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)

Mark A. Albanese, PhD; William R. Clarke, PhD; Harold P. Adams, Jr, MD; Robert F. Woolson, PhD; and TOAST Investigators

Background and Purpose Ensuring the reliability and validity of outcome measures used in clinical trials is essential to the success of the trial. The Trial of Org 10172 in Acute Stroke Treatment (TOAST) is a multicenter clinical trial that is recruiting patients with acute ischemic stroke seen at medical centers across the United States.

Methods This paper describes an approach to train physicians to use three clinical measures: the National Institutes of Health (NIH) Stroke Scale, a supplemental motor examination, and the Glasgow Outcome Scale. The program included education, certification, remediation when needed, monitoring, and reliability assessment. The goal was to ensure that interrater assessments were as equivalent to one another as possible.

Results Of the first 95 clinicians who began the certification process, 75 passed during the first evaluation. Eighteen of the other physicians were able to complete the process after remediation. The intraclass correlations of both the NIH Stroke Scale and supplemental motor examination exceeded 0.95. The values for the Glasgow Outcome Scale were 0.61 and 0.62 for the first and second ratings of the videotape, respectively.

Conclusions Our experience suggests that a program that includes educational and certification processes can be performed as part of the design of a multicenter clinical trial. The method of providing educational and testing videotapes to each site so that physicians can be trained and certified is an effective, inexpensive, and practical approach for enhancing and certifying the expertise of the large number of physicians involved in a multicenter study. (Stroke. 1994;25:1746-1751.)

Key Words • clinical trials • stroke assessment • stroke outcome

A clinical trial testing alternative treatments for acute ischemic stroke needs to ensure the quality of all outcome data. Even relatively objective data such as blood concentrations of drug are dependent on the method of determination and have been inconsistent when sampled at different times. The problem becomes even more complex when assessments depend on judgments based on clinical observations. Differences in expertise and other characteristics specific to an observer can interfere with the accuracy of such ratings. For example, Azen et al reported that interrater agreement (value) among experienced ophthalmologists who independently rated photographs of the ocular fundus for proliferative retinopathy ranged from 0.41 to 0.53 on a scale of 0 to 1.0. We previously reported that interphysician agreement in the diagnosis of a subtype of ischemic stroke can be difficult to achieve even when all physicians are reviewing identical data.

While a value of 0.4 to 0.6 might be considered "moderate" agreement, it reflects far less than perfect consistency of interpretation between experts.

Keeping such inconsistencies to a minimum can be highly advantageous in the performance of a clinical trial. Bellamy et al developed a program to standardize use of a battery of assessments for osteoarthritis among rheumatologists. The program included a 2-hour orientation session, evaluation of six patients using the rating instruments, a meeting to discuss and resolve discrepancies, and another examination of the six patients. This process resulted in marked increases in consistency among raters. The authors concluded that such a rigorous training process might allow the reduction of sample sizes for clinical trials by as much as 50% without any lessening of statistical power.

The key outcome measures of patients with stroke are clinical, including reduction in mortality and a better quality of life among survivors. The benchmarks of these outcome measures are global judgments of disability, neurological impairment, and the extent to which the patient is able to carry on with life as before stroke. Determinations of the validity and reliability of judgments of these outcomes are important if the results of a trial are to be credible. Similarly, because of the large number of patients needed to test any intervention for acute ischemic stroke, a multicenter project that may span a continent and last several years is usually man-
dated. In this context, establishing and maintaining interrater agreement across sites and over time as examiners are replaced from attrition is critical.

The purposes of this article are to (1) describe the program used for training physicians to use the clinical measures (National Institutes of Health [NIH] Stroke Scale, supplemental motor examination, and Glasgow Outcome Scale) for the clinical trial of stroke medication ORG 10172, (2) report reliability data from its first year of operation, (3) compare the reliability results with findings from other smaller-scale studies that used approaches that would be impractical on a larger scale, and (4) discuss the feasibility of using this type of reliability program in future clinical trials.

Methods

The Trial of ORG 10172 in Acute Stroke Treatment (TOAST) is a multicenter clinical trial that is recruiting patients with acute ischemic stroke at several medical centers across the United States. The goal of the trial is to test the utility of ORG 10172 in improving outcome after acute ischemic stroke. ORG 10172 is a low-molecular-weight non-heparin-containing glycosaminoglycan that has selective antithrombotic effects. As clinical measures, TOAST uses a modified Barthel Index, NIH Stroke Scale, a supplemental motor examination, and the Glasgow Outcome Scale.

Instruments

A modified Barthel Index, Glasgow Outcome Scale, and NIH Stroke Scale were selected based on an extensive review of the literature and considerations of the expected effect of administering ORG 10172. The selection of these rating instruments was aimed at providing data on global outcome (Glasgow Outcome Scale), information about disability (modified Barthel Index), and neurological impairments (NIH Stroke Scale and supplemental motor examination). These are the data that Candelisa etc. concluded are the true benchmarks for efficacy in patients with acute stroke. The scales were selected also because of their simplicity and ease in administration.

The modified Barthel Index is a 10-item checklist that is given to the patient or to a person who is aware of the patient's living conditions. It can be performed during a face-to-face interview or through a telephone call. It asks straightforward questions about the patient's ability to perform activities of daily living such as going to the bathroom or dressing. Because it is an interview with specified alternatives, it requires no special expertise to complete. Little training is needed beyond familiarizing personnel with the questions and the criteria for each response.

For this reason, it was not included as part of the reliability studies in the trial and will not be considered further.

The Glasgow Outcome Scale provides a global rating of the patient's status. It is a single item that has five possible classifications of patient status: (1) no or minimal disability or handicap, (2) moderate disability, (3) severe disability, (4) persistent vegetative state, and (5) death. Although this rating instrument was developed to rank outcomes after head injuries, it also has been used in trials of interventions for stroke. It is very well known by clinicians, and it provides a global classification of the patient's status. Although it may not be as sensitive as other scales, it does differentiate clinically important differences in outcome. Wade etc. recently concluded that it is a reasonable measure to estimate outcomes among patients with acute neurological diseases. While each of the grades in the Glasgow Outcome Scale has been defined, it still requires physician interpretation; thus, a trial should include methods to teach physicians on the discriminating features for each grade and to ensure interrater agreement.

The NIH Stroke Scale quantifies the neurological deficits commonly found in patients with stroke. It consists of 15 independently scored items, such as level of consciousness, motor responses, or language. Studies of the reliability and validity of the scale have shown interrater agreement ($\kappa$ value) on the various items to range from 0.45 to 0.95. 11 The scores on the baseline administration of the rating instrument have also been correlated with computed tomographic measurements of lesion volume at 7 days, a finding that demonstrates the concurrent validity and the predictive validity of the scale. Yet, physicians who use this scale must be familiar with the subtleties of each item being examined.

Because many patients with stroke may have different degrees of weakness between proximal and distal limb muscles and because the NIH Stroke Scale does not include independent assessments of proximal and distal motor function, the supplemental motor examination also was included in the TOAST study. This eight-item scale has been used in other clinical trials in stroke. It scores six grades of movement from normal to none at both shoulders, wrists, hips, and ankles.

Reliability Studies Design

The plan for the reliability studies for the NIH Stroke Scale, supplemental motor examination, and Glasgow Outcome Scale consisted of education, certification, remediation when needed, monitoring, and reliability assessment.

Education

The educational program consisted of an instructional manual with detailed definitions and directions for administration of the three rating tools and an instructional videotape. The videotape included a physician's examination of three patients with recent stroke who had a broad range of impairments. Our goal was to train physicians to use the TOAST study. This eight-item scale has been used in other clinical trials in stroke. It scores six grades of movement from normal to none at both shoulders, wrists, hips, and ankles.

Certification and Remediation

Physicians were required to demonstrate proficiency in the use of the scales before they could examine patients enrolled in the trial. The certification process consisted of viewing and rating a videotape of six patients being examined using the NIH Stroke Scale, supplemental motor examination, and Glasgow Outcome Scale. The physician disclosed his specific scoring for each item after a short period when the observer was allowed to complete his/her own rating. This allowed the observers at each center to verify their own ratings using the common standard score as a reference. The tape could be repeated, in whole or in part, as the trainee desired.

Whereas the certification process was unambiguous, setting the standard of performance was complicated. The standard was based on the scores of the physician who examined the patient. Because the scores were obtained during the examination of a live patient and might be influenced by features that were not obvious on the videotape, the scores were adjusted by the marks made by other physicians at the clinical coordinating center who independently evaluated each item on the three rating instruments while viewing the tapes. After reviewing their differences in scoring, the physicians decided what were acceptable deviations from the standard.

Criteria for unacceptable deviations in scoring the three rating instruments were: (1) any total score derived from the
three different rating instruments that deviated by four or more points in absolute value from the standard on any of the six patients, (2) scores of individual items that deviated by three or more points in absolute value from the standard on any of the six patients, (3) any individual item scores that deviated from the standard by any amount for three or more of the six patients, or (4) Glasgow Outcome Scale scores that deviated from the standard by any amount for two or more of the six patients.

The individual physician's ratings were compared with the standard using a form that displayed the four criteria for unacceptable deviations. Deviations in excess of an absolute value of more than three points in any of the six patients (criterion 2) were highlighted for discussion. Any physician whose ratings had not met any of the criteria for unacceptable deviations was certified.

Reliability Assessment

As part of the data collection and review process, the reliability and reproducibility of scores were determined. Interrater agreement and reproducibility were calculated by comparing each physician's scores with those of others. Intrarater reproducibility reflected score drift across time and, as described above, was measured by comparing two sets of scores of the six patients that were obtained approximately 1 to 3 weeks apart. The second ratings were used to determine the stability of ratings over time and were not part of the certification process. Reliabilities for the overall scores derived from the NIH Stroke Scale and the supplemental motor examination by summing individual item ratings were intraclass correlations. The values computed were estimates of the reliability of a single rater evaluating a single patient. The reliabilities of scores for individual items were computed using the $k$ value, a statistic. The intraclass correlations of both the NIH Stroke Scale and supplemental motor examination exceeded 0.95. The $k$ value for the Glasgow Outcome Scale is based on the $k$ statistic. The intraclass correlations of both the NIH Stroke Scale and supplemental motor examination exceeded 0.95. The $k$ value for the Glasgow Outcome Scale was 0.62. These values meet most definitions of acceptable reliabilities for research purposes, although higher values for the Glasgow Outcome Scale would be desirable.

We also compared our data with those obtained in earlier reliability studies, which were based on examination of live patients. Table 2 compares the $k$ values for each item on the NIH Stroke Scale in the present study with those for live patients reported by Brott et al. In general, the reliabilities from Goldstein et al were lower than ours, which in turn were lower than those of Brott et al. The items showing the consistently highest reliabilities across the three studies were assessments of the right arm and right leg. Those showing the consistently lowest reliabilities were assessments of facial movement, limb ataxia, and dysarthria.

Discussion

Our study suggests that a program that includes educational and certification processes is feasible for ensuring the quality of data collected as part of the design of a multicenter clinical trial. Furthermore, by providing educational and testing videotapes to each site, physicians can be trained and certified at their convenience. This is a relatively inexpensive and practical approach to certifying the expertise of the large number of physicians involved in a multicenter study. The results shown in Table 2 suggest that the rating of videotaped examinations can achieve a degree of reliability that is comparable to the live clinical setting.

We concluded that the physicians in TOAST are generally well prepared to evaluate the videotaped patients using the protocols for the NIH Stroke Scale, supplemental motor examination, and Glasgow Outcome Scale. Because this was not an experimental study

<table>
<thead>
<tr>
<th>Scale</th>
<th>No. of Items</th>
<th>Standard Error</th>
<th>Interrater Reliability</th>
<th>Intrarater Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH Stroke Scale</td>
<td>15</td>
<td>1.81</td>
<td>.96</td>
<td>.97</td>
</tr>
<tr>
<td>Supplemental motor examination</td>
<td>8</td>
<td>1.29</td>
<td>.98</td>
<td>.98</td>
</tr>
<tr>
<td>Glasgow Outcome Scale</td>
<td>1</td>
<td>...</td>
<td>.62</td>
<td>.75</td>
</tr>
</tbody>
</table>

NIH indicates National Institutes of Health.

*Standard error reflects the amount of variation in a single patient's score that could be expected on repeated assessments.

†Intraclass correlations for NIH Stroke Scale and supplemental motor examination, $k$ values for Glasgow Outcome Scale.

Results

Compliance by physicians with this demanding protocol was very good. Of the first 95 clinicians who began the certification process, 75 passed during the first evaluation. Eighteen of the other physicians were able to complete the process after remediation. While compliance was good, it required a substantial amount of staff time to remind the clinicians to complete the process. Reliability data are reported for only the 75 physicians who had a complete set of ratings. (Physicians who failed to be certified on the first testing were omitted because their second evaluations of the entire videotape were not in the same time frames as those who were certified on their first attempt.)

Table 1 shows the results of the reliability assessments for the three rating instruments. The values for the NIH Stroke Scale and supplemental motor examination are intraclass correlations based on summing all of the items on the respective instruments. The values for the Glasgow Outcome Scale are based on the $k$ statistic. The intraclass correlations of both the NIH Stroke Scale and supplemental motor examination exceeded 0.95. The $k$ value for the Glasgow Outcome Scale was 0.62. These values meet most definitions of acceptable reliabilities for research purposes, although higher values for the Glasgow Outcome Scale would be desirable.
employing a control group, we cannot say that the cause of the physicians' preparedness was our training program. However, it is likely that at least some of the physicians benefited from the training program. Furthermore, the results reported by Bellamy et al. document the potential gains in rater reliability to be had by use of a rigorous training program. Our intent was simply to develop a feasible approach for extending this concept to a large-scale clinical trial.

The lower values of $k$ for the Glasgow Outcome Scale are troubling, but given that the scale consists of only a single five-option item, values in this range are not unexpected. Furthermore, the six patients included in the videotape were all representatives of the middle three values; obviously, no patients in grade 5 (dead) were included, and no patients without deficits were videotaped. Thus, our reliability assessment focused on the three ratings that would be the most difficult to discriminate. Therefore, we believe the determinations of outcomes using the Glasgow Outcome Scale by physicians in our study will be accurate within tolerance.

The reliabilities of scoring of the NIH Stroke Scale in the study of Goldstein et al. (Table 2) tended to be lower than ours, which in turn tended to be lower than those reported by Brott et al. Although the results led to comparable conclusions, the three studies have significant differences in design. The data reported by Brott et al. were collected from a team of four healthcare professionals, including nonphysicians, who observed a neurologist examine each of 24 patients. The Goldstein et al. data were collected by four neurologists who were randomly paired to then independently rate each patient, up to a total of 20 patients. Our data were collected from 75 neurologists who viewed a videotape of single examinations of six patients. The differences in testing conditions may explain the variations in results among the three projects. Any transient progression of signs or fatigue may in part explain the lower agreement in the study of Goldstein et al. Conversely, being able to detect other cues during performance of the examination of an actual patient may explain the better results of Brott et al. Our use of a videotape may have partially explained some of our discrepancies on individual items.

The poorest reliability occurred in the determination of impairments in facial movement or limb ataxia. We initially attributed the low reliability for scoring of these two items to the use of the videotapes. However, Brott et al. and Goldstein et al. only "best arm" and "best leg" ratings were collected.

### Table 2. NIH Stroke Scale Item Reliabilities ($k$ Values)

<table>
<thead>
<tr>
<th>Item</th>
<th>Present Study</th>
<th>Earlier Studies*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interrater</td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Rating</td>
<td>Interrater</td>
</tr>
<tr>
<td></td>
<td>Second Rating</td>
<td></td>
</tr>
<tr>
<td>Level of consciousness</td>
<td>.52 .44</td>
<td>.75 .49 .50</td>
</tr>
<tr>
<td>Questions</td>
<td>.69 .65</td>
<td>.81 .80 .64</td>
</tr>
<tr>
<td>Commands</td>
<td>.77 .71</td>
<td>.87 .56 .41</td>
</tr>
<tr>
<td>Gaze</td>
<td>.74 .74</td>
<td>.84 .82 .33</td>
</tr>
<tr>
<td>Visual fields</td>
<td>.80 .80</td>
<td>.89 .81 .57</td>
</tr>
<tr>
<td>Facial movement</td>
<td>.30 .28</td>
<td>.59 .57 .22</td>
</tr>
<tr>
<td>Right arm</td>
<td>.96 .95</td>
<td>.97 .85 .77</td>
</tr>
<tr>
<td>Left arm</td>
<td>.77 .82</td>
<td>.86 † †</td>
</tr>
<tr>
<td>Right leg</td>
<td>.81 .79</td>
<td>.87 .83 .78</td>
</tr>
<tr>
<td>Left leg</td>
<td>.54 .49</td>
<td>.72 † †</td>
</tr>
<tr>
<td>Limb ataxia</td>
<td>.48 .47</td>
<td>.70 .57 -.16</td>
</tr>
<tr>
<td>Sensory</td>
<td>.58 .59</td>
<td>.77 .60 .50</td>
</tr>
<tr>
<td>Best language</td>
<td>.66 .65</td>
<td>.83 .64 .79</td>
</tr>
<tr>
<td>Dysarthria</td>
<td>.52 .51</td>
<td>.76 .55 .32</td>
</tr>
<tr>
<td>Neglect</td>
<td>.69 .66</td>
<td>.83 .58 .61</td>
</tr>
</tbody>
</table>

NIH indicates National Institutes of Health.

*The designs of the reliability studies and the methods used to estimate $k$ values were different between the three studies. Thus, differences in $k$ values may be due to either design issues or $k$ computational procedures.

In Brott et al. and Goldstein et al. only "best arm" and "best leg" ratings were collected.
continually reminded to score it even in the presence of weakness.

In summary, our study suggests that with minimal instruction, physicians can use the NIH Stroke Scale, supplemental motor examination, and Glasgow Outcome Scale to assess patients with stroke reliably and reproducibly. The possibility of using written materials and videotape to provide this instruction makes these methods well suited to large-scale clinical trials of stroke management.

Appendix

TOAST Participating Clinical Centers

University of Iowa Hospitals and Clinics, Iowa City, Iowa:
B.H. Bendixen, PhD, MD (Principal Investigator(s) [PI]). H.P. Adams, Jr, MD; R.H. Davis, MD; M.R. Jacoby, MD; F.J. Gomez, MD; M.E. Dyken, MD; E.Y. Uc, MD; J.M. Wojciech, MD; and L.J. Kappelle, MD (Co-Investigator(s) [Co-I]). A.B. Tanna, RN, and V.L. Mitchell, RN (Study Coordinator(s) [SC]).

St Louis University Medical Center, St Louis, Mo: C.R. Curtin, MD (PI). D.L. Maikoff, MD; R. Tulpyanochonette, MD; C.M. Sauer, MD; G. Riaz, MD; J.G. Schmidt, MD; and M.M. Malik, MD (Co-I). G.A. Banet, RN, MSN (SC).

Marshfield Clinic, Marshfield, Wis: P.N. Kankanja, MD, MRCPS (PI). K.P. Madden, MD; K.H. Ruggles, MD; S.F. Mickel, MD; P.G. Gottschalk, MD; P.L. Hansotia, MD; R.W. Sorensen, MD; D.M. Jacobson, MD; and B.C. Hiner, MD (Co-I). K. Mancl and E. Lukasik (SC).


Clinical Coordinating Center

University of Iowa Hospitals and Clinics, Iowa City, Iowa: H.P. Adams, Jr, MD (Project Director). B.H. Bendixen, PhD, MD; P.H. Davis, MD; and B.B. Love, MD (Medical Monitors). J.K. Grimson, RN (Center Coordinator). J.D. Olson, MD, PhD; and B.J. Pennel (Central Laboratory). K. Johnson, RPh (Central Pharmacy). S.H. Cornell, MD; D.L. Crosby, MD; and T.M. Simpson (Central Radiology). V. Krumbholz (Financial Administrative Assistant). C.R. Zalesky (Secretary).

Data Management Center

University of Iowa, Iowa City, Iowa: R.F. Woolson, PhD (Director). W.R. Clarke, PhD (Associate Director). P.A. Wasek, BA (Center Coordinator). J.A. Dieleman, BA (Systems Coordinator). J.M. Paulsen, BS, and J.P. Boreen, BS (Programmers). M.F. Jones, MA; B.M. Robb, BA; L.A. Oberbroeckling, BS; and M.D. Hansen, MS (Research Assistants). K.M. Hicklin (Secretary). Committees

Advisory Committee: M.L. Dyken, MD; R.F. Franksowski, PhD; C.S. Greenberg, MD; L.A. Harker, MD; and J.P. Whisnant, MD (Members). H.P. Adams, Jr, MD, and R.F. Woolson, PhD (Members Ex Officio). In-house Safety Committee: W.R. Clarke, PhD; R.W. Fincham, MD; T.C. Kisker, MD; J.D. Olson, MD; B.B. Love, MD; and R.F. Woolson, PhD. NIH Safety and Monitoring Committee: H.J. Day, MD; K.M.
Detre, MD, DrPH; J.C. Grotta, MD; E.C. Haley, Jr, MD; W.T. Longstreth, Jr, MD; and J.R. Marler, MD.

Acknowledgments

This study was sponsored by US Public Health Service grants NIH1-NINDS RO1-NS-27863 and NIH-NINDS RO1-NS-27960. Support, including supply of the study drug, was also provided by Organon Inc, West Orange, NJ.

References

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).

M A Albanese, W R Clarke, H P Adams, Jr and R F Woolson

*Stroke*. 1994;25:1746-1751
doi: 10.1161/01.STR.25.9.1746

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1994 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/25/9/1746

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/