Contemporary Outcome Measures in Acute Stroke Research : Choice of Primary Outcome Measure and Statistical Analysis of the Primary Outcome in Acute Stroke Trials

Barbara C. Tilley

Stroke 2012, 43:935-937: originally published online March 15, 2012
doi: 10.1161/STROKEAHA.111.644864
Stroke is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
 Print ISSN: 0039-2499. Online ISSN: 1524-4628

Contemporary Outcome Measures in Acute Stroke Research

Choice of Primary Outcome Measure and Statistical Analysis of the Primary Outcome in Acute Stroke Trials

Barbara C. Tilley, PhD

See related articles, p 1163 and 1171.

The 2 articles in this issue[1,2] are based on the recent European Stroke Organization Outcomes Workshop and provide an excellent basis for discussion of outcomes used in stroke trials. Bath, Lees, and colleagues raise important points to consider when choosing an outcome measure or an analysis plan. The authors speculate that trials to identify therapeutic benefits of effective drugs may fail to find benefits because of problems inherent in choice of outcome measures or methods of analysis. Early in their article, Bath and colleagues give examples of multiple ways to define an outcome and show that trial results vary based on the definitions used. The emphasis throughout the article is on increasing power to detect differences so effective treatments will not be missed. There are many ways to increase power, but not all lead to the identification of clinically meaningful differences. The authors give little mention to the possibility of a Type 1 error, that is, declaring an ineffective therapy effective. Type I errors can result from searches for outcomes and analytic approaches to increase the power to detect treatment differences after a trial is completed. It is as important to patients that ineffective treatments are kept out of clinical practice as it is to assure that effective treatments are not missed.

The authors recognize that 1 outcome and 1 analytic approach may not fit all trials. In choosing an outcome or approach, the most important consideration is the research question of interest. It is the trial hypothesis that should drive the choices. As an example, at a workshop to identify outcomes and an analytic approach for the National Institute of Neurological Disorders and Stroke Tissue Plasminogen Activator (NINDS tPA) Stroke Trial,[3] Thomas Brott stated that he would like "an outcome that tells me if my mother is still my mother." By implication, the question of interest became "Does treatment with recombinant tissue-type plasminogen activator result in recovery with minimal or no disability?" The workshop participants determined that the latter was an appropriate question given the risk of cerebral hemorrhage and the expected benefit of recombinant tissue-type plasminogen activator. The workshop participants decided that success on only 1 outcome scale provided insufficient evidence of benefit and that a 1- or 2-point mean difference on an ordinal scale would not be convincing. The formulation of the trial question and choice of outcomes of interest led to the decision to use a binary global statistical test for analysis. With respect to power, Bath and colleagues[1] indicate that a dichotomous scale may be more efficient than an ordinal scale when treatment effects are expected to be clustered at a single state as was expected for recombinant tissue-type plasminogen activator. Others have shown that power generally increases when a global statistical test is used rather than a single outcome.[4] Bath and colleagues were also concerned that severe adverse events could be masked by dichotomization. However, a treatment-related imbalance in severe adverse events would not be ignored by Data and Safety Monitoring Boards or regulatory agencies.

When modest treatment differences are expected and/or the risks are low, the questions of interest may be quite different. Minimal or no disability may be too high a bar. Thus, investigators in other trials have chosen to dichotomize the modified Rankin Scale (mRS) at 0 to 2 and >2, or 0 to 3 and >3, or 0 to 4 and >4 generally based on the effects they expect for their therapies. These categorizations result in a lack of clarity regarding the clinical question of interest and often little rationale is given for the choice of cut points.[5] To avoid making a choice of a cut point, the authors recommend an ordinal scale or a sliding scale. Bath and colleagues[1] note that when studying a therapy expected to have modest benefits and limited risk, using the entire scale rather as a continuum rather than dichotomizing would provide greater power to detect differences if differences exist.

The authors make the important point that cut points for the sliding scale must be defined before randomization. Certainly, prespecification helps to avoid bias and Type I errors. To design a study with a sliding scale, an investigator must have extensive knowledge of the effect of various baseline measures on subsequent outcomes using the therapy being tested. This approach makes it difficult to take into account the baseline covariate interactions such as those found in at least 1 previous trial.[6] A more reasonable approach that does not lock investigators into a set of prespecified dichotomies based on baseline covariates is to use an ordinal or multiple logistic regression analysis depending on assumptions. This approach uses the full scale and allows for adjustment for baseline covariates. It is also possible to use a global statistical test based

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From the University of Texas School of Public Health at Houston, Houston, TX.

James C. Grotta, MD, was the Guest Editor for this paper.

Correspondence to Barbara C. Tilley, PhD, University of Texas School of Public Health at Houston, 1200 Hermann Pressler Drive, RAS E833, Houston, TX 77030. E-mail Barbara.C.Tilley@uth.tmc.edu

(*Stroke*. 2012;43:935-937.)

***Stroke* is available at http://stroke.ahajournals.org**
DOI: 10.1161/STROKEAHA.111.644864

on ranks[7] as is currently being used in a long-term Parkinson disease trial.[8] The outcome scales can be ordinal, continuous, or binary allowing use of the full spectrum of the scales and the inclusion of multiple aspects of stroke recovery as was done in the binary global statistical test for the NINDS tPA Stroke Trial. The global statistical test on ranks can be adjusted for covariates.

With respect to baseline covariates, a major omission in the articles was consideration of time from stroke onset to treatment. Many baseline covariates have been identified as confounders leading to poorer outcomes across placebo and intervention groups, but all appear to be unrelated to treatment response. Despite multiple analyses,[6] only time from stroke onset to treatment was identified as an effect modifier of the recombinant tissue-type plasminogen activator treatment effect[9] implying that the effect of recombinant tissue-type plasminogen activator varies by time from stroke onset. This interaction with time was validated in pooled analyses that included other trials.[10,11] Given the current concept that "time is brain," it is likely that time from stroke outcome to treatment could be an effect modifier for other therapies that have more modest effects and should be considered as an important covariate in analyses.

Last, although the authors of the 2 articles make a strong case for the use of the mRS as the outcome scale of choice, they acknowledge its limitations for hospitalized patients and for short-term assessments of change. Lees and colleagues puzzle over why the mRS becomes less reliable when structured interviews are used or when structured interviews are delivered over the telephone.[2] A structured approach to using the mRS was developed to provide better differentiation between treatment and intervention group. The structured mRS is currently used in some of the training programs for investigators in pharmaceutical and National Institutes of Health trials. Unfortunately, the structured mRS moves away from the simple construct of a generic scale designed to measure overall disability.[12] The Rankin Scale was not designed to separate other aspects of disability from "current stroke-related" disability. Although developed for patients with stroke, the scale was not stroke-specific. In an early article on the Rankin Scale discussing ways to improve agreement on the middle categories of the mRS, van Sweiten and colleagues[13] wrote: "It is important to include all causes of handicap in patients with TIA or minor stroke because they may suffer from other complications such as angina, myocardial infarction, intermittent claudication, or retinal infarction. Even nonvascular events may be side effects of the preventive treatment that is under study and ought to be included in the assessment." Such a "call it as you see it" approach was used in the NINDS Stroke Trial for the National Institutes of Health Stroke Scale and for the mRS as well as the other scales measuring stroke outcome. In a small study that used an unstructured mRS, the κ for agreement was high between the mRS based on clinical and telephone interviews but there was less agreement between different observers.[14] In addition to the Rankin Scale, modification has been made of the National Institutes of Health Stroke Scale to measure only the results of the current stroke. Psychometricians continually counsel against modifying existing scales without extensive psychometric testing. An example was the extensive testing of the new Movement Disorders Unified Parkinson's Disease Rating Scale (MDS-UPDRS) adapted from the UPDRS by adding a small number of questions and modifying rating categories.[15] There have been limited comparisons between the structured and unstructured approaches for either the mRS or National Institutes of Health Stroke Scale.

Conclusions

It is clear that the question of interest must be clearly specified in the trial planning stage and used to choose the measure of outcome and approach to analysis. There should be a strong rationale for the choice of outcome measures as reinforced by the discussion of Lees and colleagues[2] and careful consideration of the approaches to analysis as encouraged by Bath and colleagues.[1] Investigators need to be concerned both about power to detect potential beneficial effects and about the potential for Type I errors that can be introduced if outcomes and primary analyses are modified to increase power after a trial is completed. There is little evidence that rigorous psychometric validation has been conducted for the structured mRS or the newer version of the National Institutes of Health Stroke Scale. Given the continuing debates over the choice of outcome measures, this type of validation should be conducted before the structured mRS is considered a recommended scale.

Sources of Funding

Support provided by the University of Texas School of Public Health (salary) and National Institute of Neurological Disorders and Stroke NINDS 9U01NS043127.

Disclosures

The author is a Stroke Thrombolysis Trialists' Collaboration member and serves on Data and Safety Monitoring Committees for Pfizer and Novartis.

References

1. Bath PM, Lees KR, Schellinger PD, Altman H, Bland M, Hogg C, et al, and for the European Stroke Organisation Outcomes Working Group; Statistical analysis of the primary outcome in acute stroke trials. *Stroke*. 2012;43:1171–1178.
2. Lees KR, Bath PM, Schellinger PD, Kerr DM, Fulton R, Hacke W, et al, and for the European Stroke Organisation Outcomes Working Group. Contemporary outcome measures in acute stroke research: choice of primary outcome measure. *Stroke*. 2012;43:1163–1170.
3. Tilley B, Marler J, Geller N, 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 tPA trial. *Stroke*. 1996;27:2136–2142.
4. Tilley BC, Huang P, O'Brien, PC. Global assessment variables. In: D'Agostino RB, Sullivan L, Massaro J, eds. *Wiley Encyclopedia of Clinical Trials*. New York: John Wiley & Sons, Inc; 2007.
5. New PW, Buchbinder R. Critical appraisal and review of the Rankin Scale and its derivatives. *Neuroepidemiology*. 2006;26:4–15.

6. NINDS t-PS Stroke Study Group. Generalized efficacy of tPA for acute stroke: sub-group analyses of the NINDS tPA Stroke Trial. *Stroke*. 1997; 28:2119–2125.
7. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984;40:1079–1087.
8. Huang P, Goetz CG, Woolson R, Tilley BC, Kerr D, Palesch Y, et al. for The Parkinson Study Group. Using global statistical tests in long term Parkinson's disease clinical trials. *Mov Disord.* 2009;24:1732–1739.
9. Marler JR, Tilley BC, Lu M, Brott TG, Lyden PC, Grotta JC, et al. for the NINDS rtPA Stroke Study Group. Early stroke treatment associated with better outcome. *Neurology*. 2000;55:1649–1655.
10. The ATLANTIS, ECASS, and NINDS rt-PA Study Group Investigators. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rtPA stroke trials. *Lancet.* 2004;363: 768–774.
11. Lees KR, Bluhmki E, von Kummer R, Brott TG, Toni D, Grotta JC, et al. for the ECASS, ATLANTIS, NINDS, and EPITHET rt-PA Study Group Investigators. Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. *Lancet*. 2010;375:1695–1703.
12. Rankin J. Cerebral vascular accidents in patients over the age of 60: II. Prognosis. *Scot Med J*. 1957;2:200–215.
13. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*. 1988;19:604–607.
14. Candelise L, Pinardi G, Aritzu E, Musicco M. Telephone interview for stroke outcome assessment. *Cerebrovasc Dis*. 1994;4:341–343.
15. Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stebbins GT, et al. Movement Disorder Society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): process, format, and clinimetric testing plan. *Mov Disord*. 2007;22:41–47.

KEY WORDS: acute stroke ■ modified Rankin Scale ■ outcomes ■ scales