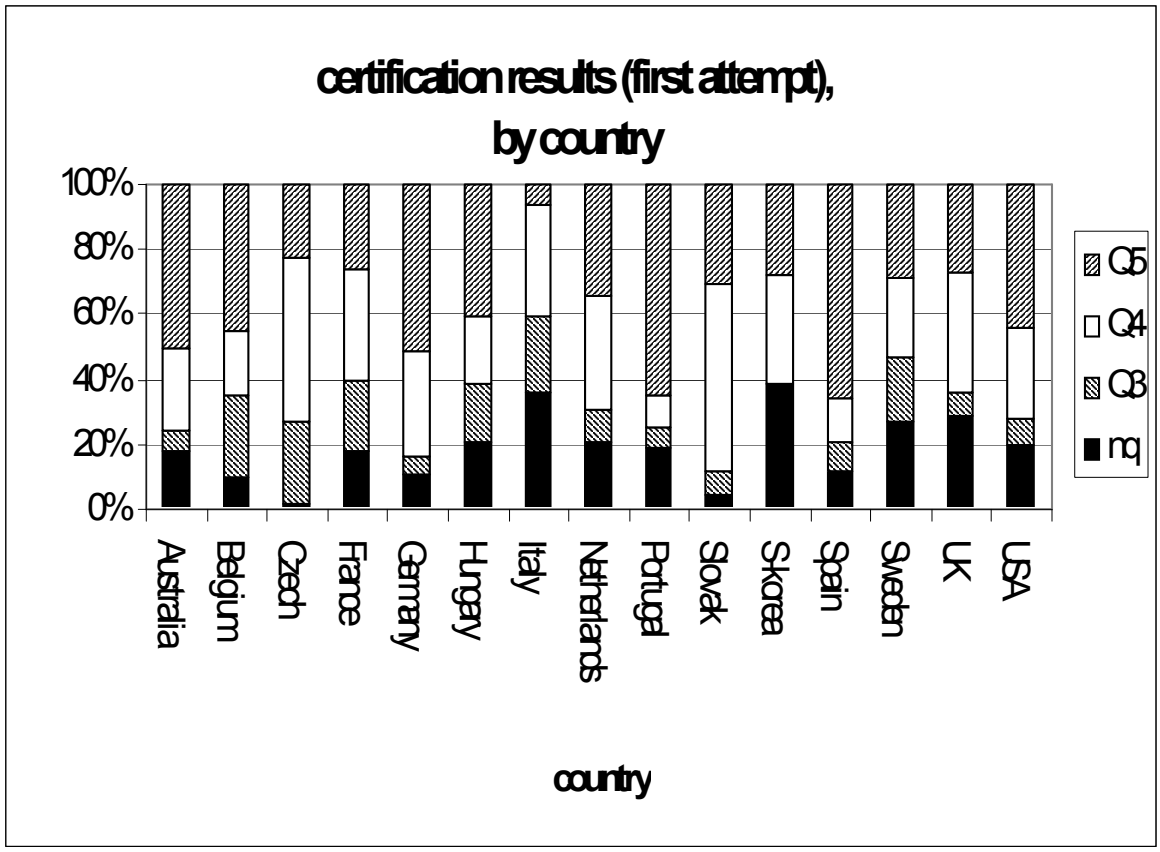


Figure 5b: Performance on the mRS certification exercise, limited to researchers who passed the certification exam.

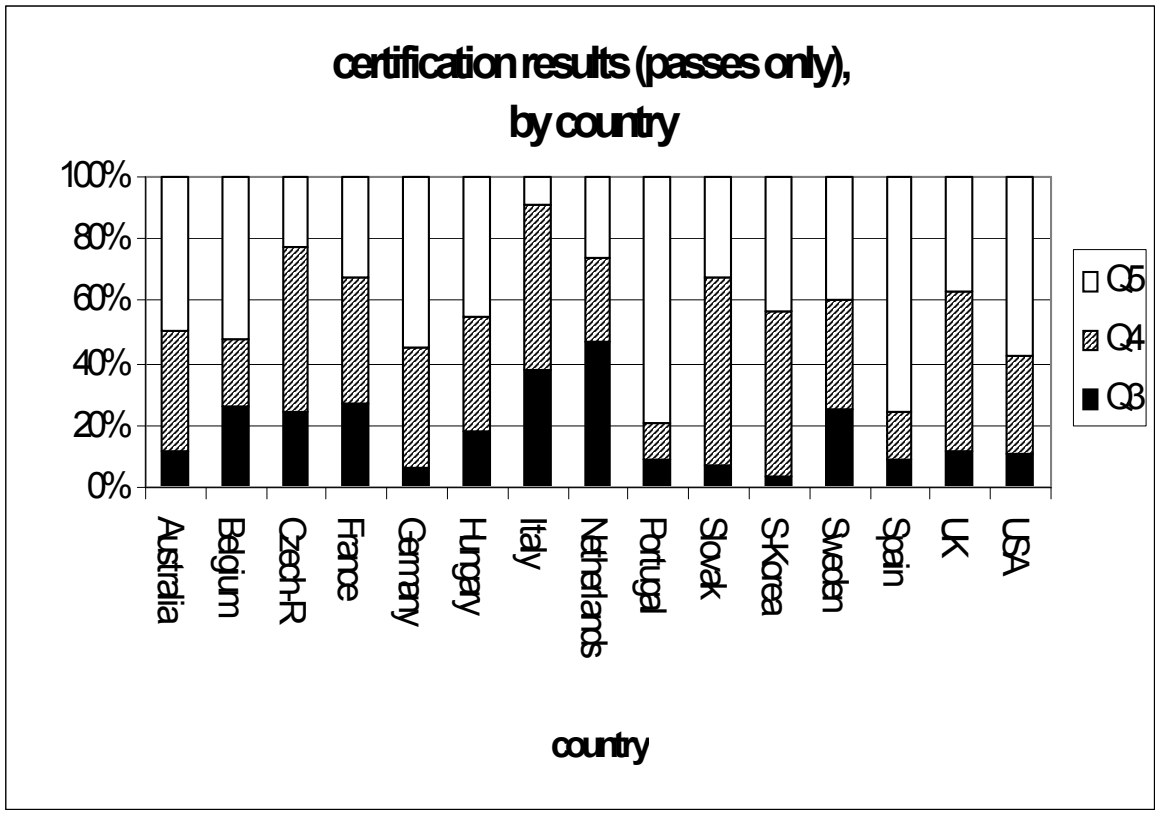
Data are for all countries submitting more than 50 assessments.

Key:

Q3=qualified 3/5 answers correct;

Q4=qualified 4/5 answers correct;

Q5=qualified 5/5 answers correct.



Inter-observer variability and variability against the “correct” grade are presented for all countries submitting more than fifty certification assessments, along with total variability across the cohort (Table 11). Other countries included in the overall analysis comprise a mix of Asian, European, Canadian and African centres. Variability by country ranged from fair to very good. Variability for the entire cohort was good, with no difference between English speaking and non-English speaking countries. For the complete cohort of assessors variability at each level of mRS grading is presented (Table 12).

UK assessors comprised a mixed group of disciplines. Although reliability in mRS grading varied with professional background (Table 13), there was no significant difference in certification assessment results (P=0.321). All groups tended to underscore disability. Heterogeneity in reliability was also present across the various UK centres. For those centres with greater than 5 raters completing the examination (anonymous) results were: centre 1 (n=22) *k*: 0.59; centre 2 (n=12) *k*: 0.70; centre 3 (n=11) *k*: 0.74; centre 4 (n=11) *K*: 0.43; centre 5 (n=6) *k*:0.67; centre 6 (n=6) *k*:0.80. For this analysis there was a significant difference in performance across the UK centres P=0.001.

Table 11: Modified Rankin Scale variability by country of assessor, tabulated for all countries submitting more than 50 certification attempts.

Variability is measured using un-weighted kappa statistics (k) and is presented as both inter-observer variability and variability against a standard of pre-defined "correct" grade.

Two analyses are presented:

a) observers first attempt at assessment exercise

b) limited to those assessors who achieved an overall "pass".

Reliabilities scored as "moderate" or poorer (using standard scoring criteria) are highlighted in bold text.

	(a) 1st attempt	(b) Passes only
Australia n=111 attempt 1/ n=110 pass	*k*=**0.60** (0.77 standard)	*k*=0.79 (0.86 standard)
Belgium n=49 / 46	0.64 (0.73)	0.72 (0.78)
Czech Republic n=49 / 48	0.70 (0.68)	0.74 (0.70)
France n=72 / 67	**0.60** (0.64)	0.83 **(0.52)**
Germany n=162 / 159	0.66 (0.78)	0.77 (0.84)
Hungary n=57 / 54	**0.60** (0.70)	0.75 (0.79)
Italy n=147 / 130	**0.55** **(0.34)**	0.62 **(0.38)**
Netherlands n=58 / 49	**0.53** **(0.50)**	0.75 (0.76)
Portugal n=63 / 53	0.66 (0.80)	0.84 (0.91)
Slovakia n=42 / 40	0.72 (0.75)	0.75 (0.77)
South Korea n=55 / 30	**0.52** (0.67)	0.74 (0.81)
Spain n=314 / 299	0.73 (0.84)	0.83 (0.90)
Sweden n=56 / 41	**0.55** (0.65)	0.71 (0.74)
UK n=109 / 95	**0.59** (0.69)	0.74 (0.77)
USA n=172 / 162	0.61 (0.73)	0.77 (0.84)
Native English n=580 total / 389 pass	0.66 (0.77)	0.69 (0.77)
Non-native English n=1769 / 1251	0.67 (0.76)	0.71 (0.78)
All n=2942 / 2151	0.67 (0.76)	0.71 (0.78)

Table 12: Variability in mRS scoring across a large cohort.

Columns a) and b): Variability described using standard kappa statistics (k) at each grade of modified Rankin Scale,

Columns c) and d): Median and mean mRS at each "correct" grade for all attempts at mRS; restricted to those who achieved certification.

NB original assessment exercise did not include patients from full range of disabilities.

mRS	Inter-observer variability	Variability against standard	Median (IQR) Mean mRS (SD) All assessors	Median (IQR) Mean mRS (SD) "Passes" only
0	-	-		
1	0.19	-		
2	0.48	0.56	2 (1 - 2) 1.7 (0.46)	2 (2 - 2) 1.8 (0.47)
3	0.74	0.79	3 (3 – 3) 2.8 (0.51)	3 (3 - 3) 2.8 (0.40)
4	0.95	0.97	4 (4 - 4) 3.9 (0.43)	4 (4 - 4) 3.9 (0.31)
5	-	-		

Table 13: Inter-observer variability (*k*) with variability against "standard" mRS (standard) in mRS scoring and median / mean submitted grade for UK assessors by background profession.

Variability is described using standard kappa statistics.

Data are presented as median (Inter-quartile range) and mean (95% confidence interval)

	k ***standard***	**Median (IQR)** **Mean (95% C.I)** **mRS 2**	**Median (IQR)** **Mean (95% C.I)** **mRS 3**	**Median (IQR)** **Mean (95% C.I)** **mRS 4**
Gen. Med. (n= 13)	0.66 0.77	2 (1 – 2) 1.67 (1.35, 1.98)	3 (3 – 3) 2.87 (2.72, 3.02)	4 (4 – 4) 3.95 (3.87, 4.04)
Geriatrics (n= 23)	0.54 0.68	2 (2 – 2) 1.91 (1.73, 2.10)	3 (3 – 3) 2.91 (2.73, 3.10)	4 (4 - 4) 3.89 (3.79, 3.98)
Neurology (n= 16)	0.56 0.65	2 (2 – 2) 2.00 (1.80, 2.19)	3 (3 – 3) 3.03 (2.78, 3.28)	4 (4 – 4) 4.00 (4.00, 4.00)
Res. Nurse (n= 58)	0.65 0.72	2 (2 – 2) 1.79 (1.67, 1.91)	3 (2 – 3) 2.77 (2.66, 2.88)	4 (4 – 4) 4.00 (3.97, 4.02)

Discussion

I have demonstrated considerable inter-observer variability in mRS grading across a large cohort of international investigators. Variability was apparent for all included countries, however within the cohort there was substantial heterogeneity with a number of countries achieving only fair – moderate reliability. These data confirm the potential for end point misclassification in a clinical stroke trial and suggest some reasons for this variability in scoring.

The statistical techniques used to describe variability demand some discussion. As has been previously discussed, there is no accepted standard test for measurement of reliability. Use of kappa statistics has been criticised, as the basic assumptions underlying the calculations rely on observer independence.(152) I recognise that complete rater independence can never be guaranteed in a trial setting but chose to use kappa statistics in this study as clinicians are familiar with the test and previous studies of mRS reliability have used a similar approach. Traditional kappa statistics do not allow for comparative analysis. To assess whether the differences seen between countries were significant I compared proportions achieving a “fail” or one of the “pass” grades using accepted techniques. Having thus established a significant difference in mRS grading between countries, I then described the inter-observer variability in terms of kappa statistics.

I present two analyses of inter-observer variability that describe differing aspects of mRS reliability. Analysis of assessors' first attempt at mRS gives an approximation of reliability across all potential stroke investigators. The second analysis, limited to assessors who successfully completed the training exercise, provides a better representation of the variability that would be seen in a contemporary stroke trial (the majority of trials that make use of the mRS training resource demand successful certification before the investigator can assess trial patients). Although this second measure is consistently better than the first, there is still substantial variability with heterogeneity across countries. By defining "correct" mRS grades for each patient interview, I was able to compare variability against a pre-defined standard. For certain countries there was a marked discrepancy between inter-observer variability and variability against the standard: this suggest that cohorts of raters were consistent in their grading but that this grading was inaccurate.

These data compare favourably with previous estimates of mRS inter-observer variability described in chapter two. I would hope that any improvement in reliability will in part represent the beneficial effect of the mRS training resource and perhaps increasing familiarity with mRS as a method of functional outcomes assessment. However, there is no room for complacency – even those assessors who successfully completed the assessment demonstrated variability. Thus although my results are encouraging there is still some way to go before mRS variability is minimised.

Across the complete cohort of assessors, inter-observer variability was greatest for grades 1 and 2. This finding has been demonstrated both in my departmental study of mRS reliability and in the other mRS studies. For clinical trial endpoint analysis, mRS outcomes are often dichotomised at grades of 1,2 or 3. Thus, in this training cohort, variability is potentially greatest for those mRS grades most likely to influence final trial result. Increased variability at these middle grades is well described. It may be attributable to better definition of the highest and lowest categories, or to the potential for misclassification in one direction only at extremes of mRS.(31)

The substantial variation in reliability observed between countries is intriguing. These data suggest that this is not purely a function of language as countries with native English performed similarly to other countries. I acknowledge the global nature of contemporary medical practice where a given institution may have a number of international staff and that most centres recruiting for international trials will be staffed by teams familiar with English. It is possible that socio-cultural factors related to perceptions of disability and handicap may influence patterns of grading. For this reason, in collaboration with the University of Glasgow, international centres are producing new assessment cases in native language and featuring local stroke survivors. I recognise that my data will include countries that may have a number of centres relatively inexperienced in stroke trials / mRS assessment. Increasing use and familiarity with the scale may help to remove some of this variability. The available data do not allow for assessment of training effects, I encourage stroke trialists to continue to use available and forthcoming recertification materials as data from

these resources will allow for future analysis of training effect and hopefully should demonstrate greater improvements in reliability.

My analyses of UK assessors suggest that heterogeneity in measured reliability is not only accounted for by nationality. Accepting the smaller numbers of certification attempts available, there was again heterogeneity between centres and between professions, although only the analysis across differing centres revealed a significant difference in performance. Inter-observer variability between differing professional backgrounds has been demonstrated for other neurological outcome scales including the Barthel Index(153) and the Unified Parkinson's disease rating scale motor examination.(132)

Some aspects of my methodology demand discussion. Although describing variability through performance on a video based assessment allows for standardisation across a large number of raters, there is the potential to overestimate reliability using this method. Variability recorded in a series of traditional face to face mRS gradings would likely be substantially higher, as assessors can employ various approaches to patient interview. In creating the mRS assessment I had to strike a balance between including a broad range of cases at varying levels of disability and having an assessment that was short enough to be acceptable to a large population of researchers. The final choice of included cases deliberately focuses attention on those mRS grades known to demonstrate greatest variability i.e. mRS 2 -4. For future training packages more cases from extremes of the mRS spectrum will be included.

The strengths of my analysis are its size and international scope. Major acute stroke trials may involve as many as 400 hospitals at international sites, each with 2-5 raters. Previous descriptions of the clinimetric properties of mRS have been limited to few observers often from single centres. Certification in mRS grading was required for a number of recently completed and ongoing multi-centre stroke trials and investigators from many of the leading stroke research centres are included in this analysis. The initial users of the novel resource may well have over represented enthusiasts and specialists in the field who already have considerable experience of mRS. For this reason I waited until the mRS teaching package was well established before attempting any analysis of the training data.

Having demonstrated this inter-observer variability in mRS, it is incumbent upon stroke trialists to take steps to improve reliability in outcomes assessment. The DVD training package was designed for this purpose and has been a success. However, even those raters who achieve certification still demonstrate a degree of inter-observer variability and I suspect that training alone will not eradicate variability. Other methodologies to improve mRS assessment are available or are being developed and certain will be described in the following chapters of this thesis.

I have demonstrated inter-observer variability in a large representative cohort of international researchers. Although country and background profession seem to influence this variability, the significant differences between similar local UK centres suggest that reasons for variability are more complex than simple socio-geographic differences. As variability between assessors may never be fully explained the stroke community should continue to pilot novel methodologies to minimise inter-observer variability in clinical trials.

Chapter seven

Deriving modified Rankin Scale grades from patient case records

Introduction

In my discussion on the use of functional outcome measures in contemporary stroke trials (chapter three) it was established that mRS is the preferred disability outcome scale and that mRS data have been collated using a variety of assessment methodologies. Traditionally mRS grading has been based on face to face(20) or telephone interview.(64) Such an approach is possible for a prospective trial, but does not allow for retrospective disability grading. Previous observational studies have attempted to derive mRS using information contained in patient case-records.(133;154) The clinimetric properties of such an approach have not been described. We should be cautious of assuming that novel methods of outcomes assessment will provide robust data.

Assessment of functional capacity is an important element of stroke clinic review. As each mRS grade describes a broad range of disability, reasonable estimation from narrative case-record information should be possible. Several stroke assessment scales in common usage can be successfully derived from routinely collected data. The NIHSS(155); the Canadian Neurological Scale (CNS)(156) and the Scandinavian Stroke Scale (SSS)(157) score have all been derived with acceptable validity and reliability.

I hypothesised that mRS could be derived accurately and reliably from information recorded at outpatient follow up.

Methods

Patients were recruited through the local stroke unit. The Western Infirmary acute stroke unit manages all patients with known or suspected cerebrovascular disease admitted to the main hospital. The department is based in a large teaching hospital with a predominantly Caucasian, urban patient demographic. At discharge from the unit, patients are allocated three-month outpatient hospital clinic follow up. A sequential series of these outpatients consented to participate in the study of video based mRS assessment described in chapter four. From this trial population I further selected patients for inclusion in the mRS derivation study using an online random sampling process (www.random.org). [last accessed January 2010] The study had full ethical approval, with patients or proxies providing written consent.

Patients attending the clinic for their routine consultation were first seen and managed by a stroke clinician according to normal practice. Doctors leading the outpatient consultation were not aware that their case-record notes would be used for retrospective analysis. The unit provides no guidance on documentation during clinic review and does not use pre-specified "pro-formas" in the clinical setting.

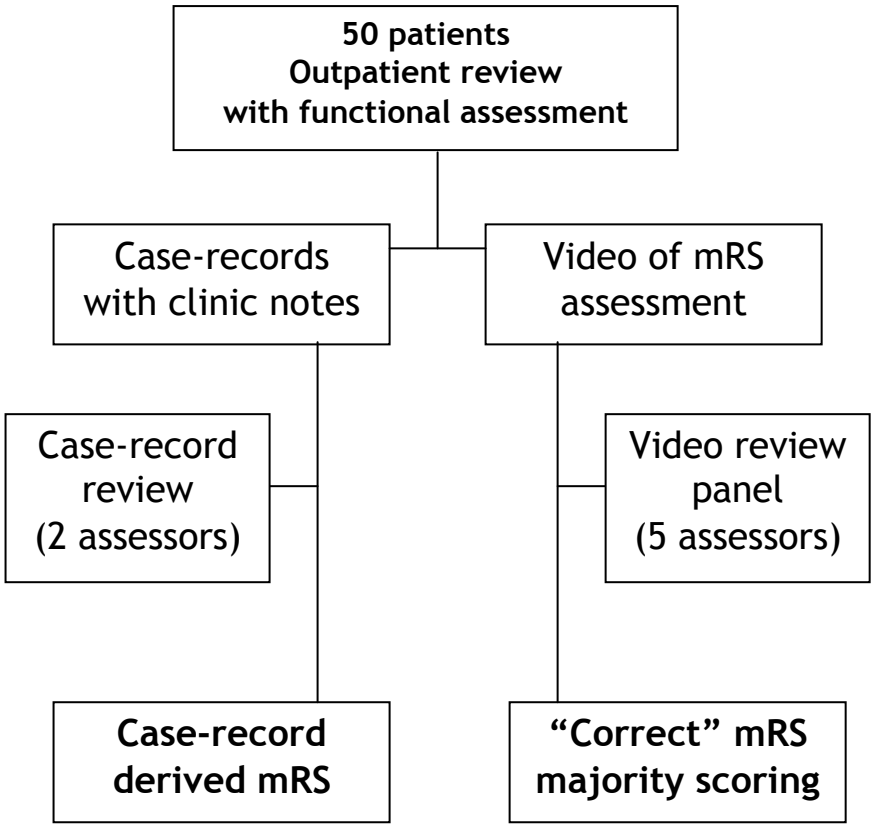
The paper based case records for the study patients were collated. Two independent stroke physicians then derived mRS grades from these case-records, blinded to each others mRS grades. They were given access to complete case-records with no external editing unless explicit reference was made to mRS. In addition to mRS grading, researchers documented degree of confidence in their assessment using a 5 point Likert scale that ranged from 0="not at all confident"

to 5="extremely confident". All clinicians involved were fully trained in mRS assessment, using the previously described digital mRS training package.

For each outpatient full mRS assessment was video recorded, immediately following the routine consultation. These interviews were performed according to the recommendations of the mRS training programme by certified raters half of the interviews followed a structured format as described originally by Wilson.(62) Four stroke physicians and one research nurses later reviewed these video recordings and independently assigned mRS grades, with final "correct" mRS decided by group consensus.(Figure 6) A more detailed description of this video based central mRS adjudication is given in chapter eight.

I calculated agreement between "correct" (ie. group derived, consensus) mRS and derived mRS and the corresponding inter-observer variability using attribute agreement analysis. Accuracy of mRS grading was described by calculating median and IQR of actual mRS for derived mRS grades. I performed all statistical analysis using Minitab software (version 13.1, Minitab Inc, PA, USA).

Figure 6: Schematic diagram of evaluation process for case-record derived mRS versus "correct" mRS.



Results

Fifty patients were selected, median age was 78 years (range: 30-92); median mRS was 2. The group comprised a variety of stroke subtypes (8 TACS; 17 PACS; 6 POCS; 19 LACS). Patients were reviewed at a median of 16 weeks (range 2-56) from index stroke event. One patient withdrew consent after interview and was not included in the final analysis. To ensure there was no recall bias, I excluded four patients where one or both of the case-record reviewers had been involved in their care.

Both reviewers were confident in their grading (median confidence 3, [IQR 3 – 4]; reviewer 1 = 3, [IQR 2 – 3]; reviewer 2 = 4, [IQR 3 – 4]). There was no statistically significant relationship between certainty of derived mRS and proportion of correct grades (p=0.727). Derived mRS showed poor agreement with correct grade (overall *k*=0.34; appraiser 1 *k*=0.35; appraiser 2 *k*=0.31) and between observers (*k*=0.33).(Table 14) Agreement was greatest at extremes of mRS. Case-record reviewers tended to underscore disability (Table 15 and Figure 7).

Table 14: Agreement with "correct" (standard) mRS and agreement between observers for case-record derived mRS.

Variability is described using standard kappa statistics (k).

N/A indicates not applicable

No patients with mRS 5 were included in the study

mRS	Agreement with standard (*k*)	Inter-observer agreement (*k*)
0	0.52	0.49
1	0.24	0.06
2	0.27	0.15
3	0.34	0.48
4	0.28	1.00
5	N/A	N/A
Total	0.33	0.34

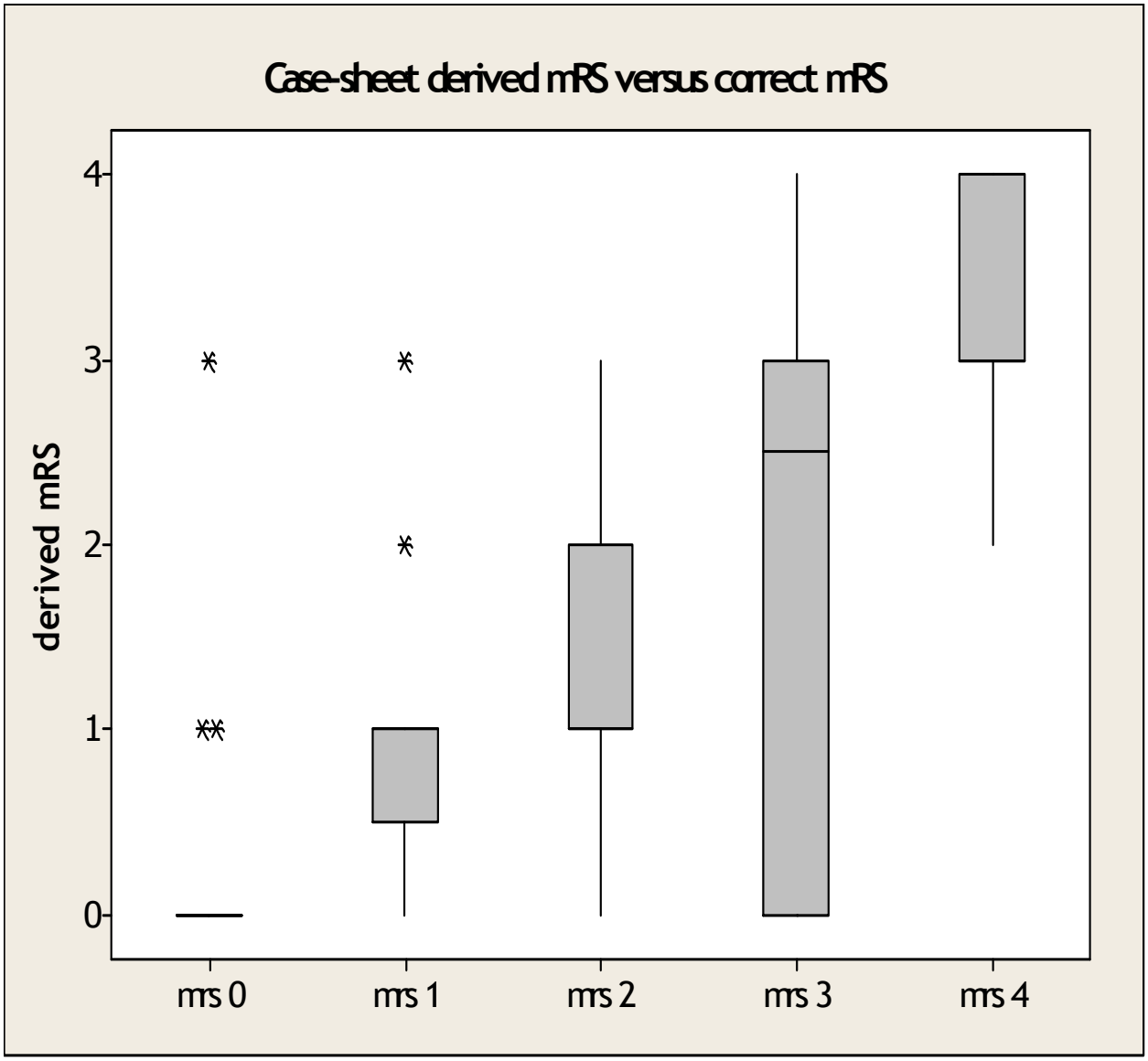
Table 15: Accuracy (median and IQR) for derived mRS versus "correct" mRS.

N/A indicates not applicable

No patients with mRS 5 were included in the study

"Correct" mRS	Combined derived mRS	Reviewer 1 derived mRS	Reviewer 2 derived mRS
0 **n=8**	0 (0 - 0)	0 (0 - 0)	0 (0 - 0)
1 **n=7**	1 (0 - 1)	1 (0 – 1)	1 (1 – 2)
2 **n=17**	1 (1 - 2)	1 (0.5 – 1.5)	2 (1 – 2.5)
3 **n=9**	2.5 (0 - 3)	2 (0 – 3)	3 (0.5 – 3)
4 **n=5**	3 (3 - 4)	3 (2.5 - 4)	4 (3 - 4)
5 **n=0**	N/A	N/A	N/A

Figure 7: Box and whisker plot of accuracy (median and IQR) for derived mRS versus "correct" mRS for both raters.



Discussion

Modified Rankin Scale data derived from patient case-records have unacceptable accuracy and reliability for use in clinical research. This contrasts with other commonly used stroke scales, where quantitative outcome data have been reliably described using qualitative case-record information.(155;156)

Those scales that have been successfully derived from case-record have measured physical impairment only. Transforming bedside neurological examination into a quantitative scale is straightforward if comprehensive physical exam is recorded. As a global disability scale, mRS review requires measures of physical, cognitive, emotional and functional status. Such data may not always be recorded during a busy outpatient assessment.

Although case-record reviewers were poor at deriving mRS, it is interesting that they felt able to derive a score for every patient and were confident in their grading for the majority. This may explain why previous trialists have been happy to use case record abstracted mRS without first testing the validity or reliability of this approach.

My results should be generalisable to other centres involved in stroke trials. Patients were reviewed at approximately 3 months from event, the period when mRS is traditionally assessed. Review was performed by practising stroke physicians trained in use of mRS. Specific pro-formas are not used for data capture and consulting doctors can document as much information as they wish. The department's clinical staff comprises internal medicine physicians with a

stroke interest. It is possible that in the context of a specific rehabilitation, or privately funded service, consultations may be longer with more emphasis on disability. Patients included in the study represented a spread of functional ability. It is noted that no patients with mRS 5 were included in the analysis, this reflects the outpatient setting of the study. As I was interested in reliability of case record derived mRS for use in a clinical trial, omission of the most severely disabled stroke survivors seems reasonable – these patients would be unlikely to be recruited to a trial.

The poor reliability inherent in standard mRS assessment has now been well described. To ensure that the "correct" mRS grading was suitably robust, I used the data generated from the previous study of video based mRS assessment, utilising multiple independent raters with final mRS chosen by consensus.

Although reliability of case record derived mRS was poor, these results must be interpreted in the context of other published literature discussed in Chapter one. From the previous review I described studies reporting an equally concerning poor reliability of traditional mRS using direct interview: k=0.34 for derived mRS; k=0.25 in one study of traditional mRS.(63) I would suggest that improved reliability of mRS assessment is needed across all modalities, however any novel methodologies should have clinimetric properties studied prior to clinical trial use.

Accurate mRS for clinical trial use cannot be derived from standard hospital records. Even amongst those cases where the appraiser was “certain” in their scoring, the proportion of correct grades was no better than chance. Deriving mRS from hospital records or other data sources should not be encouraged - a directed interview remains essential.

Chapter eight

Pilot trial of remote adjudication for modified Rankin Scale assessment in clinical stroke trials

Introduction

I have discussed a number of methods that may assist in reducing the inter-observer variability associated with mRS assessment. The original modifications to Rankin's eponymous scale were in response to a perceived subjectivity in its application.(20) I have discussed the initial experiences of a training resource, although this package should improve mRS application robust evidence of efficacy is yet to be demonstrated. Use of a structure approach has been studied and discussed in this thesis (Chapter two) and other published works, with variable results.(63;64) In essence, studies to date suggest that this approach may reduce but does not completely negate inter-observer variability.

Remote review of clinical trial endpoints, often by central committee, is common in contemporary multi-centre clinical trials.(158;159) Distancing endpoint assessment from the study centre, allows for better "blinding" and less potential for bias.(160) Other potential benefits of central adjudication are numerous and include: greater consistency in grading; opportunity for "quality control" of trial data and experience and diligence of the "expert" assessment panel.(161;162) Thus, addition of a group review approach should result in more robust endpoint data. In stroke trials, remote adjudication of imaging, laboratory and clinical endpoints is well established, while remote assessment of functional outcomes is uncommon.

Sources of variation in mRS include questions asked by the interviewer, the responses provided and the interpretation of these responses by the rater. In a multi-centre international trial involving many hundreds of raters, the grading process may carry greatest scope for variation. Separating mRS assessment from grading with remote review of mRS interview could aid standardisation. Remote grading of functional assessment presents a greater logistic challenge than review of "hard" data such as laboratory results.

A possible solution is to record video footage of the assessment for later "off-line" playback. Such an approach is not without precedent: in the development of a video based mRS training package described in chapter five, incorporation of "live" recordings of patient assessments allowed for mass training and certification. This resource has proven popular and demonstrates that remote review of mRS is feasible. Similarly, a Japanese group used remote assessment of video recorded mRS interview to study efficacy of an mRS questionnaire.(85) Although encouraging, these studies of video based mRS assessment do not necessarily extrapolate to the large scale clinical trial setting, as they incorporated only a limited number of assessments of selected patients and made use of professional video recording facilities. As mRS assessment is principally based on patient interview a less technically demanding option for remote review would be audio recording of the assessment for later playback, although again such an approach has not been formally studied.

Although video based remote mRS assessment seems intuitively attractive, there is a danger in prematurely assuming the clinical utility of what is a novel assessment technique. In this regard, a salutatory lesson can be learned from the practice of retrospectively deriving mRS from patients' case records. As described in chapter seven - this technique became established in trial methodology before any formal clinimetric assessment and has subsequently been shown to have unacceptable reliability. Although trialists are principally concerned with mRS reliability a full clinimetric assessment should also include measures of validity, acceptability and cost-effectiveness.

I hypothesised that remote review of video based mRS assessment by endpoint adjudication committee would be feasible, would demonstrate acceptable clinimetric properties in a clinical trial setting and would be superior to audio only recordings.

Methods

Study participants

Subjects were recruited from the Western Infirmary Cerebrovascular Clinic. Details of the clinic have been described in the preceding chapters. Stroke survivors attending the clinic for routine three-month review were approached and consented between August 2006 and February 2007. All post stroke patients were considered, regardless of demographics; co-morbidity or functional ability. For those patients with language or cognitive impairments, the carer accompanying the patient to clinic was included in the mRS assessment to act as proxy.

I used a stratified, sequential design to test properties of remote video mRS assessment (Figure 8). The basic study design, builds upon the study of inter-observer variability of traditional mRS described in chapter four. All randomisations (assessors; assessment for video recording; cases for audio only assessment) were performed using an online resource (www.random.org). [last accessed January 2010] and allocation was concealed from participants, using an opaque envelope system.

I designed studies to address the following: Inter-observer and intra-observer variability of mRS (reliability); validity of group (consensus) mRS; and comparison of remote review of mRS interviews using full video playback or audio only. Separate components of the study will be described in turn. The local ethics committee approved the study protocol and arrangements for use and storage of patient data.

Analysis of reliability

Seven assessors comprising four stroke physicians and three research nurses performed the initial mRS assessments. All assessors were trained and certified in mRS assessment and had considerable experience of use of the scale in clinical trials.

Patients were approached following routine clinic review. Consenting patients had an independent mRS assessment performed and captured on video by a randomly selected researcher from the pool of seven. Following interview, mRS grading was recorded and sealed in an opaque envelope. The researcher conducting the interview also managed the video recording process. No formal training in use of video recording equipment was given.

Three months later, all seven researchers reviewed the complete set of video mRS recordings, including their own interviews. They were asked to grade mRS based on the recorded assessments. Video review was performed independently and blinded both to original mRS grading and to colleagues' grades. A delay of three months was chosen to minimise recall bias while allowing timely completion of the study. These individual video based mRS grades were used to describe inter-observer variability within the group. This inter-observer variability was compared with data from previous study of traditional mRS conducted by the same research team and in the same unit.

After a further three months, a group of 5 researchers from the original cohort (4 physicians; 1 study nurse) reviewed and graded the video recordings for a second time. As before, video based mRS grading was performed independently and blinded to previous and colleagues' scores. These second grades were compared with original video mRS grades to assess intra-observer variability.

To maintain consistency with my own and others group studies, kappa (*k*) statistics were used to describe agreement, using accepted definitions of poor (*k* = 0.00-0.20); fair (*k* = 0.21 – 0.40); moderate (*k* = 0.41 – 0.60); good (*k* = 0.61 – 0.80) and very good (*k*= 0.81 – 1.00) agreement.(34) Traditionally, *k* values of > 0.61 are taken as acceptable for a clinical test.(32) The number of interviews in which raters agreed exactly on mRS, was calculated and expressed as percentage agreement. Analysis was performed using Minitab software (version 14.0, Minitab Inc, PA, USA).

Group (consensus) review

At a final group viewing of the mRS videos, the panel's 5 researchers were allowed to discuss their grading with mRS scoring by consensus. The group were not obliged to score and in cases where no consensus could be reached, this was recorded. The group was also asked to grade quality of the video recording using free text and document any technical faults. As many of my proposed measures of validity require a single scale, consensus mRS results were used for this purpose.

To assess convergent validity, agreement between video mRS and original face to face mRS interview score was described using attribute agreement analysis. I further compared individual video mRS scores to the consensus score. For individual raters, proportions of original mRS grades that agree with final consensus were calculated and compared using chi-square analysis.

As part of the group exercise, randomly selected video interviews were first presented as audio only playback. The group graded these audio cases and then watched the original complete video and re-graded as necessary. Variability within the group's audio mRS scores and variability between audio mRS grading and full video mRS were described using attribute agreement analysis.

Technical specifications

In consultation with the Media Services Department, University of Glasgow I developed a suitable recording system for clinical trial use. Final choice of hardware for video recording was: HDVR-HC1E 1080i digital HD camera recorder (Sony, Japan) and ATR97 omni-directional condenser boundary microphone (Audio-technica, Ohio, USA). Digital recordings were transferred to video-disc using Windows Movie Maker (Microsoft, Washington USA). The accompanying Windows Media Player application (Microsoft, Washington USA) was used for playback during video review.

Results

Of 104 patients approached, 102 consented to video recording of mRS interview with 1 patient subsequently withdrawing consent following the recording. Full demographics of the study population have been described in chapter four.

Of completed recordings, 99 were of sufficient technical quality to allow assessment, 7 videos had technical problems that did not preclude mRS grading (3 poor sound quality; 4 poor visual quality). Mean duration of recorded mRS assessment was 4.9 (SD 2.4) minutes. After watching 99 videos suitable for review, the group was able to reach a consensus score in 96. Based on group consensus, final distribution of disability was: mRS 0=14 patients; mRS 1=25; mRS 2=23; mRS 3=20 mRS 4=11; mRS 5=3.

Reliability of remote video review of mRS

Inter-observer variability for remote review of video mRS was good on first and second viewing (*k1*=0.67, 47% matched (95% CI: 36.1 – 57.5); *k2*=0.66 44% matched (95%CI:33.4-54.3)), variation was least at extremes of grading (Table 16, Figure 9). Reliability of remote video mRS assessment was favourable compared to my previous estimate of mRS variation using traditional face to face interview (k=0.57; 67% matched (95% CI: 56.6 – 75.7)). Differences in trial methodology do not allow for direct comparative analysis.

Intra-observer variability was good (*k*=0.64, 28% matched (95% CI: 14.0-37.9) again with least variation at extremes of the scale. (Table 16)

Validity of remote video review of mRS

Group review of video mRS showed moderate agreement with traditional face to face mRS *k*=0.57 (65 matched 95% CI: 57.6 – 72.0)). Agreement with individual remote video scores was good *k*=0.77 (47 matched 95% CI: 36.1-57.6) (Table 17). Differences between consensus and individual scores could not be attributed to a single rater systematically scoring differently to the group, as percentage agreement with consensus was similar across the group (range 70% - 76% agreement; p=0.651).

Audio only mRS

Forty one cases were randomly selected (again using the online random number generator previously described) for analysis of audio only mRS assessment and mRS was derived for 39. All raters agreed that the remaining 2 cases could not be assessed without corresponding visual information. There is no method to incorporate such missing data within traditional kappa statistics and as such these cases were excluded from analysis, with the resulting statistics likely over estimating reliability of an audio only approach. Accepting this limitation, reliability of audio only mRS review was *k*=0.62 (31% matched (95%CI: 17.0 – 41.0)). Comparing responses for audio only and the equivalent full video review, agreement between scores was *k*=0.67 (matched 26%, 95% CI: 20.1 – 30.0).(Table 18)

Table 16: Inter-observer and intra-observer variability for video based modified Rankin Scale (mRS) assessment.

Variability is described using standard kappa (k) statistics.

mRS	Inter-observer variability (*k*) first video review	Inter-observer variability (*k*) second video review	Intra-observer variability (*k*) video review
0	0.83	0.80	0.80
1	0.64	0.60	0.60
2	0.52	0.52	0.53
3	0.65	0.65	0.60
4	0.78	0.77	0.72
5	0.94	0.94	0.94
overall	0.67	0.65	0.64

Table 17: Variability comparing group consensus modified Rankin Scale (mRS) to traditional mRS assessment and individual video assessment.

Variability is described using standard kappa (k) statistics.

mRS	Variability (*k*) consensus v. traditional mRS	Variability (*k*) consensus v. individual review
0	**0.75**	**0.88**
1	**0.52**	**0.75**
2	**0.58**	**0.64**
3	**0.58**	**0.73**
4	**0.71**	**0.83**
5	**0.83**	**0.97**
Overall	**0.63**	**0.77**

Table 18: Reliability of modified Rankin Scale (mRS) across a number of modalities.

Agreement is measured by kappa (k) and quadratic weighted kappa (kw) statistic, presented as k (95% confidence interval).

For reference k=0.61 is considered acceptable for a clinical scale.

** Is agreement across the group (greater than 2 raters compared)*

mRS modality	Variability (*k*)	Weighted (*kw*)	Agreement (%)
Standard Face to Face	0.57 (0.46 – 0.69)	0.83 (0.63 – 1.0)	67%
Video mRS (first viewing)	0.68 (0.64 – 0.72)	0.92 (0.85 – 0.98)	47%*
Video mRS (second viewing)	0.66 (0.63 – 0.70)	0.91 (0.85 – 0.98)	44%*
Video mRS (intra-observer)	0.67 (0.61 – 0.71)	0.89 (0.80 – 0.98)	75%
Audio mRS	0.67 (0.61 – 0.73)	0.88 (0.78 – 0.96)	31%*

Discussion

Variability in mRS grading limits its utility as a clinical trial endpoint. Measures to improve reliability have been developed but no perfect solution has been described. I developed a novel video based mRS assessment for use in clinical trials and have demonstrated acceptable clinimetric properties of this approach.

As has been discussed, variability in assessment implies a degree of misclassification of endpoints. Removing this inherent "noise" from the data should decrease the likelihood of a type II error and ultimately improve statistical power.(39) Attempts to standardise mRS assessment by creating and distributing training resources for stroke researchers have already been discussed. In this pilot I attempted a converse but complementary strategy – limiting mRS assessment to a selected group of individuals.

Analysis of inter and intra-observer variability of remote mRS was encouraging. Variability in video mRS assessment was equivalent or less than that seen locally for traditional face to face mRS assessment. The apparent improvement in variability was welcome but unexpected and demands some consideration. It is well recognised that subjects perform a task differently if they know they are being observed – the "Hawthorne effect".(163) The good reliability seen with video review may in part be related to interviewers conducting a more complete assessment "for the camera". If this effect can be sustained and video assessments remain rigorous, I would expect increasing use of video based technologies to improve overall standards in mRS interview. The objectivity

afforded by review of interview, distant from time constraints and distractions of clinical work, with the ability to review the consultation repeatedly until happy with a final grading may also have impacted on reliability.

Although my prime concern was mRS reliability, for this novel assessment methodology it was important to test as many clinimetric properties as possible. There is no standard methodology for describing validity.(145) In the absence of a recognised gold standard, I assessed convergent validity using direct comparison of the scale to other measures of the outcome of interest. My results suggest that remote group review of video mRS is a valid measure of post stroke disability.

There are other clinimetric properties that can be assessed. Responsiveness, the ability to detect meaningful change over time, was not assessed. As a single measurement of mRS is usually analysed as a trial endpoint, most often at 90 days post ictus, I felt that analysis of responsiveness was of lesser importance than other properties. A literature describing responsiveness of stroke outcomes scales is available.(52) I did not formally measure acceptability or cost-effectiveness but can extrapolate from the data collated. Only three (2.9 %) patients refused video recording of mRS, a good recruitment rate for any intervention and certainly suggestive of acceptability; in a trial, patients consent to outcome assessment at the outset. The audio-visual equipment used in the trial was deliberately chosen to be economical and I made use of widely available computer software. When compared to the costs reimbursed to a centre for each patient recruited to an industry sponsored trial, the initial expense in setting up such a system is minimal.

My sub-study of remote mRS assessment based on audio playback suggests that this approach may not be suitable for clinical trial use. I could find no previous study utilising an audio only approach; the closest equivalents have concerned "live" telephone based mRS.(64;67;70) Acceptable reliability of telephone based assessment has been demonstrated for scales measuring Glasgow Outcome Score (164) and cognitive impairment(68). Studies of telephone based disability assessment, particularly mRS have yielded conflicting results. Independent groups have demonstrated poor reliability (k=0.30)(64) and good reliability (k=0.74)(70), although results of this later analysis are likely overestimated as single dichotomised mRS was used. In summary it seems variability is increased by use of telephone based mRS assessment and as such my finding of poor reliability of audio only feedback is in keeping with the available literature.

As well as the potential to improve trial data quality, there are numerous other possible benefits to video recording of mRS assessment. The video format allows for repeated review and discussion of challenging cases. Following a difficult interview the assessor could review their interview and re-grade if necessary. Feedback from central reviews may help educate colleagues less experienced in mRS application and raise overall standards. Finally, preserving a recording of interview facilitates quality control and provides another barrier to research fraud. Although this pilot was concerned only with mRS there is no reason that my approach could not be applied to other disability assessment scales in stroke or other areas of neurology.

My aim was to develop a system for future use in a multi-centre trial setting. A possible option would have been to "add" video mRS assessment to an ongoing clinical trial. To avoid "contamination" of results with potential treatment effects I opted for the single pilot study of remote mRS described. Patient selection criteria were deliberately inclusive. I included a large representative cohort of stroke patients including patients for whom assessment of disability may be challenging. The low baseline stroke severity of this cohort represents the outpatient setting of the study and is representative of a clinical trial population. I selected a panel of assessors from different clinical backgrounds as previous work has suggested that profession and training can impact on reliability of outcomes assessment.(132)

Previous reports describing development of "tele-stroke" systems emphasise the importance of high definition audio and visual recording to ensure reliability.(165) The ideal would be use of professional recording studio facilities to achieve optimal audio-visual capture. Such an approach is not practical in a clinical setting. Thus in choosing audio-visual recording equipment for video mRS I needed a compromise between hardware that was portable, allowing for use in clinic or at bedside and a system that provided high fidelity recording. Critics may argue that with two videos unable to be graded because of technical issues and a further seven videos possibly compromised, I have not demonstrated technical feasibility. However, no formal training in use of hardware was given prior to video recording and the majority of technical errors were made early in the study. As such I would expect a much lower rate of technical problems if full training and support were offered. In a trial, even if some video recordings were to be found useless and time did not allow re-

assessment, the original rater's contemporaneous grading could stand as a substitute for analysis.

I acknowledge the preliminary nature of my study of remote mRS assessment. There were certain weaknesses in methodology that could be addressed in future work. My study was based in a single centre, using a team of researchers who have worked together for many years. All the researchers involved in this work and the majority of patients speak English as a first language. Differences in culture, language and health care systems may compromise reliability if remote assessment is performed across multiple international centres as would be required for a large scale clinical trial. The limited number of local researchers experienced in mRS assessment precluded creation of a separate review panel not involved in the original mRS assessments. I took measures to limit the effect of recall bias but for future study of remote mRS assessment an independent review body would be a prerequisite.

In conclusion I have demonstrated feasibility and acceptable properties of remote video based assessment of mRS. These pilot data now need prospective confirmation, along with more detailed analyses of feasibility; economics and data safety in a multi-centre clinical trial setting.

Figure 8: Schematic of remote modified Rankin Scale (mRS) assessment methodology.

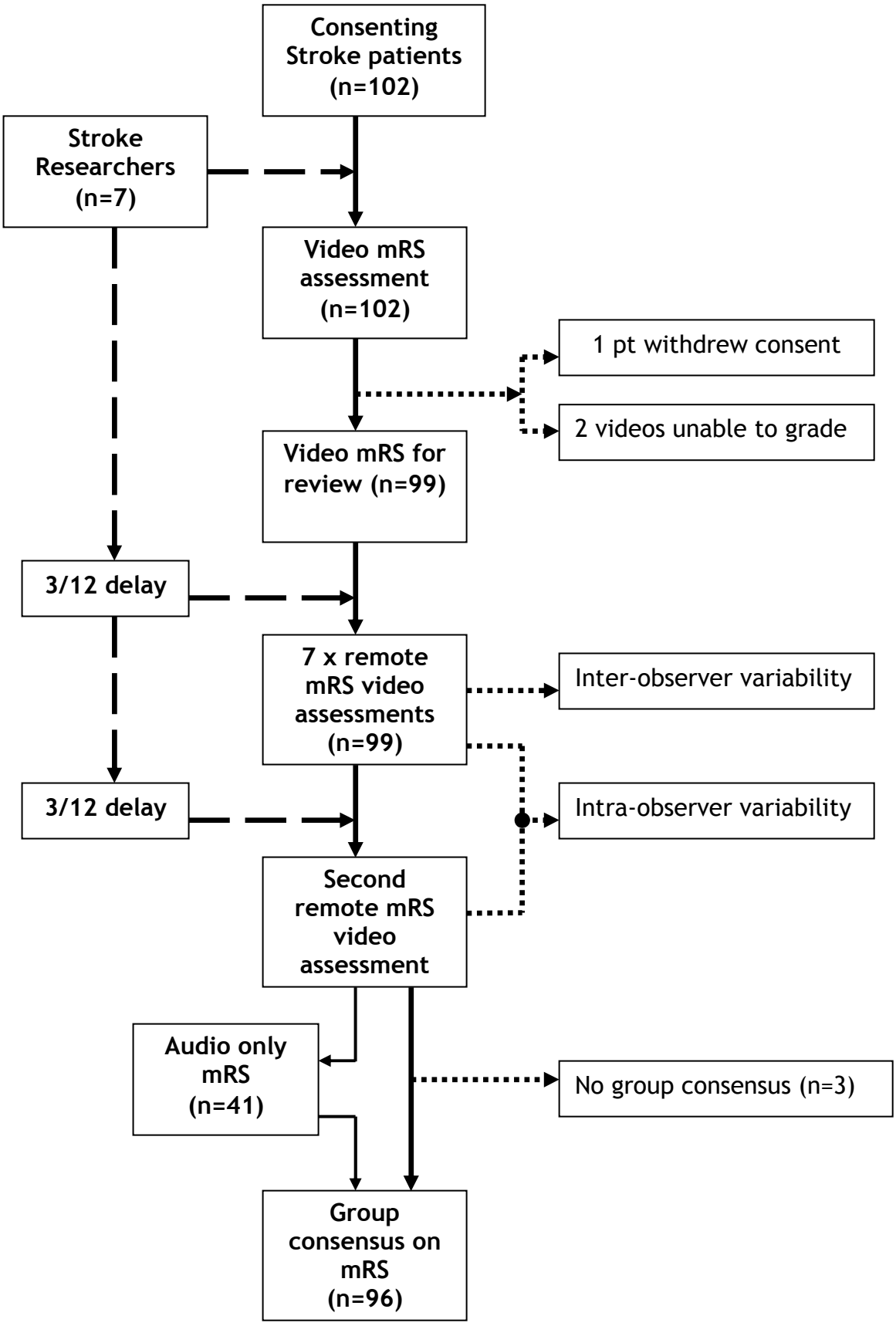


Figure 9: Spread of video mRS scores for differing mRS grades.

mRS	0	1	2	3	4	5
0	102	11	9			
1	12	152	33			
2	3	41	144	26	2	
3			32	95	9	
4				21	71	4
5						27

Chapter nine

Time spent at home post stroke

“Home-time” – a meaningful and robust outcome measure for stroke trials

Introduction

The thesis has demonstrated the need for robust assessment of post-stroke recovery in both clinical and trial work. Of the many tools that exist to quantify functional outcome, I have chosen to focus on the mRS as this is the most frequently used functional outcome scale and has been used in a number of landmark stroke trials.(120;151) Despite its popularity with trialists, mRS has limitations – in particular I have now demonstrated potential for substantial inter-observer variation. Although I have outlined certain approaches to improve reliability, training in administration; use of a structured interview, no single approach has negated the effect of inter-observer variability.

The ideal outcome measure would be simple to understand and apply with acceptable validity, variability, and responsiveness. To date, no stroke outcome measure adequately meets these criteria. Creation of a novel instrument is unlikely to achieve immediate widespread acceptance.

An alternative is to derive surrogate outcome measures from routinely collected patient data. For example, duration of inpatient stay lends itself to health economic analysis because inpatient days account for much of the expenditure associated with stroke. A relationship between mRS and hospital bed occupancy has been demonstrated.(56) However, inpatient stay is not an ideal measure of post stroke functional outcome. Despite significant advances in acute stroke treatment there remains substantial mortality, rates within the first few weeks

remain high and in survivors many require ongoing care in Nursing home or equivalent.(166) Early mortality and transfer to a long-term care setting after severe stroke will each be associated with shorter stay and thus potentially skew outcomes data based on hospital “bed days”.(167)

I hypothesized that duration living independently in the community could serve as an appropriate outcome measure less likely to be confounded by the survival and transfer issues discussed. To explore this hypothesis, I measured duration of stay in the patient’s own home following stroke event —"home-time"— using data from a comprehensive, prospectively gathered stroke outcomes trial.

Methods

I analyzed data that had previously been extracted from the original records of the GAIN International trial conducted in 1998 to 1999 and first reported in 2000.(168) GAIN was a multi-centre randomized, double-blind, placebo-controlled trial of the putative neuro-protectant "gavestinel" administered in acute stroke. Patients were aged greater than 18 years old with symptoms of acute stroke, including limb weakness. Prior to stroke event, all were previously independent and were randomized and treated within 6 hours of onset of stroke. All original data extraction and preparation for statistical analysis were performed by two researchers (Tau-Pin Chang and Jennifer S. Lees).

The GAIN trialists collected outcomes data across a number of functional domains: mRS, BI, and NIHSS. Functional outcomes scores were collated at 1 month and 3 month intervals from recruitment. Outcomes on mRS were assessed by local observers according to a standard scoring system. No formal training or certification was offered. To assist consistency, observers were offered advice on scoring of BI in the form of a video based demonstration but no certification procedure was in place. All observers had been trained and were certified in use of the NIHSS scoring system.

Resource use data were recorded at 90 day follow-up visit based on interviews with patients or a proxy (relative / caregiver) these data were supplemented by review of hospital records. In particular, detailed data regarding duration of hospitalization or time spent in nursing facilities were gathered. Data on non-hospital placement post discharge were also collected and categorised using the

following six labels: own home, relative's home, intermediate-care facility, nursing/convalescence home, rehabilitation facility, or undefined.

For the purposes of this analysis, all high-dependency and medical bed-days were grouped together and considered separately from days spent in nursing or institutional residential care. Days spent in the patient's own home were grouped together with time spent in a relative's home, these data were used to calculate "home-time."

Only those outcomes actually recorded at 90 +/- 17 days were used in this analysis; missing data were not imputed. Resource use was censored at 90 days. When final follow up occurred earlier than 83 days in a patient who survived past 90 days, last known placement was extrapolated to 90 days; patients in whom dates of placement were not recorded were excluded from analysis.

For this preliminary analysis, I used one-way analysis of variance (ANOVA) to assess home-time trends for mRS, NIHSS, and BI comparing adjacent categories by Bonferroni testing. All analyses were performed using StatsDirect statistical software version 2.4.5 (StatsDirect Ltd, Cheshire, UK).

Results

Full outcome data were available for 1717 of the 1788 intention to treat patients. Data were incomplete for 15 patients and 56 withdrew from the study before 90 days. Mean age was 69.7 (S.D 12.2) years, 737 (42.9%) were female, 321 (18.7%) had intracranial haemorrhages as the index event, and mean admission NIHSS score was 13.1 (S.D 6.2).

Mean time in the hospital was 28 days and mean home-time was 31 days. Home-time was significantly associated with changes across mRS (P<0.0001), NIHSS (P<0.0001), and BI (P<0.0001). On analysis of between-category differences, home-time was significantly associated with change across all mRS categories except mRS 4 to 5. (Figure 10, Table 19)

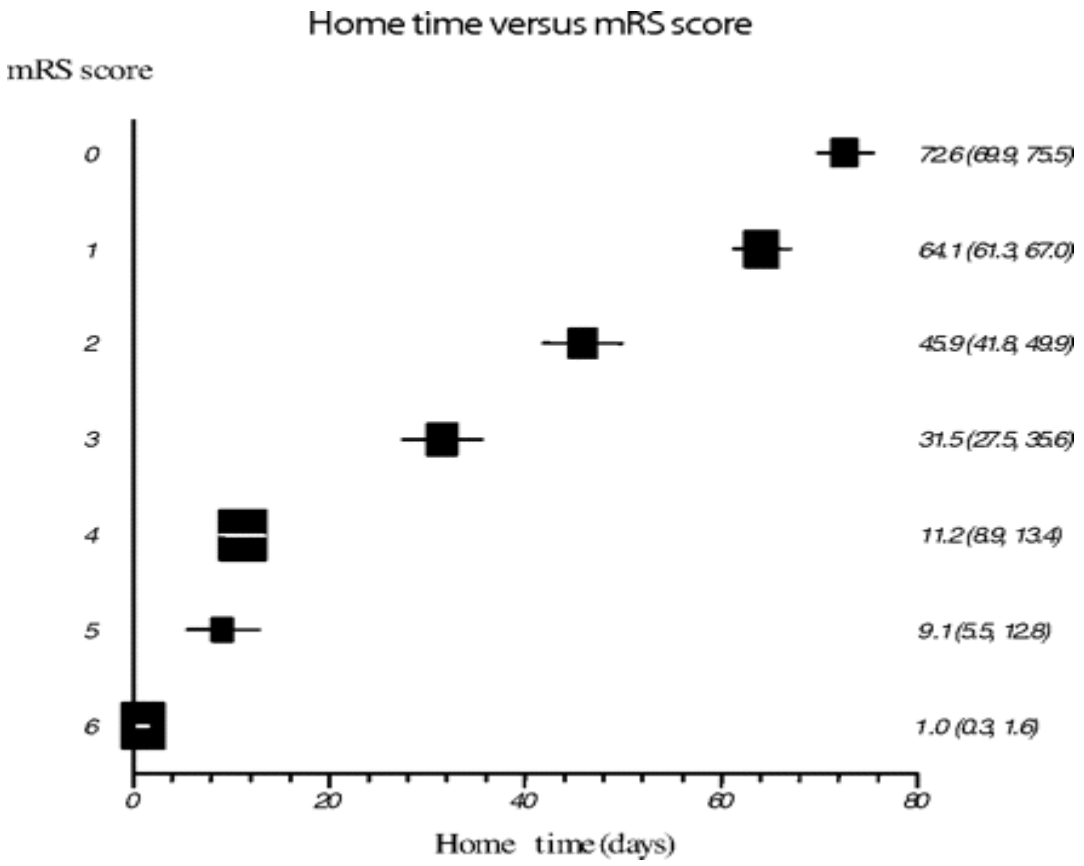


Figure 10: Mean 90-day home-time ±95 CI versus mRS.

P<0.0001 comparing adjacent categories except

mRS 4 to 5 (P=0.37)

and mRS 5 to 6 (P=0.0003).

Table 19: Relationship between Home-time and mRS.

	mRs 0 **197**	mRs 1 **268**	mRs 2 **205**	mRs 3 **214**	mRs 4 **366**	mRs 5 **143**	mRs 6 **324**
Home-time (mean)	72.6 -	64.1 *	45.9 *	31.5 *	11.2 *	9.1 P=0.37	1.0 p=0.0003
95% CI	69.9 – 75.5	61.3 – 67.0	41.8 – 49.9	27.5 – 35.6	8.9 – 13.4	5.5 – 12.8	0.3 – 1.6

** P<0.0001 comparing preceding column*

Analysis of NIHSS and home-time revealed a significant association with NIHSS categories 0 to 1 (*P*<0.0001), 1 to 2 (*P*<0.0017), 2 to 3 (*P*<0.0013), and 4 to 5 (*P*<0.0001). Home-time was significantly associated with change across BI categories 100 to 95 (*P*<0.0001) and 95 to 90 (*P*<0.0001). Change across all other categories of NIHSS and BI were non-significant. (Figures 11 and 12)

Discussion

I have shown that home-time has a significant association with post stroke disability as measured by mRS, particularly across the better recovery levels. Although intuitive, this relationship has not been demonstrated previously. I have already established that mRS is the preferred outcome measure in acute stroke trials and my demonstration of the relationship between home-time and mRS outcomes, provides strong evidence of the validity and potential utility of home-time as a novel trial endpoint.

Home-time has potential advantages over mRS in application and interpretation. Given the problems of inter-observer variability associated with mRS, an objective measure such as home-time should give better reliability and would not require the formal training that has been described for mRS. "Perfect" reliability would not be possible, even for a measure such as home-time. Historical data regarding discharge, transfer to rehabilitation etc was first collated from patient or proxy and such is prone to recall bias. Using hospital records to check these dates should help remove some inaccuracy. The continuous nature of home-time data lends itself to more powerful statistical techniques than traditional dichotomized or ordinal outcome measures.(136) Home-time is generalizable and although I have used stroke trial data in this analysis it could be applied to any disabling condition.

A further strength of home-time is its immediacy. Discussion of possible treatment benefit is essential for informed consent. The abstract outcome measures used in trials make this already challenging task more difficult. Some centres are piloting the use of patient videos to illustrate mRS grades to patients and carers in an effort to improve informed consent for cerebrovascular intervention (personal communication, University of Heidelberg). Home-time offers an outcome measure that should be easily understood by the lay public and other medical professionals.

Association of home-time with NIHSS and BI was less convincing. I do not interpret this as a failing of home-time; rather, home-time accentuates the limitations of NIHSS and BI. "Floor and ceiling" effects of the BI are well recognized and were discussed in chapter one.(51) NIHSS measures physical impairment, it takes no account of ability to compensate for functional deficit. At ninety days post event, it is likely to be responsive to change only at extremes of outcome.(169) In support of this I have shown that home-time change across grades was significant only at extremes of the scales that denote almost complete recovery.

Patients in the GAIN study were broadly representative of a clinical trial population. However, all patients in GAIN were independent at baseline. Home-time may be less valid as an outcome measure if applied to a more disabled population such as seen in routine clinical practice. For clinical trials, pre-morbid residence at home could be an objective entry criterion.

Rehabilitation is often necessary post stroke, reducing home-time in the short term to achieve longer-term improved outcomes.(170) As such, a 90-day cut off is most appropriate for home-time analysis. For the majority of stroke survivors, it is unlikely that meaningful inpatient rehabilitation will continue past 90 days. Ninety-day outcome assessment has become standard in clinical trials and so ascertainment of home-time would not necessitate changes to study protocols.

I do not claim home-time to be the perfect measure of outcome; it is prone to many of the same limitations as other accepted outcome scales. By measuring home-time at specific cut offs, data may be biased by "early" and "late" responders, those patients whose recovery time from disabling stroke is substantially longer or shorter than average. Although important at the individual patient level, such influences are less important in the context of large multi-centre trials and it is in this area that I propose the use of the home-time instrument.

It is assumed that increasing home-time is a positive outcome because return home is desired by most patients and will reduce total costs. Home-time makes no measure of level of care required to facilitate discharge; a large package of supplemental carers and home modifications may allow return home but at substantial economic expense. Potential for provision of care by state or family will vary between countries, this could potentially bias home-time data if recorded in a number of international centres as is the norm for a large scale clinical trial. The data available from the GAIN trial did not allow for meaningful sub-analysis of home-time by country. However, the influence of country and culture is not unique to home-time; significant differences across

countries are also seen for the disability measures of mRS and BI in the GAIN data set.(154) Other socio-demographic factors external to the patient such as marital status, dependents, and healthcare insurance may influence home-time, but such data are not routinely collected and so appropriate analysis could not be performed.

Despite its plausibility, I recognize that the potential utility of home-time needs to be confirmed. I have derived home-time retrospectively from previous trial data and thus can make no assessment of its use in real-time clinical practice. However, GAIN was a typical, multi-centre, intention-to-treat trial and, as such, I assume that the findings would hold for future trials. Rather than replace established instruments, home-time could complement other trial end points. Integration with other scales could generate a powerful global outcome statistic.(82) It would be of interest to examine home-time in existing trial data sets of thrombolysis or haemostasis.

In summary, home-time assessment offers robust, objective, easily communicated information on stroke outcomes. Trialists are encouraged to measure home-time and consider its inclusion as a clinical end point.

Figure 11: Median “Home-time” versus Barthel Index

as measured at ninety days post stroke event

Data are presented graphically as :

Median (diamond)

IQR (rectangle)

Range (line)

Figure 12: Median “Home-time” versus National Institutes of Health Stroke Scale

as measured at ninety days post stroke event

Data are presented graphically as :

Median (diamond)

IQR (rectangle)

Range (line)

Note small numbers of patients at higher NIHSS levels.

Figure 11: Median 90-day home-time versus Barthel Index

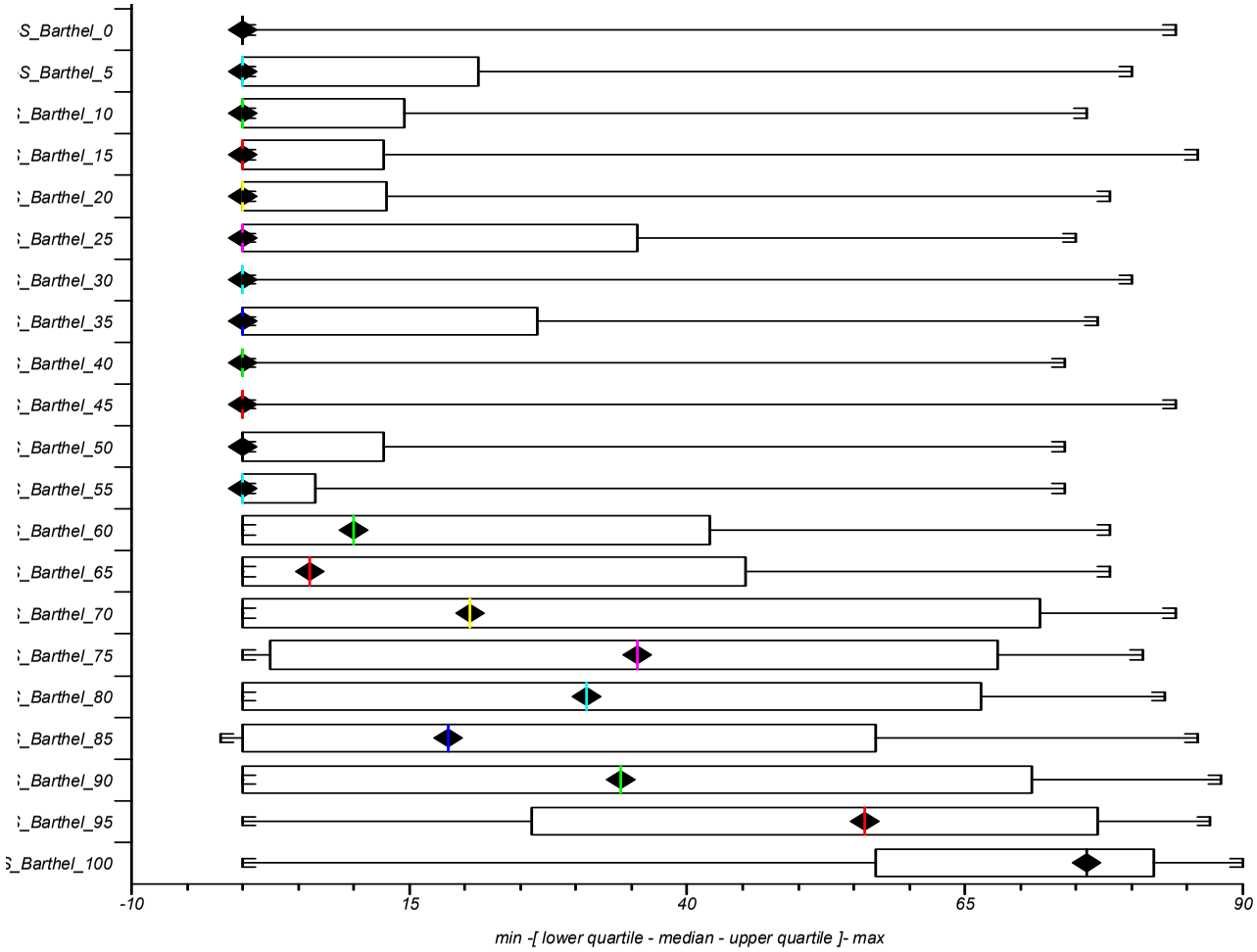
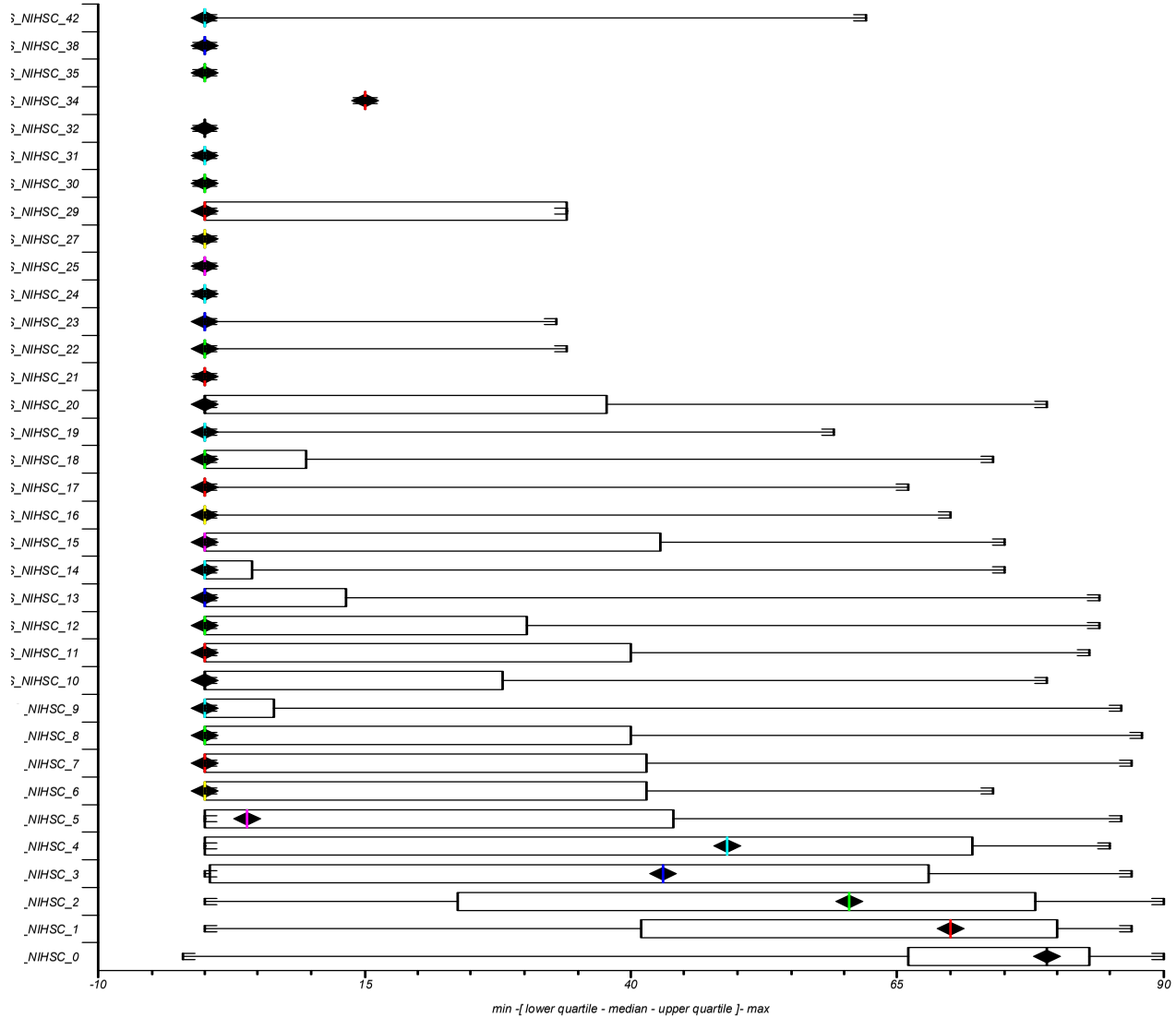


Figure 12: Median 90-day home-time versus NIHSS



Chapter ten

Dr John Rankin; his life, legacy and the 50th anniversary of the Rankin stroke scale – a historical review

Introduction

Dr John Rankin (1923-1981) is one of the many distinguished alumni of the former University Department of Materia Medica and Therapeutics, Stobhill Hospital Glasgow. While his varied international career encompassed pulmonary physiology, occupational medicine and public health, he remains best remembered in the UK for his early stroke publications. In a series of articles published 50 years ago in the *Scottish Medical Journal* he described early rehabilitative stroke medicine using a novel grading system.

Half a century on and Rankin's eponymous stroke scale has become the endpoint of choice in acute stroke trials. This thesis has concentrated on current and future use of the modified Rankin scale. To put this work in a historical context, this chapter describes Rankin's remarkable career and the legacy of his work, with a particular focus on his stroke research and grading systems. The historical data presented are gathered from Rankin's published research manuscripts and editorials; newspaper articles and other papers from University of Glasgow and University of Wisconsin, Madison.

Rankin and the University of Glasgow

John Rankin was born in Glasgow in 1923 into an academic family, his father being a noted Professor of Physics. Rankin began his own academic studies at the Medical Faculty of the University where he achieved several prizes and excelled in pathology. Successfully completing his medical degree, he was awarded the prestigious Rockerfeller scholarship to pursue post-graduate study at University of Wisconsin, Madison. He achieved his MD in 1947.(171) During his year in Madison Rankin forged strong transatlantic research links, which remained prominent throughout the rest of his career. Rankin was always thankful for the opportunities afforded by his scholarship and it is fitting that University of Wisconsin now offer the "John Rankin Travel Award" to facilitate international medical research.

Rankin left Madison in 1948, completing training in London before returning to his native Glasgow to work within the University Department of "Materia Medica and Therapeutics", Stobhill Hospital. Stobhill was unusual for its time, having an established academic department within a municipal hospital that predominantly cared for older and chronically unwell patients.(172) Working in this environment of scholarly research and real-life clinical medicine clearly influenced the young Rankin. When he eventually led his own department of medicine, he vociferously opposed the traditional culture of elitism that separated academia from routine patient care.

Stobhill was a general hospital providing a range of medical and surgical services. Originally designed to house the large number of patients with

"encephalitis lethargica" following the first world war, a number of beds remained specifically set aside for "chronic sick" patients. During Rankin's time at Stobhill these beds mostly comprised patients with rheumatic heart disease or stroke.(173) The combination of an active academic department, a young physician with innovative research ideas and a cohort of long term stroke inpatients was potent. Despite serving only three years on the staff at Stobhill, Rankin collected unique and unparalleled observational data on cerebrovascular diseases.

In a period where active intervention for stroke was uncommon and where therapies that were employed (such as barbiturate coma) often contributed to mortality, Rankin's optimistic attitude to stroke was highly unusual. He argued that positive results could be achieved through rehabilitation, time and encouragement, and that there was no place for the therapeutic nihilism exhibited by his peers.(174) Rankin's belief in early mobilisation was strongly influenced by primate work simultaneously being carried out at University of Wisconsin by Travis and Woolsey.(175) Together the three researchers developed theories of the brains ability to regain function following insult - effectively an early model of neural plasticity.(92)

Within the chronic sick beds of Stobhill, Rankin effectively created a prototypic stroke unit. It is certainly true that Rankin's ideals of multidisciplinary working and early rehabilitation remain core principles of contemporary stroke care.(176) We can only assume that other faculty staff shared Rankin's views on the value of rehabilitative services. Of the eight doctors who comprised the

Department of Materia Medica in the late 1940s, three went on to achieve chairs in the nascent speciality of Geriatric Medicine.

Rankin's early stroke manuscripts gave an indication of the themes that would come to characterise his future career: a belief in providing a scientific evidence base to clinical work; recognition of the social aspects of disease and the need for preventative rather than curative medicine. Although remembered for his scholarly achievement at Stobhill, Rankin remained a practical physician, popular with his patients and respected by his junior staff. It is telling that Rankin's final manuscript on stroke described a bespoke apparatus for prevention of drop foot in the paralysed limb.(177) However, in some less enlightened passages he did comment that only poor outcomes could be expected in "artisans" or "women at an age when the call of the family or the home no longer exist."(92)

Figure 13: Picture of Dr John Rankin during his time at Stobhill Hospital, Glasgow (circa 1951).

Figure 14: Department of Materia Medica Stobhill Hospital, Glasgow.

Rankin in Madison

It has been commented in neurology texts, that following his time at Stobhill, Rankin's subsequent career passed unnoticed.(157) It is true that Rankin did not progress his original cerebrovascular work, rather he took an unconventional career "side-step" - returning to University of Wisconsin in 1953 as an undergraduate physiology instructor and eventually pursuing post-doctoral research in pulmonary alveolar diffusion. The University of Wisconsin was founded on principles of equality of access and multidisciplinary delivery of education, and as such it seems appropriate that the idealistic Rankin remained in Madison for the remainder of his life.

The reasons for Rankin's return to Madison and the basic sciences are not clear. It has been suggested that frustration at treating what he saw as a preventable disease, encouraged Rankin to forsake Stobhill and stroke medicine. Primary prevention of morbidity was the driving force for the rest of his distinguished career. Combing his knowledge of pulmonary physiology, clinical medicine and his passion for public health he conducted large scale, long-term field research. Along with researcher Helen Dickey he was instrumental in defining the pathogenesis of the farmers lung type of hypersensitivity pneumonitis.(178) In later work he described other industrial lung diseases(179) and alerted the public to the possible environmental and health effects of the industrial rubber industry.(180)

During his 23 years at the University of Wisconsin medical school Rankin rose through the ranks to achieve chairs in Medicine (1964); Preventative Medicine (1968) and become chief of Pulmonary Medicine (1970). In recognition of the part that the state played in the health of the nation, Rankin served on an increasing number of governmental organizations – over 112 in his lifetime. In spite of the considerable time he gave to administrative and state matters Rankin remained a prolific researcher and devoted teacher. He published over 70 scientific papers, 11 book chapters and 40 abstracts, many posthumously. The most tangible aspect of Rankin's legacy at University Wisconsin is the small-scale physiology laboratory he developed into a world class research centre – now the John Rankin Pulmonary Medicine Laboratory. One can only guess as to what he may have achieved had he stayed in Glasgow and pursued the study of stroke.

Rankin died in his home aged 57, having taken his own life. It was speculated in the press at the time that increasing state driven budget cuts to his research program precipitated this tragic event. If this is the case, it is grimly ironic for a man who gave so much of his time to governmental committees and who passionately believed that medicine and state should work together. Rankin himself had prophetically commented, "whenever you have a conflict between economics and health, health loses out". Tributes were many; from patients, students, professional colleagues and friends. The University of Wisconsin gave a fitting epitaph when in obituary they described him as "the near-ideal model of scholarship, service and humanitarianism".(181)

Rankin, Stroke Medicine and Development of the Stroke Scale

Despite his many laudable achievements it is for his early stroke work that Rankin is best remembered in the UK - in particular his tool for describing post-stroke disability. Rankin shared his experiences of stroke care while in Glasgow in a series of papers submitted to the newly published periodical – The Scottish Medical Journal (which had emerged from the Journal of the Royal Medico-Chirurgical Societies of Glasgow and Edinburgh and the Edinburgh Obstetrical Society). He summarised his research in a seminal manuscript, which was subsequently presented across three papers in the journal: Cerebral vascular accidents in patients over the age of 60. Volume 1: General considerations(174); Volume II: Prognosis(92); Volume III Diagnosis and Management.(182) In these papers he presented critical review of an embryonic stroke medicine literature; described his failed attempts at establishing a West of Glasgow stroke registry and reported his observations of 206 stroke patients admitted through his department and followed to death or discharge. It was in this work that Rankin described his eponymous stroke scale, a tool that was to become instrumental in future stroke studies and that has formed the basis of much of this thesis. Despite Rankin's impressive ground-work, a "Materia Medica" stroke unit with a comprehensive database of cerebrovascular disease was not realised in the Western Infirmary until the early 1990s.

For any physician with an interest in cerebrovascular disease, these papers remain fascinating and prove Rankin's exemplarily knowledge of stroke medicine. It is evidence both of Rankin's forward thinking and of the disappointingly slow progress that has been made in stroke since, that many of the observations made in these original papers remain true today. As an example, although no evidence-based acute stroke therapies were available during Rankin's lifetime, he correctly surmised: "at the onset when treatment is likely to be of most value, accurate diagnosis is often difficult and sometimes impossible." Unfortunately in this age of sophisticated imaging and increasing numbers of proven and potential acute interventions, diagnosis of acute stroke remains a clinical challenge.(183)

Comparison of observations made by Rankin with recent cerebrovascular literature provides further salutary evidence of the prescient nature of Rankin's essays:

John Rankin 1957 "*The importance of these lesions (stroke) is widely realised but is hardly reflected in the volume of research devoted … compared even to a disease as rare as myasthenia gravis.*"

Rothwell et al 2004 "*Stroke and stroke research remains depressingly under-funded.*"(184)

John Rankin 1957 "*the number of deaths increases yearly … and in many instances admission to the appropriate ward is delayed because of shortage of beds*."

UK stroke audit 2004 "*Despite evidence of efficacy…two thirds of stroke units are having to ration access to this limited resource*."(185)

Rankin described good outcomes in the majority of patients cared for using his unorthodox methods of holistic stroke care. To aid his descriptive work he formulated a novel outcome scale. Rankin's instrument consisted of five hierarchical grades of "functional recovery" from Grade I – no significant disability to Grade V severe disability (appendix F). In early stroke work it was not uncommon for authors to describe a bespoke outcome measure. Although some other authors made use of the scale(186), during Rankin's lifetime there was little to distinguish his scale from others cited in the literature. It was not until the development of intervention trials that the scale was "rediscovered".

For the first multi-centre trial in neurology - the UK TIA study(187) - trialists needed an easily administered measure of stroke outcomes. Rather than develop a de-novo instrument they turned to Rankin's eponymous scale. Following initial pilot work, the UK TIA team revised the wording of Rankin's original gradings to allow for better reliability – the modified Rankin scale (mRS)(20). The mRS was subsequently used in the first International Stroke Trial (IST)(188). The success of these trials alerted the stroke community to the utility of Rankin's scale.

The UK TIA authors attempted to further refine the mRS through the development of the Oxford Handicap Scale (OHS)(93). However, OHS was felt to add little to the mRS, was infrequently used by trialists and has been largely abandoned by the stroke community. In comparison mRS has seen increasing use in the scientific press. A review of the literature in the late 1990s reported mRS as the second most popular disability outcome measure in stroke work.(57) My own review of recent stroke trials has demonstrated that mRS is now the preferred disability outcome measure of UK and international researchers. In many of the pivotal clinical trials that have shaped modern stroke practice, mRS has been used often as primary end-point. Recent examples include "landmark" studies of thrombolytic therapy(189), neuro-protectant(190) and surgical treatment of stroke(138). We must assume that Rankin's original work describing this stroke outcome scale is one of the most referenced articles in the SMJ's 50 year history.

Rankin's scale was developed to aid his descriptive analysis of the natural history of stroke and its putative treatments. It is not clear if he intended for others to use the scale for trial outcome. Indeed, the multi-centre clinical trials that characterise contemporary cardiovascular medicine were unknown during Rankin's tenure at Stobhill. It was many years after the Rankin scale had become popular as a stroke research tool that researchers began to analyse its clinimetric properties. A subsequent body of literature, including much of the work in this thesis, has demonstrated the limitations of standard mRS administration and in particular the potential for substantial inter-observer variability.

In recognition of this potential weakness of the scale, stroke researchers are now using modern technologies to improve application of the mRS and examples include the DVD based training resources and digital recording of mRS interviews for later off-line assessment discussed in this thesis. Given that Rankin's scale was developed during his time in Stobhill hospital it is appropriate that much of this work is being conducted by a team based in the University of Glasgow.

John Rankin excelled as scholar, teacher, administrator and physician. The legacy of his stroke scale continues to influence research and it is likely we will continue to refer to Rankin's early Scottish Medical Journal publications well into the 21st century.

Conclusions and future directions

In this thesis I have presented a body of work, exploring the use of outcome scales for stroke trials. From systematic review of the literature I have shown that the most prevalent functional outcome measure is the mRS, most often performed at ninety days post ictus and via telephone interview. Further systematic review and meta-analysis has shown the potential for substantial inter-observer variability in mRS grading. This finding is supported by my own departmental study of mRS reliability. Potential interventions to improve mRS reliability include standardising the interview process; training assessors and using novel methodologies for data collection. Through literature review and my own study I have demonstrated that a structured approach to mRS assessment does not consistently improve the quality of the data. I have further outlined the development of a DVD based training resource for mRS assessment that has been used by thousands of international researchers. Analysis of these submitted training data have shown that mRS variability is present across countries and across medical disciplines. Novel methods of collecting mRS outcomes should be tested prior to clinical use. My study of deriving mRS from patients' case records demonstrated poor reliability. However, pilot study of using remote video based assessment has suggested initial efficacy.

Although I have presented and discussed a number of studies of mRS assessment, using a broad range of research techniques, there is still much that is unknown regarding the properties and application of this scale and there is substantial potential for further important studies in this area.

Where methodology is sufficiently described, mRS is most often performed using telephone based assessment. The properties of a telephone based mRS interview are poorly described in the literature, while properties of other mRS assessment methodologies such as postal questionnaire have never been described. Use of an assessment technique that does not rely on direct interview is attractive in terms of convenience and economics. However, as demonstrated by my study of case record derived mRS, reliability of a technique should not be assumed. Using a "mock clinical trial" design, as described in chapter four, a prospective blinded assessment of telephone mRS (or other assessment) would be possible and would provide useful data for planning future trials.

In this thesis, "success" of the video based training package is evidenced through its popularity with trialists and rates of satisfactory performance on the final certification assessment. Although encouraging, these data do not prove the efficacy of the resource as a teaching aid. A study of untrained researchers could test the impact of the resource on mRS scoring. The researchers would grade a series of mRS interviews ("live" or recorded) and scores would be collated. They would then attempt the training materials and certification exercise. On successful completion of training, they would grade a further series of mRS interviews. Inter-observer agreement and agreement with (group adjudication panel) "correct" mRS could be described in the standard fashion. To account for possible learning effect of repeated mRS interview a control group, not exposed to the teaching materials, could be used. Such an approach would lend itself to the online environment of current mRS training.

In chapter eight I outlined pilot data describing use of remote video based mRS assessment. I acknowledge the preliminary nature of this work but initial results are encouraging. To robustly test the properties of video based assessment in a clinical trial setting would require a larger multi-centre study. Future studies comparing reliability between adjudication panels would help better define the clinimetrics of video based group mRS. Informed by the pilot work presented in this thesis, a multi-centre study of central adjudication of video based mRS has been devised and with funding from the Chief Scientist Office is due to begin this year (2009). (Successful grant application included as appendix)

Appendix A: modified Rankin Scale

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

TOTAL (0-6): _____

Appendix B: Barthel Index

Feeding
0 = unable
5 = needs help cutting, spreading butter, etc., or requires modified diet
10 = independent

Bathing
0 = dependent
5 = independent (or in shower)

Grooming
0 = needs to help with personal care
5 = independent face/hair/teeth/shaving (implements provided)

Dressing
0 = dependent
5 = needs help but can do about half unaided
10 = independent (including buttons, zips, laces, etc.)

Bowels
0 = incontinent (or needs to be given enemas)
5 = occasional accident
10 = continent

Bladder
0 = incontinent, or catheterized and unable to manage alone
5 = occasional accident
10 = continent

Toilet Use
0 = dependent
5 = needs some help, but can do something alone
10 = independent (on and off, dressing, wiping)

Transfers (bed to chair, and back))
0 = unable, no sitting balance
5 = major help (one or two people, physical), can sit
10 = minor help (verbal or physical)
15 = independent

Mobility (on level surfaces)
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 any aid; for example, stick) > 50 yards

Stairs
0 = unable
5 = needs help (verbal, physical, carrying aid)
10 = independent

TOTAL (0–100):

The Barthel ADL Index: Guidelines

1. The index should be used as a record of what a patient does, not as a record of what a patient could do.
2. The main aim is to establish degree of independence from any help, physical or verbal, however minor and for whatever reason.
3. The need for supervision renders the patient not independent.
4. A patient's performance should be established using the best available evidence. Asking the patient, friends/relatives and nurses are the usual sources, but direct observation and common sense are also important. However direct testing is not needed.
5. Usually the patient's performance over the preceding 24-48 hours is important, but occasionally longer periods will be relevant.
6. Middle categories imply that the patient supplies over 50 per cent of the effort.
7. Use of aids to be independent is allowed.

Appendix C: Manuscripts reviewed and excluded from systematic study of modified Rankin Scale reliability.

Papers recovered from original search

Albanese MA. 1994

Excluded as does not use mRS based assessments.

Albanese MA. Clarke WR. Adams HP. 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). Journal of Head Trauma Rehabilitation 1994;9:1746-51.

Davidson I. 2001

Excluded as does not use mRS based outcome assessments.

Davidson I. Booth J. Hillier VF. Waters K Inter-rater reliability of rehabilitation nurses and therapists. British Journal of Therapy and Rehabilitation 2001;8:462-7.

De Haan 1993

Excluded as no inter / intra-observer mRS comparison.

De Haan R. 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;24:1178-1181.

Duncan PW. 2002

Excluded as no inter / intra-observer mRS comparison.

Duncan PW. Lai SM. Tyler D. Perera S. Reker DM. Studenski S.Evaluation of proxy responses to the Stroke Impact Scale. Stroke. 2002;33:2593-9.

Celani MG 2002

Excluded as uses Oxford Handicap Scale and dichotomised outcomes only.

Celani MG. Cantisani TA. Righetti E. Spizzichino L. Ricci S. Italian International Stroke Trial (IST) Collaborators. Different measures for assessing stroke outcome: an analysis from the International Stroke Trial in Italy. Stroke. 2002;33:218-23.

Halkes PH 2006

Excluded as no inter / intraobserver mRS comparison.

Halkes PH. van Gijn J. Kappelle LJ. Koudstaal PJ. Algra A. Classification of cause of death after stroke in clinical research. Stroke. 2006;37:1521-4.

Hantson L 1994

Excluded as no inter / intraobserver mRS comparison.

Hantson L. De Weerdt W. De Keyser J. Diener H.C. Franke C. Palm R. Van Orshoven M. Schoonderwalt H. De Klippel N. Herroelen L. Feys H. The European Stroke Scale. Stroke.1994;25:2215-9.

Merino JG 2005

Excluded as uses only dichotomised mRS.

Merino, Jose G. Lattimore, Susan U. Warach, Steven.

Telephone assessment of stroke outcome is reliable. Stroke. 2005;36:232-3.

Meyer B.C. 2002

Excluded as no inter / intraobserver mRS comparison.

Meyer B.C. Hemmen T.M. Jackson C.M. Lyden P.D. Modified National Institutes of Health Stroke Scale for use in stroke clinical trials: prospective reliability and validity. Stroke. 2002;33:1261-6.

Oveisgharan S 2006

Excluded as no inter / intraobserver mRS comparison.

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. Cerebrovascular Diseases. 2006;22:350-4.

Quinn TJ 20008

Excluded as mRS derived with no direct patient contact.

Quinn TJ, Ray G, Atula S, Walters MR, Dawson J, Lees KR. Deriving Modified Rankin Scores from Medical Case-Records. Stroke. 2008;39:3421-3.

Quinn TJ 2008

Excluded as mRS derived with no direct patient contact.

Quinn TJ, Dawson J, Walters MR, Lees KR. Variability in Modified Rankin Scoring Across a Large Cohort of International Observers. Stroke. 2008;39:2975-9

Reeves MJ 2008

Excluded as mRS derived with no direct patient contact.

Reeves M.J. Mullard A.J. Wehner S. Inter-rater reliability of data elements from a prototype of the Paul Coverdell National Acute Stroke Registry.BMC Neurology. 8, 2008. Article Number: 19.

Shinohara Y. 2006.

Excluded as mRS derived with no direct patient contact.

Shinohara Y. Minematsu K. Amano T. Ohashi Y. Modified Rankin scale with expanded guidance scheme and interview questionnaire: inter-rater agreement and reproducibility of assessment. Cerebrovascular Diseases. 2006;21:271-8.

Visser MC. 1992

Excluded as does not review stroke patients.

Visser MC. Koudstaal PJ. van Latum JC. Frericks H. Berengholz-Zlochin SN. van Gijn J.Inter-observer variation in the application of 2 disability scales in heart patients. Nederlands Tijdschrift voor Geneeskunde. 1992; 136:831-4.

Papers recovered from bibliographic review of retrieved reports and additional searches

Atiya 2003

Excluded as mRS derived with no direct patient contact.

Atiya M, Kurth T, Berger K, Buring JE, Kase CS. Inter-observer agreement in the classification of stroke in the Women's Health Study. Stroke. 2003; 34: 565-567

Bamford 1989

Exclude as does not use mRS based outcome assessments.

JM Bamford, PA Sandercock, CP Warlow, and J Slattery Inter-observer agreement for the assessment of handicap in stroke patients Stroke. 1989;20:828.

Cup EH 2003

Excluded as does not use mRS based outcome assessments.

Cup EH. Scholte op Reimer WJ. Thijssen MC. van Kuyk-Minis MA. Reliability and validity of the Canadian Occupational Performance Measure in stroke patients. Clinical Rehabilitation 2003;17:402-9.

Côté R 1988

Excluded as provides only review data.

Côté R, Batista RN, Wolfson CM, Hachinski V. Stroke assessment scales: guidelines for development, validation and reliability assessment. Can J Neurol Sci 1988;15:261-265.

Jaillard AS

Exclude as does not use mRS based outcome assessments.

Jaillard AS on behalf of the MAST-E group - value of the phone interview in stroke outcome assessment. (abstract) Cerebrovasc Dis 1995; 5:269

Loewen SC 1988

Excluded as did not use mRS based outcome assessments.

Loewen SC, Anderson BA Reliability of the modified motor assessment scale and the Barthel index. Phys Ther 1988;68:1077-1081.

Appendix D: Functional outcome measures used in contemporary stroke trials

Action research activities index
Author's own (three)
Barthel Index
Berg balance
Canadian stroke scale
Composite / Global scale incorporating more than one of these tools
Composite spasticity index
Disability assessment scale
EuroQOL
Fugyl meyer
Frenchay Activites Index
Glasgow Outcomes Scale
Geriatric depression scale
Global health
Global health status
Human activity profile
Katz Activities of Daily Living Scale
Late Life Functional Dependence Index
Lidcombe test plate
Modified Ashford scale
Modified Rankin Scale
Motor activity log
Motor assessment scale
Motor skill performance
MRC strength scale
National Institutes of Health Stroke Scale
Optotrak assessment
Orgogozo scale
Oxford Handicap Scale
Physician assessment scale
Physiological cost
Porch Index of Communicative Ability

Resource use questionnaire

Rivermead Mobility Index

Scandinavian stroke scale

Short from 36

Sickness impact profile

Single question

Six minute walk

Stroke impact scale

TEMPA (Upper Extremity Performance Test for the Elderly)

Timed up and go

Timed walk

Tinetti scale

Upper limb of BFM test

Walking impairment questionnaire

Wolf motor function test

Appendix E: Pro-forma for assessment of video mRS.

Assessor initials: **Date:**

Patient ID: **Audio only: Y N**

Modified Rankin 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, 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	5

Were you able to confidently score the interview: **Y N**
If no, tick all the contributory factors:

Technical
Poor visual quality _____
Poor sound quality _____
Incomplete recording _____
Other (please specify) _____

Interview specific
Incomplete mRs interview _____
Lack of clarity on a specific _____
Other (please specify) _____

Patient specific
Speech problem _____
Cognitive impairment _____
Insufficient information _____
Inconsistent answers _____
Other (please specify) _____

Consensus reached on mRs **Y N**

Final group mRs score

Appendix F: Original Rankin Stroke Scale and derivations.

Rankin Stroke Scale

	Description
Grade I	No significant disability, able to carry out all usual duties
Grade II	Slight disability, unable to carry out shome of previous activities but able to look after own affairs without assistance
Grade III	Moderate disability, requiring some help but able to walk without assistance
Grade IV	Moderately severe disability, unable to walk without assistance and unable to attend to own bodily needs without assistance
Grade V	Severe disability, bedridden, incontinent and requiring constant nursing care and attention

Oxford Handicap Scale

Handicap	Lifestyle	Grade
none	no change	0
minor symptoms	no interference	1
minor handicap	some restrictions but able to look after self	2
moderate handicap	significant restriction; unable to lead a totally independent existence (requires some assistance)	3
moderate-to-severe handicap	unable to live independently but does not require constant attention	4
severe handicap	totally dependent; requires constant attention day and night	5

Appendix G: Application for funding to support multi-centre study of video based mRS

Chief Scientist Office Form 2

Grant application Form	**CSO reference number: CZB/4/595**

Please complete this form in Verdana 10 point font size

Project title (not more than 25 words):

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

Duration of Project *(months)*: 18

Chief Investigator:

Name and title	Position	Institution
Professor Kennedy R Lees	Professor of Cerebrovascular Medicine	University of Glasgow

Project summary *(not more than 150 words)*:

Clinical trials in acute stroke require assessment of functional outcome. Unfortunately, wide inter-rater variation hinders use of such assessments leading to frequent misclassification of the trial endpoint, which impacts upon statistical efficiency and trial power. Central reading of endpoint assessments using video recordings could reduce misclassification by improving reliability of assessments, reduce the number of trial observers and will allow central adjudication of "misclassified" patients. This could enhance ability to assess treatment effects with smaller required sample sizes providing major cost savings to study sponsors, research councils and industry.

We will demonstrate feasibility, reliability and acceptability of centrally adjudicated endpoints in acute stroke trials via a multi-centre study of at least 300 patients over 2 years. Then, through both simulations and practical application to ongoing clinical trials, we will estimate the improvements brought to sample size and power of stroke trials.

1. Application for a research grant in: *(please tick)*

	Full grant	Small grant
Biomedical & Therapeutic Research	√	
Health Services Research		

2. Project category: *(please tick)*

New project	√
Resubmission	
Clinician Scientist	
Request for Supplementary Funding	

3. Research category: *(please tick)*

Clinical Trial subject to the Clinical Trial Regulations	
Other	√

4. **Keywords** *(please use suggested lists)*:

Primary:

Stroke

Secondary:

Outcomes	Disability	Telemedicine		

5. Dates:

Proposed start date	1/8/2008
Proposed finish date	1/2/2010

Section 1

(Not more than 8 pages)

Proposed research project:

1. Title
2. Introduction *(citing key references, searches used, etc.)*
3. Results of any pilot studies
4. Aims
5. Research questions
6. Plan, methods, expertise available, statistical power
7. Timetable
8. Existing facilities
9. Justification of requirements
10. Research outcomes relating to NHS implementation potential
11. Dissemination
12. Key references
13. Relevant additional material

1. CENTRAL ADJUDICATION OF MODIFIED RANKIN SCALE DISABILITY ASSESSMENTS IN ACUTE STROKE TRIALS

2. INTRODUCTION - Stroke is among the most important causes of severe disability, death and health care resource consumption in the Western world (1,2,3). Despite this there are few effective treatments and many neuroprotectant trials have failed (4,5,6); up to 50 drugs have been discontinued after initial studies. While this may be partly explained by lack of treatment efficacy, inadequate study design has contributed. Assumptions regarding anticipated treatment effect and event rates have been overambitious and inaccurate (7) leading to trials that were underpowered (8). Even with recent innovations in trial design (9), standard power calculations still suggest, because outcomes are variable and there is marked heterogeneity of causes, that several thousand patients are required for functional outcome trials of neuroprotectant strategies.

Acute stroke trials therefore require being large and expensive. The need for a robust measure of functional outcome makes them yet more challenging. 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). However, there are concerns with its use - considerable inter-observer variability is recognised (10,11) and traditional dichotomised methods of outcome analysis disregard important differences between adjacent mRs groups (12).

Table 1 – The Modified Rankin Score

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

Trial power is influenced by many factors and inter-observer variability is a particular concern. Single centre studies typically give weighted kappa statistics for inter-observer agreement of 0.7 to 0.8 with use of the mRs (10). A three-site comparison which involved 15 raters found an un-weighted kappa statistic of only 0.25 (10). Such variability increases the risk of assigning patients to the wrong outcome group (endpoint misclassification) which can introduce bias and increase type two error rates (13,14) and reduce trial power. This is likely to feature in large multi-centre stroke trials which involve many hundred observers and this potential for misclassification of endpoint represents a major design flaw in acute stroke trials.

The Importance of Endpoint Misclassification - Data exist to support a significant detrimental effect of endpoint misclassification. For example, in a trial of pneumococcal vaccine (13) in which accuracy of identification of fatal respiratory tract infection (the trial endpoint) was erroneously assumed to be 100%, trial power was compromised: modest misclassification of the cause of death occurred and reduced trial power by 40% (from 93% to 54%). Analysis of data from a trial in neurotrauma (14) reveals that erroneous misclassification of patient outcome has a significant impact on estimates of effect size: without misclassification, the treatment effect was 7.5% (p=0.039). With 10% misclassification, this dropped to 6% (p=0.102), while with 20% misclassification it was only 4.5% (p=0.228). We have performed preliminary analysis to show this is likely to be an important factor in acute stroke trials. Using the distribution of mRs scores seen in the SAINT trial (9), an 18% rate of

mRs misclassification (19) for categories of 1 and 2 would have a minor impact on the treatment effect seen if favourable outcome were defined as mRs 0 or 1 (absolute treatment benefit of 3.64% with misclassification and 4% without). Although these appear little different, the corresponding sample sizes required for 80% power under a continuity-corrected chi-squared test are 2605 and 2218 per group; a large difference which could bring significant time and cost savings.

Neurosurgical and Other Interventional Trials - Several neurosurgical and neuro-radiological trial programmes have yielded encouraging results. For example, the recently reported MELT study (MCA Embolism Local fibrinolytic intervention Trial) (15) suggests that the intra-arterial approach represents a viable treatment option in those with middle cerebral artery (MCA) occlusion: this observation is supported by registry data of patients treated with mechanical thrombectomy using the Multi-MERCI device (16). There is a clear ethical and scientific need to build on these observational data with prospective randomised controlled trials comparing intervention with best medical therapy. Further, a pooled meta-analysis of three randomised controlled trials (DECIMAL, DESTINY and HAMLET) (17) revealed startling evidence of benefit of decompressive hemicraniectomy in those with malignant MCA syndrome. Important questions raised during these trials (for example, identification of patients most likely to benefit and the optimal timing of surgery) now need to be addressed.

These interventions are, by their nature, difficult to study and treatment masking is difficult to ensure. Ideally, a prospective randomised open-label blinded end point evaluation (PROBE) design would be employed. This could be employed in acute stroke trials if mortality or recurrent event rate was the endpoint but is not feasible when measures of post stroke disability such as the mRs assessment are used. These measures mandate a clinical assessment of the participant which for practical reasons is typically performed in the local trial centre by a research nurse or doctor. It is unlikely that these individuals can guarantee being blinded to treatment allocation, making a PROBE design impractical.

Central adjudication of endpoint assessment could reduce misclassification and ensure blinded endpoint assessment. This is not entirely without precedent in stroke trials and is routinely employed in the context of imaging related endpoints. We hypothesise that central adjudication of endpoint assessment is feasible and can be achieved by digital recording of mRs assessments.

Potential Benefits of Central Adjudication - Digital video recording of mRs assessments in a large clinical trial 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, help ensure quality of data (via source data verification and by ensuring adherence to interview procedures) and improve blinding of endpoint assessment.

Central adjudication may also afford examination of more subtle gradations of disability. We believe a central outcomes adjudication panel could compare and rank a large series of patients, thereby giving a measure of spread within, as well as between, mRs categories. Again, this is not without precedent. When using a subjective ranking technique (18), raters rank each patient in a clinical trial according to their trial experiences and the distribution of ranks between treatment groups is compared. This method has been applied to data from the Systolic Hypertension in the Elderly Program (18) and provides a sensitive measure of treatment effect. This novel approach requires careful evaluation and is an important secondary aim of our project. As an example, such an approach could include information derived from other assessments such as the NIHSS.

Before video recording of outcomes and central adjudication could be widely adopted, it must be rigorously assessed. Even though the technique is based upon an adaptation of a commonly used method there are several areas of note. First, the mRs

by nature is subjective and whether important extra information (such as how the patient travelled to hospital or other background details) contributes and by how much is unclear. This approach will add to complexity of trial design and although we feel by an insignificant amount, the technique must yield benefit before it could be deemed worthwhile.

3. Pilot Work. Pilot Work to Clarify the Extent of Variability and Endpoint Misclassification - We have reviewed the results of over 1500 observers who assigned mRs scores to the example cases used in our mRs training DVD – a unique dataset (19). Substantial interobserver variability was confirmed – only 81% of observers achieved a "pass" on the first attempt and less than one third of these individuals scored all 5 cases correctly. Disagreement was especially evident in cases where the correct mRs score was in the range of 2 to 3. In three separate example cases approximately 40% of observers misclassified the patient. The junction between mRs scores 2 and 3 is critical; it defines independence or dependence after stroke and is a commonly used cut-off for dichotomised endpoints. Major acute stroke trials involve as many as 400 hospitals, each with 2-5 raters, ensuring that endpoint misclassification will be a frequent and major confounding issue.

Pilot Work to Support Use of Video mRs Assessments and Central Adjudication - We have performed a pilot study of this technique in our centre on 100 post-stroke patients (manuscript in preparation, abstract published (20). Patients were graded by two independent assessors, one of whom was randomly assigned to record their assessment on video. These recordings were reviewed at least 3 months later by four experienced researchers and three experienced research nurses who independently assigned an mRs score. This was done blinded to both the initial mRs score and that of other observers. We showed the technique to be feasible with a high technical success rate and excellent precision compared to "correct mRs". Agreement between the seven observers was good on review of the recorded mRs assessment (k=0.64) and superior to agreement during the standard mRs assessments (k=0.57) and was as good or better than that seen in previous studies using traditional mRs assessments. Thus, remote review of a recorded mRs assessment is not inferior to standard techniques, could be employed in a PROBE design and interestingly, performed better than use of a structured interview. Work to establish methods of central adjudication of cases of disagreement is ongoing.

4. AIMS - We aim to establish whether central adjudication of locally recorded mRs assessments can be performed in a multi-centre trial setting, to clarify the extent of endpoint misclassification and to assess whether this can be addressed with demonstrable improvements in trial statistical power. We will also explore whether a subjective ranking technique can be applied to mRs outcomes and thus provide a more sensitive measure of post-stroke disability.

5. RESEARCH QUESTIONS - Does central adjudication of video recordings of mRs assessments;

1. Provide a feasible method of measuring outcome in a multi-centre 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. Exert meaningful effects on statistical power and required sample size in clinical trials?
4. Allow measurement of more subtle effects on outcome through grading of outcomes within mRs categories, and if so, which statistical approaches allow this to be undertaken practically?

6. PLAN, METHODS, EXPERTISE AVAILABLE, STATISTICAL POWER - We will perform a "virtual" acute stroke trial. Baseline data will be gathered and endpoint assessments performed as in an interventional trial. Briefly, patients with a diagnosis

of stroke (ischaemic or haemorrhagic) presenting within 48 hours who have a demonstrable deficit on the NIHSS will be studied. Exclusion criteria will include a pre-morbid modified Rankin score of ≥3. Follow up will occur at 30 and 90 days where an mRs assessment will be performed and recorded. All observers will be trained and certified in the use of the mRs and the video equipment. We will also perform video NIHSS assessments at 90 days.

We aim to recruit a minimum of 300 patients from between 5 and 10 centres. This is the minimum number of patients likely to be required in a phase III acute stroke trial of a reperfusion strategy and will provide sufficient video assessments for us to evaluate the technique and its impact on trial design.

The mRs Assessment – These will be performed on survivors in standard fashion according to each centre's normal practice. This is likely to be in a clinic room, by a patient's bedside or at home if they are unable to attend the hospital. The mRs assessment will be recorded using a digital video camera. The local investigator will assign an mRs score which will be inserted to the eCRF (the standard mRs score). They will also be asked to comment whether there is significant dysarthria or dysphasia. When possible, investigators who have been closely involved in management of a patient should not assess the patient for outcome. Whenever possible, the assessor should remain constant across the follow-up period for a given patient. We recognise these restrictions may be impractical for smaller sites but should be possible for the major centres.

Video Equipment - A high definition video camera will be used (we currently use a HDR-HV1E 1080i digital HD video camera recorder (Sony, Japan; specifications: CCDType: Single CMOS Chip; Video Recording Format: DV Tape; Optical Zoom: 10 times) or equivalent system). Cameras will record to digital video tape. In conjunction, a desktop omni-directional condenser boundary microphone will be used (we currently use an ATR97, Audio-technica, Ohio USA; Specifications: Frequency response: 50-1500Hz). An easily portable tripod will be used to mount the video camera (we currently use a Manfrotto 117B movie tripod, Italy).

Transfer of Recording and Upload to Co-ordinating Centre - The digital recordings will be transferred to a secure computer hard drive via a FireWire (IEEE) or USB 2 connection. Windows Movie Maker software (Microsoft, USA) will be used to anonymise and code the recording and convert it to MPEG format. A title will be placed at the start of the clip stating only the patient's study number, study centre, mRs assessor's initials and investigator code. No other editing will take place, unless required to maintain patient anonymity. If this is required, it will be recorded in the eCRF along with the length and nature of clip removed.

The edited clip will be uploaded to the Rankin Outcome Adjudication web portal. It will also be recorded to compact disc (CD) which, along with the digital video tape, will be archived locally. From our experience with the technique and equipment, clips will be between 5 and 10 minutes long and 7 and 15 MB in MPEG format. Central archiving and storage of copies of clips will be as for other trial related data specified in the protocol and University of Glasgow procedures.

The Rankin Outcome Adjudication Web Portal – This will provide tools for investigators to enter their subjects' modified Rankin Scale assessments and upload accompanying videos. The portal will be administered by the Robertson Centre for Biostatistics. The web portal will include a system that will make new videos available to the outcomes manager for quality checks and pre-review editing and transcription, assign new videos and data to assessors, permit them to make notes and to complete an adjudication form online. In addition, allocation to endpoint review committee for further assessment can be implemented in the case of disagreement. Assessors and committee members will be able to record periods of leave to facilitate timely

turnaround of assessments. Standard performance metrics will be created and stored. The system will allow reminders (e-mail and at next log-in) to be initiated.

The web portal will be secure and end-users will access the system by entering a username and password. Assessors will be grouped by centre. 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.

Review of the Digital mRs Assessment – Recorded assessments will be reviewed at the outcome coordinating centre (Western Infirmary Acute Stroke Unit, Glasgow, UK) which will employ an adjudication procedure and review all endpoint assessments. Upon upload of an mRs assessment, the Outcomes Manager (a clinical research fellow) will be notified by an automated email. The fellow or a designee will then review the assessment (via the web portal and within 72 hours of submission) and remove any patient identifiable information from the assessment clip using Windows Movie Maker software. The fellow will also verify that assessor is currently certified in mRs assessment. If the assessor is not trained, the assessment will need to be repeated by a trained observer. If assessor training is valid, the fellow will assign an mRs score and then release the assessment for review by the endpoint assessment committee. If any editing has occurred to ensure blinded assessment, the original clip will be maintained and the nature of this editing recorded in the eCRF. The edited clip will be re-uploaded to the web portal and noted as the clip to be used by the endpoint assessment committee. The endpoint committee members will be notified by email.

The endpoint committee will be made up of the named clinical applicants. A minimum of four will review the mRs assessment and assign an mRs score. This will be done independently to the local investigator score, blinded to all other patient information and within seven days of the fellow releasing the clip. The assessment will be viewed via the web portal and the score entered into the portal. An automated process will establish whether these scores agree with the initial local score and if so, the patient is assigned to the common mRs category. If there is any disagreement, the patient is 'misclassified' and the video clip will be submitted for further review by the entire endpoint assessment committee. Note that the website will indicate to the fellow that disagreement has occurred but not the original investigator's score or the nature of the disagreement. Once the fellow is notified of a disagreement, he or she will inform the committee by email. After group review, the committee will assign the patient to one of the following groups: technically inadequate assessment, inadequate assessment, adequate assessment with unanimous committee agreement or adequate assessment with non-unanimous committee decision (where majority opinion as to which score is correct will apply). Where committee classification is possible, the patient will be assigned to that Rankin category. Otherwise, the submitting centre may be asked for further information (for example, to put a specific additional question to a patient) or to repeat the assessment if deemed necessary.

Observer Training – **All investigators will be trained in mRs assessment using a validated DVD based training programme. During the training sessions in mRs, observers will 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 minimum requirement is that one individual from each centre will be given one-to-one training by a member of the endpoint assessment team and will therefore be able to provide further training and demonstration to other individuals should it be required. A written instruction manual and summary pamphlet will be delivered with the video equipment to each active centre. Repeat training will be offered if video quality is below the required standard or on request from sites. Technical questions will be answered by email and/or telephone.**

Development of The 'outcome ranking' Technique - It is not possible for an observer to watch and accurately rank a large series of patients - a full review for a trial of 1700 patients would require nearly 1.5 million pairwise comparisons ($[n^2-n]/2$).

However, patients are already partially ordered by mRs category and sorting algorithms exist to identify the pairwise comparisons required to rank the outcomes. Because this is entirely novel, it is not possible to state exactly which technique will be most suitable for use in an acute stroke trial, although the *shell, library* and *quicksort* algorithms are possibilities. Identifying the most suitable technique is an express and key secondary aim of this project.

Expertise Available – This study will be a collaboration between departments of the University of Glasgow and we expect several acute centres of the UK Stroke Research Network. Personnel at the coordinating centre have extensive experience in acute stroke trial design and outcome assessment and have developed the mRs training programme outlined above. The team is led by Professor KR Lees. The endpoint assessment committee members are Professor Lees (chair), Dr J Dawson (Lecturer in Medicine), Dr MR Walters (Senior Lecturer in Medicine), Dr K Muir (Senior Lecturer in Neurology), Prof P Langhorne (Professor of Stroke Care) and Dr T Quinn (Research Fellow in Stroke). Statistical expertise is provided by Dr C Weir from the Robertson Centre for Biostatistics. The team will also include an Outcomes Manager (a clinical research fellow) who will be specifically employed to handle all trial outcome data and video assessments and will be fully trained in mRs assessment.

We have a strong research record in acute stroke and received a 5 rating in the last RAE. We have performed extensive research into outcome assessment after stroke (8,9) and have developed a DVD based training for mRs assessment and contributed to development of a structured interview for the mRs (10,11). We have led the largest acute stroke trial programmes (SAINT, IMAGES and GAIN trials). Our statistical team is also vastly experienced, has been involved in several large projects (WOSCOPS, IMAGES) and has contributed extensively to stroke trial design (21,22). The Robertson Centre for Biostatistics is an internationally renowned centre with extensive experience in eCRF design and data handling procedures in accordance with ICH Good Clinical Practice and industry regulatory guidelines. The Centre is accredited for ISO 9001:2000 for its quality systems and has TickIT accreditation for its software development.

The UK SRN is a recently launched DoH & CSO initiative with the express aim of facilitating conduct of randomised prospective trials and other well-designed studies of stroke. We are one of the coordinating centres and hold two Associate Director posts (Prof Lees and Ford), chair of the acute care group (ACG) (Prof Lees) and positions on the ACG (Dr Muir) and rehabilitation committees (Professor Langhorne). We are certain that we have all the required expertise and resources to make this project a success.

Statistical Analysis - We will assess agreement, using the weighted kappa statistic, between the standard mRs score and central video assessment and establish misclassification rates. We will then assess the effect of local misclassification rates, and the potential benefits of our approach, on trial power and treatment effect estimates via standard statistical formulae.

7. TIMETABLE - The study will take 18 months to complete. It is feasible to recruit 300 plus patients in one year. We anticipate involvement of at least 5 centres and expect to recruit well in excess of 100 patients at the coordinating centre (we recruited this figure in a year during our pilot study). Our collaborators at the Glasgow Royal Infirmary (Prof Langhorne) and the Southern General Hospital (Dr Muir) will also recruit significant numbers and with the aid of the other centres (a minimum of 2) we will reach our target.

8. EXISTING FACILITIES - All facilities and staff required to make this project a success are already available (except the video equipment for which we seek funding). Our pilot work ensures project implementation will be quick and smooth. All involved staff have extensive experience of acute stroke trials and have published on outcome

assessment after stroke. Training materials for mRs use are available and validated. We expect the project will be adopted by the UK SRN which will ensure recruitment targets are met.

9. JUSTIFICATION OF REQUIREMENTS - The total funding sought is £183289. This will provide the required equipment for up to 10 active centres, cover the gathering and uploading of patient information, ensure the availability of required statistical time and that of a dedicated Clinical Research Fellow. The salaries in this application have been calculated using current salary scales which came into effect on 1 August 2005. Assimilation to new pay scales will occur either before or during the period of the proposed research and, as agreed with RCUK, it is expected that a request for remuneration of additional net costs due to this restructuring will be made at the reconciliation stage of any award. Total equipment costs are £13533 with a further £1500 required for consumables (digital video tape).

A dedicated clinical research fellow is required. The fellow will recruit a large number of patients at the coordinating and other local centres, have day to day responsibility for liaising with the other centres, coordinate the review committee, review all assessments, issue data queries and have primary responsibility for the writing of any manuscripts. This easily necessitates a full time commitment for the project duration. We believe a fellow, rather than a trained nurse is required to perform these duties and this represents little extra cost.

The other applicants will make up the endpoint assessment committee. We estimate that up to 20% of cases (approximately 120) will be "misclassified." Allowing approximately 10 minutes per case, this will amount to 20 hours per committee member with a further 30 hours available for the initial review and performing assessments. Extra time (a further 50 hours) is required for Prof Lees and Dr Dawson, who have primary responsibility for project implementation and for assisting in preparing the manuscript and meetings. They will also directly support the new Fellow and to allow the project to continue during periods of annual leave.

We estimate that 3 months of principal study statistician time will be spent on methodological development, simulation work and study data analysis. Due to cost constraints the scope of the statistical methodological work will be restricted to the key components of this proposal but we have ensured availability of senior statistician time to ensure success of the project. Support is required to cover the development and use of case report forms, and to edit and store the digital video assessments. These are estimated at equivalent to 225, 525 and 112.5 hours months of administrative, IT and data management and management support.

The costs of staffing at each local centre will be met by the local host institution and with the aid of the SRN. However, we must reimburse centres for actual expenditure incurred during patient follow-up (telephone, stationery and patient taxi costs). From our experience, £100 per patient recruited is the minimum necessary – amounting to £30000 for the study.

30% of our costs relate to the virtual stroke trial. We have explored whether we could collect sufficient data from any ongoing trial but the costs of incorporating this to a pharmaceutical trial as an add-on would be prohibitive unless it was sponsored by a company; until our methodology is established, no company will wish to risk compromising their trial recruitment. Academic trials are recruiting in UK too slowly to deliver the recruitment needed within a practical timescale. By excluding the need for experimental treatment or testing of other hypotheses, inclusion criteria can be wide, consent will be readily obtained and this important study can be performed promptly.

10. RESEARCH OUTCOMES RELATING TO NHS IMPLEMENTATION POTENTIAL - It is vital that we improve endpoint assessment in acute stroke trials. If this method is successful we hope that the increase in trial efficiency will ease the conduct of further

research and facilitate discovery of new treatments for what is a devastating condition. Conduct of this virtual trial will also facilitate training within each SRN centre and provide evidence of expertise to attract externally funded research trials.

11. DISSEMINATION - This research will clarify important issues regarding acute stroke trial design and we envisage such an important development will be published in a major peer reviewed journal. Data will also be incorporated in to the VISTA, making it available to all academic collaborators. However, in order to ensure its dissemination and use in future clinical trials we will liaise closely with industry and hold training sessions at investigator meetings; we have a clear record of dissemination at academic meetings.

Reference List

(1) Hallstrom B, Jonsson AC, Nerbrand C, Norrving B, Lindgren A. Stroke incidence and survival in the beginning of the 21st century in southern Sweden: comparisons with the late 20th century and projections into the future. Stroke 2008; 39(1):10-15.

(2) Asplund K, Bonita R, Kuulasmaa K, Rajakangas AM, Schaedlich H, Suzuki K et al. Multinational comparisons of stroke epidemiology. Evaluation of case ascertainment in the WHO MONICA Stroke Study. World Health Organization Monitoring Trends and Determinants in Cardiovascular Disease. Stroke 1995; 26(3):355-360.

(3) Brown DL, Boden-Albala B, Langa KM, Lisabeth LD, Fair M, Smith MA et al. Projected costs of ischemic stroke in the United States. Neurology 2006; 67(8):1390-1395.

(4) Johnston SC. The 2008 William M. Feinberg lecture: prioritizing stroke research. Stroke 2008; 39(12):3431-3436.

(5) Ali M, Bath PM, Curram J, Davis SM, Diener HC, Donnan GA et al. The Virtual International Stroke Trials Archive. Stroke 2007; 38(6):1905-1910.

(6) Kidwell CS, Liebeskind DS, Starkman S, Saver JL. Trends in acute ischemic stroke trials through the 20th century. Stroke 2001; 32(6):1349-1359.

(7) Pavlakis SG, Sacco R, Levine SR, Meschia JF, Palesch Y, Tilley BC et al. Lessons from adult stroke trials. Pediatr Neurol 2006; 34(6):446-449.

(8) Youssef MY, Mojiminiyi OA, Abdella NA. Plasma concentrations of C-reactive protein and total homocysteine in relation to the severity and risk factors for cerebrovascular disease. Transl Res 2007; 150(3):158-163.

(9) Gandhi MJ. Does 'ENHANCE' diminish confidence in ezetimibe? J Assoc Physicians India 2008; 56:665-666.

(10) Peto R, Emberson J, Landray M, Baigent C, Collins R, Clare R et al. Analyses of cancer data from three ezetimibe trials. N Engl J Med 2008; 359(13):1357-1366.

(11) Jonsson AC, Lindgren I, Hallstrom B, Norrving B, Lindgren A. Determinants of quality of life in stroke survivors and their informal caregivers. Stroke 2005; 36(4):803-808.

(12) Salter KL, Moses MB, Foley NC, Teasell RW. Health-related quality of life after stroke: what are we measuring? Int J Rehabil Res 2008; 31(2):111-117.

(13) Duncan PW. Stroke disability. Phys Ther 1994; 74(5):399-407.

(14) Roberts L, Counsell C. Assessment of clinical outcomes in acute stroke trials. Stroke 1998; 29(5):986-991.

(15) points to consider on clinical investigation of medicinal products for the treatment of acute stroke. London: European Agency for Evaluation of Medicinal Products, 2001.

(16) Gandhi GY, Murad MH, Fujiyoshi A, Mullan RJ, Flynn DN, Elamin MB et al. Patient-important outcomes in registered diabetes trials. JAMA 2008; 299(21):2543-2549.

(17) Kasner SE. Clinical interpretation and use of stroke scales. Lancet Neurol 2006; 5(7):603-612.

(18) Schepers VP, Ketelaar M, van dP, I, Visser-Meily JM, Lindeman E. Comparing contents of functional outcome measures in stroke rehabilitation using the International Classification of Functioning, Disability and Health. Disabil Rehabil 2007; 29(3):221-230.

(19) Brott T, Adams HP, Jr., Olinger CP, Marler JR, Barsan WG, Biller J et al. Measurements of acute cerebral infarction: a clinical examination scale. Stroke 1989; 20(7):864-870.

(20) van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 1988; 19(5):604-607.

(21) Harwood RH, Rogers A, Dickinson E, Ebrahim S. Measuring handicap: the London Handicap Scale, a new outcome measure for chronic disease. Qual Health Care 1994; 3(1):11-16.

(22) Ackerley SJ, Gordon HJ, Elston AF, Crawford LM, McPherson KM. Assessment of quality of life and participation within an outpatient rehabilitation setting. Disabil Rehabil 2009;1-8.

(23) Buck D, Jacoby A, Massey A, Ford G. Evaluation of measures used to assess quality of life after stroke. Stroke 2000; 31(8):2004-2010.

(24) Lyden PD, Hantson L. Assessment scales for the evaluation of stroke patients. J Stroke Cerebrovasc Dis 1998; 7(2):113-127.

(25) Candelise L. Stroke Scores and Scales. Cerebrovascular Diseases 1992; 2:239-247.

(26) Asplund K. Clinimetrics in stroke research. Stroke 1987; 18(2):528-530.

(27) Feinstein AR. Clinimetric perspectives. J Chronic Dis 1987; 40(6):635-640.

(28) Lees KR, Milia P. Halving effort in acute stroke trials. Clin Exp Hypertens 2006; 28(3-4):309-312.

(29) Bullock MR, Merchant RE, Choi SC, Gilman CB, Kreutzer JS, Marmarou A et al. Outcome measures for clinical trials in neurotrauma. Neurosurg Focus 2002; 13(1):ECP1.

(30) Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke: the impact of using different methods for measuring responsiveness. J Clin Epidemiol 2002; 55(9):922-928.

(31) 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; 38(3):1091-1096.

(32) Fleiss JL, Spitzer RL, Endicott J, Cohen J. Quantification of agreement in multiple psychiatric diagnosis. Arch Gen Psychiatry 1972; 26(2):168-171.

(33) Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992; 304(6840):1491-1494.

(34) Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin 1971; 76:378-381.

(35) Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 1968; 70:213-220.

(36) Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-174.

(37) 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. Int J Epidemiol 2003; 32(3):430-436.

(38) Choi SC, Clifton GL, Marmarou A, Miller ER. Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. J Neurotrauma 2002; 19(1):17-22.

(39) Young FB, Lees KR, Weir CJ. Improving trial power through use of prognosis-adjusted end points. Stroke 2005; 36(3):597-601.

(40) Weaver CS, Leonardi-Bee J, Bath-Hextall FJ, Bath PM. Sample size calculations in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. Stroke 2004; 35(5):1216-1224.

(41) Shuaib A, Lees KR, Lyden P, Grotta J, Davalos A, Davis SM et al. NXY-059 for the treatment of acute ischemic stroke. N Engl J Med 2007; 357(6):562-571.

(42) Hacke W, Albers G, Al Rawi Y, Bogousslavsky J, Davalos A, Eliasziw M et al. The Desmoteplase in Acute Ischemic Stroke Trial (DIAS): a phase II MRI-based 9-hour window acute stroke thrombolysis trial with intravenous desmoteplase. Stroke 2005; 36(1):66-73.

(43) Fischer U, Arnold M, Nedeltchev K, Brekenfeld C, Ballinari P, Remonda L et al. NIHSS score and arteriographic findings in acute ischemic stroke. Stroke 2005; 36(10):2121-2125.

(44) Goldstein LB, Bertels C, Davis JN. Interrater reliability of the NIH stroke scale. Arch Neurol 1989; 46(6):660-662.

(45) 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; 28(2):307-310.

(46) Lyden P, Raman R, Liu L, Grotta J, Broderick J, Olson S et al. NIHSS training and certification using a new digital video disk is reliable. Stroke 2005; 36(11):2446-2449.

(47) Woo D, Broderick JP, Kothari RU, Lu M, Brott T, Lyden PD et al. Does the National Institutes of Health Stroke Scale favor left hemisphere strokes? NINDS t-PA Stroke Study Group. Stroke 1999; 30(11):2355-2359.

(48) Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. Maryland State Medical Journal 1965; 14:61-65.

(49) Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. Int Disabil Stud 1988; 10(2):61-63.

(50) Sinoff G, Ore L. The Barthel activities of daily living index: self-reporting versus actual performance in the old-old (> or = 75 years). J Am Geriatr Soc 1997; 45(7):832-836.

(51) Dromerick AW, Edwards DF, Diringer MN. Sensitivity to changes in disability after stroke: a comparison of four scales useful in clinical trials. J Rehabil Res Dev 2003; 40(1):1-8.

(52) 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; 35(4):918-923.

(53) Granger CV, Dewis LS, Peters NC, Sherwood CC, Barrett JE. Stroke rehabilitation: analysis of repeated Barthel index measures. Arch Phys Med Rehabil 1979; 60(1):14-17.

(54) 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; 30(10):2131-2140.

(55) de Haan R, Limburg M, Bossuyt P, van der MJ, Aaronson N. The clinical meaning of Rankin 'handicap' grades after stroke. Stroke 1995; 26(11):2027-2030.

(56) Dawson J, Lees JS, Chang TP, Walters MR, Ali M, Davis SM et al. Association between disability measures and healthcare costs after initial treatment for acute stroke. Stroke 2007; 38(6):1893-1898.

(57) Sulter G, Steen C, De Keyser J. Use of the Barthel index and modified Rankin scale in acute stroke trials. Stroke 1999; 30(8):1538-1541.

(58) 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. Restor Neurol Neurosci 2005; 23(3-4):257-263.

(59) Demchuk AM, Tanne D, Hill MD, Kasner SE, Hanson S, Grond M et al. Predictors of good outcome after intravenous tPA for acute ischemic stroke. Neurology 2001; 57(3):474-480.

(60) 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):381-387.

(61) Uyttenboogaart M, Luijckx GJ, Vroomen PC, Stewart RE, De Keyser J. Measuring disability in stroke: relationship between the modified Rankin scale and the Barthel index. J Neurol 2007; 254(8):1113-1117.

(62) Wilson JT, Hareendran A, Grant M, Baird T, Schulz UG, Muir KW et al. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. Stroke 2002; 33(9):2243-2246.

(63) Wilson JT, 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; 36(4):777-781.

(64) 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):377-378.

(65) Wilson JT, Edwards P, Fiddes H, Stewart E, Teasdale GM. Reliability of postal questionnaires for the Glasgow Outcome Scale. J Neurotrauma 2002; 19(9):999-1005.

(66) Duncan PW, Reker DM, Horner RD, Samsa GP, Hoenig H, LaClair BJ et al. Performance of a mail-administered version of a stroke-specific outcome measure, the Stroke Impact Scale. Clin Rehabil 2002; 16(5):493-505.

(67) Candelise L, Musicco M, Aritzu E, Pinardi G. Telephone interview for stroke outcome assessment. Cerebrovascuar Diseases 1994; 4:341-343.

(68) Barber M, Stott DJ. Validity of the Telephone Interview for Cognitive Status (TICS) in post-stroke subjects. Int J Geriatr Psychiatry 2004; 19(1):75-79.

(69) Heuschmann PU, Kolominsky-Rabas PL, Nolte CH, Hunermund G, Ruf HU, Laumeier I et al. [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.]. Fortschr Neurol Psychiatr 2005; 73(2):74-82.

(70) Merino JG, Lattimore SU, Warach S. Telephone assessment of stroke outcome is reliable. Stroke 2005; 36(2):232-233.

(71) Chen MH, Hsieh CL, Mao HF, Huang SL. Differences between patient and proxy reports in the assessment of disability after stroke. Clin Rehabil 2007; 21(4):351-356.

(72) Duncan PW, Lai SM, Tyler D, Perera S, Reker DM, Studenski S. Evaluation of proxy responses to the Stroke Impact Scale. Stroke 2002; 33(11):2593-2599.

(73) van der Linden FA, Kragt JJ, Hobart JC, Klein M, Thompson AJ, van der Ploeg HM et al. Proxy measurements in multiple sclerosis: agreement between patients and their partners on the impact of multiple sclerosis in daily life. J Neurol Neurosurg Psychiatry 2006; 77(10):1157-1162.

(74) Hacke W, Bluhmki E, Steiner T, Tatlisumak T, Mahagne MH, Sacchetti ML et al. 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; 29(10):2073-2075.

(75) Uyttenboogaart M, Stewart RE, Vroomen PC, De Keyser J, Luijckx GJ. Optimizing cutoff scores for the Barthel index and the modified Rankin scale for defining outcome in acute stroke trials. Stroke 2005; 36(9):1984-1987.

(76) Young FB, Lees KR, Weir CJ. Strengthening acute stroke trials through optimal use of disability end points. Stroke 2003; 34(11):2676-2680.

(77) Haley EC, Jr., Thompson JL, Levin B, Davis S, Lees KR, Pittman JG et al. Gavestinel does not improve outcome after acute intracerebral hemorrhage: an analysis from the GAIN International and GAIN Americas studies. Stroke 2005; 36(5):1006-1010.

(78) Saver JL. Novel end point analytic techniques and interpreting shifts across the entire range of outcome scales in acute stroke trials. Stroke 2007; 38(11):3055-3062.

(79) Diener HC, Lees KR, Lyden P, Grotta J, Davalos A, Davis SM et al. NXY-059 for the treatment of acute stroke: pooled analysis of the SAINT I and II Trials. Stroke 2008; 39(6):1751-1758.

(80) Fisher M, Hess DC, Lees K. Issues pertaining to the critiques of the SAINT-I Trial. Stroke 2007; 38(11):e126-e127.

(81) Koziol JA, Feng AC. On the analysis and interpretation of outcome measures in stroke clinical trials: lessons from the SAINT I study of NXY-059 for acute ischemic stroke. Stroke 2006; 37(10):2644-2647.

(82) Tilley BC, Marler J, Geller NL, Lu M, Legler J, Brott T et al. 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; 27(11):2136-2142.

(83) Hacke W, Bluhmki E, Steiner T, Tatlisumak T, Mahagne MH, Sacchetti ML et al. 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; 29(10):2073-2075.

(84) Higashida RT, Furlan AJ, Roberts H, Tomsick T, Connors B, Barr J et al. Trial design and reporting standards for intra-arterial cerebral thrombolysis for acute ischemic stroke. Stroke 2003; 34(8):e109-e137.

(85) 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):271-278.

(86) New PW, Buchbinder R. Critical appraisal and review of the Rankin scale and its derivatives. Neuroepidemiology 2006; 26(1):4-15.

(87) Berger K, Weltermann B, Kolominsky-Rabas P, Meves S, Heuschmann P, Bohner J et al. [The reliability of stroke scales. The german version of NIHSS, ESS and Rankin scales]. Fortschr Neurol Psychiatr 1999; 67(2):81-93.

(88) Cincura C, Pontes-Neto OM, Neville IS, Mendes HF, Menezes DF, Mariano DC et al. 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 2008; 27(2):119-122.

(89) Barreca S, Wilkins S. Experiences of nurses working in a stroke rehabilitation unit. J Adv Nurs 2008; 63(1):36-44.

(90) Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009; 6(7):e1000097.

(91) Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000; 283(15):2008-2012.

(92) RANKIN J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. Scott Med J 1957; 2(5):200-215.

(93) Bamford JM, Sandercock PA, Warlow CP, Slattery J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 1989; 20(6):828.

(94) Quinn TJ, Dawson J, Walters MR, Lees KR. Exploring the Reliability of the Modified Rankin Scale. Stroke 2009.

(95) Quinn TJ, Ray G, Atula S, Walters MR, Dawson J, Lees KR. Deriving modified Rankin scores from medical case-records. Stroke 2008; 39(12):3421-3423.

(96) Quinn TJ, Dawson J, Walters MR, Lees KR. Variability in modified Rankin scoring across a large cohort of international observers. Stroke 2008; 39(11):2975-2979.

(97) Wolfe CD, Taub NA, Woodrow EJ, Burney PG. Assessment of scales of disability and handicap for stroke patients. Stroke 1991; 22(10):1242-1244.

(98) Gur AY, Lampl Y, Gross B, Royter V, Shopin L, Bornstein NM. A new scale for assessing patients with vertebrobasilar stroke-the Israeli Vertebrobasilar Stroke Scale (IVBSS): inter-rater reliability and concurrent validity. Clin Neurol Neurosurg 2007; 109(4):317-322.

(99) Meyer BC, Raman R, Chacon MR, Jensen M, Werner JD. Reliability of site-independent telemedicine when assessed by telemedicine-naive stroke practitioners. J Stroke Cerebrovasc Dis 2008; 17(4):181-186.

(100) Oveisgharan S, Shirani S, Ghorbani A, Soltanzade A, Baghaei A, Hosseini S et al. Barthel index in a Middle-East country: translation, validity and reliability. Cerebrovasc Dis 2006; 22(5-6):350-354.

(101) Reeves MJ, Mullard AJ, Wehner S. Inter-rater reliability of data elements from a prototype of the Paul Coverdell National Acute Stroke Registry. BMC Neurol 2008; 8:19.

(102) Visser MC, Koudstaal PJ, van Latum JC, Frericks H, Berengholz-Zlochin SN, van Gijn J. [Interobserver variation in the application of 2 disability scales in heart patients]. Ned Tijdschr Geneeskd 1992; 136(17):831-834.

(103) 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; 25(9):1746-1751.

(104) Atiya M, Kurth T, Berger K, Buring JE, Kase CS. Interobserver agreement in the classification of stroke in the Women's Health Study. Stroke 2003; 34(2):565-567.

(105) 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; 33(1):218-223.

(106) Cup EH, Scholte op Reimer WJ, Thijssen MC, Kuyk-Minis MA. Reliability and validity of the Canadian Occupational Performance Measure in stroke patients. Clin Rehabil 2003; 17(4):402-409.

(107) Cote R, Battista RN, Wolfson CM, Hachinski V. Stroke assessment scales: guidelines for development, validation, and reliability assessment. Can J Neurol Sci 1988; 15(3):261-265.

(108) Davidson I, Booth J, Hillier VF, Waters K. Inter-rater reliability of rehabilitation nurses and therapists. Br J Ther Rehab 2001; 8:462-467.

(109) de Haan R, Horn J, Limburg M, van der MJ, Bossuyt P. A comparison of five stroke scales with measures of disability, handicap, and quality of life. Stroke 1993; 24(8):1178-1181.

(110) Halkes PH, van Gijn J, Kappelle LJ, Koudstaal PJ, Algra A. Classification of cause of death after stroke in clinical research. Stroke 2006; 37(6):1521-1524.

(111) Hantson L, De Weerdt W, De Keyser J, Diener HC, Franke C, Palm R et al. The European Stroke Scale. Stroke 1994; 25(11):2215-2219.

(112) Loewen SC, Anderson BA. Reliability of the Modified Motor Assessment Scale and the Barthel Index. Phys Ther 1988; 68(7):1077-1081.

(113) Jaillard AS. Value of the phone interviewin stroke outcome assessment. Cerebrovasc Dis 1995; 5:269.

(114) de Caneda MA, Fernandes JG, de Almeida AG, Mugnol FE. Reliability of neurological assessment scales in patients with stroke. Arq Neuropsiquiatr (portuguese) 2006; 64:690-697.

(115) 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; 33(5):1261-1266.

(116) Zorowitz RD, Stineman MG. There's no place like home...for some. Stroke 2000; 31(10):2521-2522.

(117) Quinn TJ, Dawson J, Lees JS, Chang TP, Walters MR, Lees KR. Time spent at home poststroke: "home-time" a meaningful and robust outcome measure for stroke trials. Stroke 2008; 39(1):231-233.

(118) Edwards M, Freightner J, Goldsmith CH. Inter-rater reliability of assessments by individuals with or without a background in health care. Occupational therapy journal research 1995; 15:103-110.

(119) Segal ME, Gillard M, Schall RR. Telephone and in-person proxy agreement between stroke patients and caregivers for the Functional Independence Measure. American Journal Physical Medicine and Rehabilitation 1996;208-212.

(120) Lees KR, Zivin JA, Ashwood T, Davalos A, Davis SM, Diener HC et al. NXY-059 for acute ischemic stroke. N Engl J Med 2006; 354(6):588-600.

(121) Saposnik G, Webster F, O'Callaghan C, Hachinski V. Optimizing discharge planning: clinical predictors of longer stay after recombinant tissue plasminogen activator for acute stroke. Stroke 2005; 36(1):147-150.

(122) Formiga F, Mascaro J, Pujol R. Inter-rater reliability of the Barthel Index. Age Ageing 2005; 34(6):655-656.

(123) Carod-Artal FJ, Gonzalez-Gutierrez JL, Egido-Herrero JA, Varela dS. [The psychometric properties of the Spanish version of the stroke-adapted 30-item Sickness Impact Profile (SIP30-AI)]. Rev Neurol 2007; 45(11):647-654.

(124) Williams LS, Bakas T, Brizendine E, Plue L, Tu W, Hendrie H et al. How valid are family proxy assessments of stroke patients' health-related quality of life? Stroke 2006; 37(8):2081-2085.

(125) Adams HP, Jr. Trials of trials in acute ischemic stroke. The Humana Lecture. Stroke 1993; 24(9):1410-1415.

(126) DeJong G, Horn SD, Conroy B, Nichols D, Healton EB. Opening the black box of post-stroke rehabilitation: stroke rehabilitation patients, processes, and outcomes. Arch Phys Med Rehabil 2005; 86(12 Suppl 2):S1-S7.

(127) Horn SD, DeJong G, Ryser DK, Veazie PJ, Teraoka J. Another look at observational studies in rehabilitation research: going beyond the holy grail of the randomized controlled trial. Arch Phys Med Rehabil 2005; 86(12 Suppl 2):S8-S15.

(128) Salter KL, Teasell RW, Foley NC, Jutai JW. Outcome assessment in randomized controlled trials of stroke rehabilitation. Am J Phys Med Rehabil 2007; 86(12):1007-1012.

(129) Duncan PW, Lai SM, van C, V, Huang L, Clausen D, Wallace D. Development of a comprehensive assessment toolbox for stroke. Clin Geriatr Med 1999; 15(4):885-915.

(130) Young FB, Weir CJ, Lees KR. Comparison of the National Institutes of Health Stroke Scale with disability outcome measures in acute stroke trials. Stroke 2005; 36(10):2187-2192.

(131) Lai SM, Duncan PW. Stroke recovery profile and the Modified Rankin assessment. Neuroepidemiology 2001; 20(1):26-30.

(132) Post B, Merkus MP, de Bie RM, de Haan RJ, Speelman JD. Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? Mov Disord 2005; 20(12):1577-1584.

(133) Frey JL, Jahnke HK, Goslar PW, Partovi S, Flaster MS. tPA by telephone: extending the benefits of a comprehensive stroke center. Neurology 2005; 64(1):154-156.

(134) Weimar C, Kurth T, Kraywinkel K, Wagner M, Busse O, Haberl RL et al. Assessment of functioning and disability after ischemic stroke. Stroke 2002; 33(8):2053-2059.

(135) Walker MF. Stroke rehabilitation: evidence-based or evidence-tinged? J Rehabil Med 2007; 39(3):193-197.

(136) Bath PM, Gray LJ, Collier T, Pocock S, Carpenter J. Can we improve the statistical analysis of stroke trials? Statistical reanalysis of functional outcomes in stroke trials. Stroke 2007; 38(6):1911-1915.

(137) 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? J Neurol 2008; 255(6):867-874.

(138) Juttler E, Schwab S, Schmiedek P, Unterberg A, Hennerici M, Woitzik J et al. Decompressive Surgery for the Treatment of Malignant Infarction of the Middle Cerebral Artery (DESTINY): a randomized, controlled trial. Stroke 2007; 38(9):2518-2525.

(139) Murray S, Bashir K, Lees KR, Muir K, MacAlpine C, Roberts M et al. Epidemiological aspects of referral to TIA clinics in Glasgow. Scott Med J 2007; 52(1):4-8.

(140) Saver JL, Gornbein J. Treatment effects for which shift or binary analyses are advantageous in acute stroke trials. Neurology 2008.

(141) Saver JL, Kidwell CS, Liebeskind DS, Starkman S. Acute ischemic stroke trials. Stroke 2001; 32(1):275-278.

(142) Sanossian N, Ohanian AG, Saver JL, Kim LI, Ovbiagele B. Frequency and determinants of nonpublication of research in the stroke literature. Stroke 2006; 37(10):2588-2592.

(143) Gladstone DJ, Danells CJ, Black SE. The fugl-meyer assessment of motor recovery after stroke: a critical review of its measurement properties. Neurorehabil Neural Repair 2002; 16(3):232-240.

(144) Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. Ann Intern Med 1983; 99(6):843-848.

(145) D'Olhaberriague L, Litvan I, Mitsias P, Mansbach HH. A reappraisal of reliability and validity studies in stroke. Stroke 1996; 27(12):2331-2336.

(146) Burleigh E, Reeves I, McAlpine C, Davie J. Can doctors predict patients' abbreviated mental test scores. Age Ageing 2002; 31(4):303-306.

(147) Berge E, Fjaertoft H, Indredavik B, Sandset PM. Validity and reliability of simple questions in assessing short- and long-term outcome in Norwegian stroke patients. Cerebrovasc Dis 2001; 11(4):305-310.

(148) Lees KR, Hankey GJ, Hacke W. Design of future acute-stroke treatment trials. Lancet Neurol 2003; 2(1):54-61.

(149) Garraway WM, Akhtar AJ, Gore SM, Prescott RJ, Smith RG. Observer variation in the clinical assessment of stroke. Age Ageing 1976; 5(4):233-240.

(150) Lyden PD, Shuaib A, Lees KR, Davalos A, Davis SM, Diener HC et al. Safety and tolerability of NXY-059 for acute intracerebral hemorrhage: the CHANT Trial. Stroke 2007; 38(8):2262-2269.

(151) Hacke W, Kaste M, Fieschi C, von Kummer R, Davalos A, Meier D et al. 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; 352(9136):1245-1251.

(152) Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappraisal. Journal of Clinical Epidemiology 1988; 41:59-68.

(153) Richards SH, Peters TJ, Coast J, Gunnell DJ, Darlow MA, Pounsford J. Inter-rater reliability of the Barthel ADL index: how does a researcher compare to a nurse? Clin Rehabil 2000; 14(1):72-78.

(154) Hylek EM, Go AS, Chang Y, Jensvold NG, Henault LE, Selby JV et al. Effect of intensity of oral anticoagulation on stroke severity and mortality in atrial fibrillation. N Engl J Med 2003; 349(11):1019-1026.

(155) Kasner SE, Chalela JA, Luciano JM, Cucchiara BL, Raps EC, McGarvey ML et al. Reliability and validity of estimating the NIH stroke scale score from medical records. Stroke 1999; 30(8):1534-1537.

(156) Stavem K, Lossius M, Ronning OM. Reliability and validity of the Canadian Neurological Scale in retrospective assessment of initial stroke severity. Cerebrovasc Dis 2003; 16(3):286-291.

(157) Barber M, Fail M, Shields M, Stott DJ, Langhorne P. Validity and reliability of estimating the scandinavian stroke scale score from medical records. Cerebrovasc Dis 2004; 17(2-3):224-227.

(158) Ninomiya T, Donnan G, Anderson N, Bladin C, Chambers B, Gordon G et al. Effects of the end point adjudication process on the results of the Perindopril Protection Against Recurrent Stroke Study (PROGRESS). Stroke 2009; 40(6):2111-2115.

(159) Granger CB, Vogel V, Cummings SR, Held P, Fiedorek F, Lawrence M et al. Do we need to adjudicate major clinical events? Clin Trials 2008; 5(1):56-60.

(160) Adams HP, Jr., Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke 1993; 24(1):35-41.

(161) Mahaffey KW, Harrington RA, Akkerhuis M, Kleiman NS, Berdan LG, Crenshaw BS et al. Disagreements between central clinical events committee and site investigator assessments of myocardial infarction endpoints in an international clinical trial: review of the PURSUIT study. Curr Control Trials Cardiovasc Med 2001; 2(4):187-194.

(162) Mahaffey KW, Harrington RA, Akkerhuis M, Kleiman NS, Berdan LG, Crenshaw BS et al. Systematic adjudication of myocardial infarction end-points in an international clinical trial. Curr Control Trials Cardiovasc Med 2001; 2(4):180-186.

(163) Franke RH, Kaul JD. The Hawthorne experiments: First statistical interpretation. Am Soc Rev 1978; 43:623-643.

(164) Pettigrew LE, Wilson JT, Teasdale GM. Reliability of ratings on the Glasgow Outcome Scales from in-person and telephone structured interviews. J Head Trauma Rehabil 2003; 18(3):252-258.

(165) Shafqat S, Kvedar JC, Guanci MM, Chang Y, Schwamm LH. Role for telemedicine in acute stroke. Feasibility and reliability of remote administration of the NIH stroke scale. Stroke 1999; 30(10):2141-2145.

(166) von Koch L, Pedro-Cuesta J, Kostulas V, Almazan J, Widen HL. Randomized controlled trial of rehabilitation at home after stroke: one-year follow-up of patient outcome, resource use and cost. Cerebrovasc Dis 2001; 12(2):131-138.

(167) Massucci M, Perdon L, Agosti M, Celani MG, Righetti E, Recupero E et al. Prognostic factors of activity limitation and discharge destination after stroke rehabilitation. Am J Phys Med Rehabil 2006; 85(12):963-970.

(168) Leees KR, Asplund K, Carolei A, Davis SM, Diener HC, Kaste M et al. Glycine antagonist (gavestinel) in neuroprotection (GAIN International) in patients with acute stroke: a randomised controlled trial. GAIN International Investigators. Lancet 2000; 355(9219):1949-1954.

(169) Huybrechts KF, Caro JJ. The Barthel Index and modified Rankin Scale as prognostic tools for long-term outcomes after stroke: a qualitative review of the literature. Curr Med Res Opin 2007; 23(7):1627-1636.

(170) Stineman MG, Ross RN, Hamilton BB, Maislin G, Bates B, Granger CV et al. Inpatient rehabilitation after stroke: a comparison of lengths of stay and outcomes in the Veterans Affairs and non-Veterans Affairs health care system. Med Care 2001; 39(2):123-137.

(171) Lindley RI. John Rankin (1923-1981). J Neurol 2001; 248(11):1007-1008.

(172) Hull A. Hector's house: Sir Hector Hetherington and the academicization of Glasgow Hospital Medicine before the NHS. Med Hist 2001; 45(2):207-242.

(173) Watt OM. Stobhill Hospital; the First Seventy Years. 1 ed. Glasgow: Robert Maclehouse Publishers, 1971.

(174) RANKIN J. Cerebral vascular accidents in patients over the age of 60. I. General considerations. Scott Med J 1957; 2(4):127-136.

(175) Benton A, Tranel D. HIstorical notes on reorganisation of function and neuroplasticity. Cerebral reorganisation of function after brain damage 2000; (Editors Levin HS, Grafman J):3-23.

(176) Langhorne P, Pollock A. What are the components of effective stroke unit care? Age Ageing 2002; 31(5):365-371.

(177) RANKIN J. The use of skin traction in the treatment of recent hemiplegia. Scott Med J 1957; 2(9):366-367.

(178) Dickie HA, Rankin J. The lung's response to inhaled organic dusts. Arch Environ Health 1967; 15(2):139-140.

(179) RANKIN J, JAESCHKE WH, CALLIES QC, Dickie HA. Farmer's lung: physiopathologic features of the acute interstitial granulomatous pneumonitis of agricultural workers. Ann Intern Med 1962; 57:606-626.

(180) Rankin J, Kobayashi M, Barbee RA, Dickie HA. Pulmonary granulomatoses due to inhaled organic antigens. Med Clin North Am 1967; 51(2):459-482.

(181) Dickie HA, Kabler JD, Peters HA, Dempsey J. Memorial Resolution of the Faculty of University of Wisconsin on the Death of John Rankin. University of Wisconsin (Madison) Faculty Document 1981; 44:1-5.

(182) RANKIN J. Cerebral vascular accidents in patients over the age of 60. III. Diagnosis and treatment. Scott Med J 1957; 2(6):254-268.

(183) Scott PA, Silbergleit R. Misdiagnosis of stroke in tissue plasminogen activator-treated patients: characteristics and outcomes. Ann Emerg Med 2003; 42(5):611-618.

(184) Pendlebury ST, Rothwell PM, Algra A, Ariesen MJ, Bakac G, Czlonkowska A et al. Underfunding of stroke research: a Europe-wide problem. Stroke 2004; 35(10):2368-2371.

(185) Rudd AG, Hoffman A, Down C, Pearson M, Lowe D. Access to stroke care in England, Wales and Northern Ireland: the effect of age, gender and weekend admission. Age Ageing 2007; 36(3):247-255.

(186) Di Carlo A, Lamassa M, Pracucci G, Basile AM, Trefoloni G, Vanni P et al. Stroke in the very old : clinical presentation and determinants of 3-month functional outcome: A European perspective. European BIOMED Study of Stroke Care Group. Stroke 1999; 30(11):2313-2319.

(187) United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: interim results. UK-TIA Study Group. Br Med J (Clin Res Ed) 1988; 296(6618):316-320.

(188) 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; 349(9065):1569-1581.

(189) Hacke W, Kaste M, Fieschi C, Toni D, Lesaffre E, von Kummer R et al. Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke. The European Cooperative Acute Stroke Study (ECASS). JAMA 1995; 274(13):1017-1025.

(190) Lees KR, Davalos A, Davis SM, Diener HC, Grotta J, Lyden P et al. Additional outcomes and subgroup analyses of NXY-059 for acute ischemic stroke in the SAINT I trial. Stroke 2006; 37(12):2970-2978.

Modifications to Thesis of Dr. T J Quinn:

"Improving outcome assessment for clinical trials in stroke"

Following the recommendations of internal and external examiners at the oral examination on 21st January 2010, the following changes have been made to the MD thesis:

1/ All use of first person plural has been removed in favour of first person singular.

2/ Additional references have been added, in certain areas reference is to an online electronic resource rather than traditional published periodical and these are described in the main body of the text.

3/ Grammatical, formatting and typographic errors have been corrected.

4/ Certain key phrases ("Clinimetrics"; "Functional outcome"; "Convenience sample") have been better defined with corresponding references to original published work.

5/ Figures 11 and 12 have been redrawn with a corresponding legend to better explain the data presented.