Assessment of Clinical Outcomes in Acute Stroke Trials

Lucy Roberts, MBChB; Carl Counsell, MRCP(UK)

Background and Purpose—Adequate outcome assessment is crucial to randomized trials. We wished to assess the types of outcomes used in acute stroke trials and the appropriateness of these outcomes and their analyses.

Methods—Acute stroke trials from the Cochrane Stroke Group’s database were included from 1955 to 1995 if they were published in full text in English. For each trial we collected year of publication, number of patients randomized, blinding of outcome assessment, the specific outcome instruments used, the statistical methods used for analysis, and the significance of the results. The validity and reliability of each outcome measure were assessed by review of the literature.

Results—Our study included 174 trials. Outcomes were assessed blindly in 69%. Death was recorded in only 76% of trials, impairment in 76%, disability in 42%, and handicap or quality of life in only 2%. Of the trials that measured impairment, 35% used a measure of established validity or reliability. For disability and handicap, the proportions with valid or reliable measures were 70% and 25%, respectively. Impairment and handicap measures were primarily analyzed as continuous variables, while disability was mainly analyzed as a dichotomous variable. Continuous data were usually analyzed with inappropriate parametric statistics. There was no relationship between the method of analysis, the type of outcome, and the statistical significance of results.

Conclusions—Most acute stroke trials up to 1995 have used clinical outcome measures that were inadequate in terms of their content, reliability, validity, blinded assessment, and statistical analysis. This has important implications for future stroke research. (Stroke. 1998;29:986-991.)

Key Words: clinical trials ■ outcome assessment ■ stroke, acute

A randomized controlled trial aims to assess the effects of treatment, but if the outcomes are measured inappropriately, the trial cannot provide reliable results.1-3 Clinical outcome can be classified into impairment (signs of underlying pathology), disability (the functional results of impairment), and handicap (the social impact of the disease).4 In this system, impairment is the least clinically relevant to the patient and handicap the most.4,5 Health-related quality of life may be even more relevant than handicap, but it has proved difficult to define.6,7 Since there is no simple relationship between impairment, disability, handicap, or health-related quality of life,8,9 outcome measures should only include items relating to a single level because a mixture of levels is conceptually confusing and difficult to interpret clinically.10 In general, the more relevant the outcome measure is to the patient, the more difficult it is to define and assess.10

To influence clinical practice, acute stroke trials should use measures of outcome that are relevant to patients. Thus, phase III trials ought to concentrate on measuring disability, handicap, or health-related quality of life. In contrast, small phase II trials are generally not designed to influence practice but aim to establish whether a treatment influences the disease process and therefore might have an important clinical benefit. In these trials it is often appropriate to concentrate on measuring impairments. However, phase II trials should also measure more clinically relevant outcomes because these results may help in designing later phase III trials (eg, by providing estimates of realistic treatment effects) and can also be included in ongoing systematic reviews.

Other types of outcomes also need to be considered in acute stroke trials. The numbers of deaths should always be reported because both doctors and patients need to know whether a treatment is associated with more deaths even if it decreases disability in survivors. Length of hospital stay measures the economic implications of an intervention,11 while outcomes in the caregiver may also be important after the patient has been discharged from the hospital. Given the variety of outcomes that can be measured in trials of treatments for acute stroke, the outcomes should be prespecified and classified as primary (assessing the main hypothesis) or secondary (assessing the secondary hypotheses) to avoid inappropriate “data dredging” or multiple subgroup analyses to find a spurious positive result.

Outcome measures should also be valid,5 reliable,5 sensitive to important clinical changes, assessed blindly to avoid measurement bias, and analyzed appropriately. Measures of impairment, disability, and handicap almost always consist of a mixture of categories designated by numbers to form an ordinal scale in which the gaps between the different points are not necessarily equal.12 Strictly speaking, it is therefore...
statistically incorrect to treat these numbers as though they represent values on a continuous scale. The statistical methods used should generally be nonparametric and suitable for ranked data. Analysis that categorizes results into two or more groups of patients with similar outcomes is perhaps even more appropriate.

Given the above considerations, we wished to assess how appropriate the measurement of outcomes has been in acute stroke trials in terms of the types of outcomes recorded, whether they were assessed blindly, their validity and reliability, and the method of analysis.

**Subjects and Methods**

**Trial Selection**
We included all completed randomized trials of interventions in the acute phase of stroke (ie, started within 30 days of onset) provided they were published in full in English before 1996. Trials were identified from the Cochrane Stroke Group’s comprehensive register of trials, which is continually updated by means of multiple overlapping search strategies (including electronic searching of MEDLINE and other databases and hand searching of journals, books, and reference lists).13 We limited ourselves to studies published before 1996 because, at the time of our study, we could not be sure that the database was complete for 1996.

**Data Collection**
The following data were collected on each trial: the year of publication (for trials with more than one report, the year of the main report was taken); the number of patients originally randomized; the type of intervention tested; the reported clinical outcome measures and whether they were categorized as primary or secondary (non-clinical outcomes such as biochemical measurements were excluded); the method of analysis of each outcome measure (continuous or categorical/dichotomous, parametric or nonparametric); the statistical significance of each trial (results with Ρ<0.05 were interpreted as significant); the duration of follow-up; and whether outcomes were assessed blinded to treatment allocation (if the report simply stated that the trial was “double-blind,” we assumed that the assessor was blind). Data were only collected on these criteria in relation to the whole study population at the end of follow-up. Subgroup analyses were ignored.

**Definitions for the Classification of Outcomes**
After data collection, each outcome was classified into one of the following categories: impairment, disability, handicap, death, place of residence at follow-up, length of hospital stay, side effects (recorded as “side effects,” “adverse events,” or “complications” in the original trial reports), recurrent vascular events (including deep venous thrombosis, myocardial infarction, and recurrent stroke), and assessments of the patient’s caregiver. Measures were classified as impairment, disability, and handicap scales according to the definitions of Wade, with the exception of the Rankin/Modified Rankin Scale, which was classified as a measure of disability rather than handicap.14 Quality of life measures were classified as measures of handicap because there were too few of them to consider separately and, in terms of importance to the patient, they are closer to handicap than impairment.

Impairment, disability, and handicap scales were further classified into those that had been shown to have some validity or reliability and those that were neither valid nor reliable. A rigorous appraisal of the validity and reliability of each measure was beyond the scope of this study. Instead, we identified previous reviews5,15–17 that discussed the properties of outcomes measures used in stroke research and followed their recommendations about which measures were valid and reliable. For those outcome measures not mentioned in these reviews, we tried to identify primary research studies that documented their validity or reliability. Unnamed measures that were not referenced in the original trial report were assumed to be untested unless otherwise stated. The classification of impairment, disability, and handicap scales is shown in Table 1 along with the references to support our decisions on validity and reliability.

All data were collected by the two authors independently and were cross-checked. Disagreements were resolved by consensus. We did not formally measure interobserver variability.

**Analysis**
The data were entered onto a computer database (DBase IV). Percentages and the corresponding 95% confidence intervals were
calculated with the Confidence Interval Analysis Program, and differences in proportions were assessed with the $\chi^2$ test in Epi Info (version 5.01b).

Results

A total of 174 of 374 acute stroke trials on the database fulfilled the selection criteria, most of which were published from 1975 onward. Trial size ranged from 8 to 1267 patients (median, 68 patients; interquartile range, 37 to 159 patients), and the duration of follow-up ranged from 1 or 2 days to more than a year (median, 12 weeks; interquartile range, 3 to 24 weeks). The most frequently tested interventions were hemodilution (18 trials), anticoagulants (18 trials), calcium antagonists (18 trials), corticosteroids (17 trials), and thrombolytic agents (15 trials). In most studies it was impossible to tell which was the primary and which were the secondary outcomes, and therefore all outcomes were considered together. In 69% of trials the outcome was assessed blindly, 6% were not blinded, and in 25% the blinding of outcome assessment was not clearly described in the published report.

Assessment of the Type of Outcomes

Table 2 demonstrates the number of trials assessing each of the clinical outcomes considered. A substantial minority of trials did not report data on death (24%) or impairment (24%), while the majority of trials did not report data on disability or handicap. Ten trials (6%) measured only biochemical markers and did not assess death, impairment, disability, or handicap. The proportion of trials measuring impairment was significantly higher than that measuring disability (76% versus 42%; $P=0.00001$). The most common impairment measures used were a trialist’s own scale (eg, signs better, worse, or unchanged) (47/174 trials, 27%), the Mathew/Modified Mathew Scale (28/174, 16%), the National Institutes of Health Stroke Scale (11/174, 6%), the Scandinavian Stroke Scale (9/174, 5%), and the Toronto Stroke Score (7/174, 4%). For disability, the most common measures were the Barthel Index (37/174, 21%), the trialist’s own scale (20/174, 11%), and the Rankin or Modified Rankin Scale (15/174, 9%). Several trials used more than one measure of impairment or disability. Many of the measures were inadequate (Table 2): of the trials that measured impairment, only 35% (47/133) used a measure with proven reliability or validity, while only 70% of trials (51/73) measuring disability used an adequate measure. Most of the untested measures appeared to have been developed by the trialists themselves.

The relationship between the type of outcome measured and the size of the trial is demonstrated in Figure 1. Impairment was assessed in approximately 70% to 80% of trials regardless of size. Trials with more than 100 patients were significantly more likely to assess disability than smaller ones (60% versus 25%; $\chi^2=21.5, df=3, P=0.00008$), while larger trials were also more likely to report deaths ($\chi^2=18.15, df=3, P=0.0004$); approximately 80% of trials with more than 50 patients reported deaths compared with 60% of smaller trials. Figure 2 shows the changes in the proportions of trials measuring each type of outcome over time. There was a trend for more trials to record disability over time ($P=0.16$) and for fewer to report death ($P=0.1$), but neither of these was significant. However, the proportion of trials measuring impairment did increase over time ($\chi^2=21.42, df=7, P=0.003$), with fewer trials measuring it before 1964 compared with after (20% versus 80%).

Statistical Analysis of Outcomes

Table 3 shows how the main outcome measures were analyzed. Two trials did not analyze the results for impairment...
despite having recorded it, and one did not analyze disability. Impairment and handicap measures were primarily analyzed only as continuous data, while a greater proportion of disability measures were analyzed as dichotomous data (ie, patients were separated into those who were dead or dependent and those alive and independent). Continuous data were mainly analyzed with parametric statistical methods. Most of the trials that analyzed dichotomous data simply reported the percentages with a good or bad outcome in each group rather than an odds ratio or relative risk. Of the 10 trials that did report a relative risk, seven were published in 1995.

Twelve trials assessed disability with the Barthel Index and analyzed it as a dichotomous variable, but there was no standardization in the cutoff point used to define a good outcome (ie, which patients were independent). Five trials defined a good outcome as a Barthel Index of 12 or more of 20 (or 60 of 100); 4 trials as a Barthel Index of more than 14 (70 of 100); 1 trial as a Barthel Index of 18 or more (90 of 100); 1 trial as a Barthel Index of 19 or more (95 of 100); and 1 trial as an improvement in the Barthel Index from the baseline score.

Measures analyzed as continuous data were not more likely to report statistically significant results than those analyzed as dichotomous data (impairment: $\chi^2=3.3, P=0.07$; disability: $\chi^2=0.01, P=0.9$). Similarly, for continuous variables there was no significant difference between the proportion of significant results obtained with parametric and nonparametric methods (impairment: $\chi^2=1.27, P=0.53$; disability: $\chi^2=0.54, P=0.76$). There was also no significant difference between the proportion of trials with significant results when impairment was the outcome of interest compared with when disability was analyzed.

**Discussion**

This study has shown that few trials in acute stroke have measured the outcomes of most relevance to the patient. Death was measured in most trials, but given the importance and robustness of this outcome it should have been reported in all trials. Most trials also measured impairment, which has the advantage of being easy to measure and is relatively objective but has limited clinical relevance. The patient’s state is more than merely the sum of his or her signs. Handicap and quality of life are much more important to the patient, but such measures are more difficult to define, more subjective, and more difficult to validate. It is therefore not surprising that only four trials recorded these and, of these, two used measures of handicap that could be regarded as measuring mainly disability (the Frenchay and Viitanen scales). Assessment of disability has been suggested to be the most feasible compromise, but only 40% of trials assessed this.

Historically, impairment and disability measures were introduced into stroke trials about the same time (the early 1960s), but whereas the use of impairment rapidly increased,
the measurement of disability has increased only gradually. This may have been because there were so many measures of impairment or because impairment measures were perceived to be easier to use. However, several simple disability measures such as the Rankin or Barthel scales have existed for many years. Many of the trials may have been phase II studies and therefore concentrated on impairment to demonstrate some effect on the disease process. This is partly supported by the fact that larger trials were more likely to measure disability. However, few trials were formally called phase II studies, and even phase II studies should report deaths, but 25% of trials did not even do this. Moreover, only 60% of trials with more than 300 patients reported disability. Detailed measurement of impairment may also be more likely to show small but statistically significant effects that may be important to detect in exploratory phase II trials. However, we did not find that trials measuring impairment were more likely to show statistically significant results than those measuring disability. There was little evidence to suggest that reporting of important outcomes had improved over time.

Any measure used to assess outcome, whether impairment or disability, should be shown to be valid and reliable, and yet many of those used in stroke trials were not. Many trialists used their own measure of impairment or disability rather than an established and tested one. The statistical analysis of outcomes recorded on categorical or ordinal scales was also extremely poor. Most were analyzed as continuous data, but this has inherent problems. A difference of a few points in a continuous score between the treated and control groups has little clinical meaning. If such a scoring system is used, the scoring of death is also problematic (should death merely be assigned the worst possible score, or is it worse than that?). In addition, most trials analyzed continuous data with parametric statistics, which is invalid if the trial is small and the distribution is skewed. This was the case in many of the trials, particularly those in which death was scored as zero or negative on a scale.

Categorical analysis is much more appropriate because clinical meaning can usually be applied to different categories, eg, disability scores can be divided into “independent in activities of daily living” or “dependent.” Death can be added as a separate category or combined with another category (eg, “dead or dependent”). However, problems also existed in the reporting of results of measures analyzed appropriately as two categories. Very few of these trials reported a measure of relative treatment effect such as relative risk or odds ratio. In addition, there was no standardization in the point at which disability scales, such as the Barthel Index, were split into different categories. Comparison of these outcomes between different trials, for example in meta-analysis, therefore becomes difficult. Reassuringly, the method of analysis (continuous versus categorical, parametric versus nonparametric) was not associated with the proportion of statistically significant results. If it had been, it might have suggested that trialists were using inappropriate statistics to increase the significance of their findings.

We found other problems with the process and reporting of outcome assessment. Only two thirds of the trials stated that outcomes were blindly assessed, yet unblinded assessments may be biased. Even in so-called double-blind trials, the outcome assessor may in fact not have been blind, particularly if the assessor was involved in the care of the patient and the treatment was associated with particular side effects. It was usually impossible to determine from the reports which were the primary and secondary outcomes. More emphasis should be given to the results of the prespecified primary outcomes because these are the ones the trial was specifically designed to test. Secondary outcomes may be less reliable, especially if multiple outcomes are measured and only the statistically significant ones are reported. Finally, the period of follow-up was often too short to assess the full impact of treatment on the final status of the patient.

Our study has some weaknesses. First, we excluded non–English language trials, and therefore our results may not be generalizable. However, there is little evidence to suggest that non–English language trials are any better or worse than English language trials. We also did not include trials from 1996 onward, and therefore it is possible that the assessment of outcomes has improved recently. However, given the scale of the problems we found, it is unlikely that they will have been resolved since 1995. Certainly, the trends shown in Figure 2 do not suggest a dramatic improvement in the type of outcome assessed over time. Our classification of reliability and validity of outcome measures was usually only defined by previous authors’ conclusions. This was the only feasible way to assess these, however, since it is difficult to set uniform, rigorous criteria for validity and reliability. We also defined adequate outcome measures as those with proven validity or reliability and not both. In addition, one of the disability measures defined as reliable (the Glasgow Outcome Scale, which was used in three trials) has been validated in head injury but not in stroke. We may therefore have been overoptimistic in our assessment of reliability and validity. Finally, other important characteristics of outcome measures were not considered, such as sensitivity to change and feasibility. Measures that take several hours to complete are not feasible for large trials.

This study highlights several improvements that should be made in future acute stroke trials. Outcome measures should be relevant to the patient, concentrating on disability or handicap, and deaths must be reported. Even small exploratory phase II trials should assess these outcomes because the results can then be included in meta-analyses. Another important factors such as side effects, length of hospital stay, and assessment of the caregiver should also be considered. The outcome measures must be well established with proven validity and reliability and should be blindly assessed. Statistical analyses should be appropriate for ranked data or, better still, involve division of patients into clinically meaningful groups, eg, those alive and independent compared with those dead or dependent. The definitions of dependency in existing disability scales need to be standardized to facilitate comparison between trials in meta-analyses.

Acknowledgments
This study was supported by a Wellcome Trust research training fellowship in clinical epidemiology (to Dr Counsell). We thank Dr...
Peter Sandercock and the two anonymous reviewers for their comments on the manuscript.

References

Assessment of Clinical Outcomes in Acute Stroke Trials
Lucy Roberts and Carl Counsell

Stroke. 1998;29:986-991
doi: 10.1161/01.STR.29.5.986
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1998 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/29/5/986

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at: http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at: http://stroke.ahajournals.org/subscriptions/