Treatment effects for which shift or binary analyses are advantageous in acute stroke trials

🐇

Jeffrey L. Saver, MD
Jeffrey Gornbein, DrPH

Address correspondence and reprint requests to Dr. Jeffrey L. Saver, UCLA Stroke Center, 710 Westwood Plaza, Los Angeles, CA 90095
jsaver@ucla.edu

ABSTRACT

Background: In acute stroke trials, functional outcome may be analyzed by dichotomizing ordinal outcome scales or by evaluating the entire scale range (shift analysis). The conditions under which shift or binary analysis will be more efficient have not been previously well delineated.

Methods: Model randomized clinical trials employing the modified Rankin Scale of global handicap were constructed to reflect 1) mild benefits experienced across all ranges of stroke severity (neuroprotective effect), 2) substantial benefits across all ranges of stroke severity (early recanalization effect), 3) substantial benefits across wide range of stroke severity but with limited ability to achieve fully normal outcome (late recanalization effect), 4) benefits clustered at unexpected health state transitions.

Results: In neuroprotective models, shift analysis was the most efficient technique in detecting a treatment effect. In the early recanalization models, dichotomization at excellent outcome and shift analysis were of comparable efficiency, both superior to dichotomization at good outcome. In the late recanalization models, dichotomization at good outcome performed best, shift analysis less well, and dichotomization at excellent outcome poorly. In the unexpected benefits model, shift analysis substantially outperformed dichotomization analyses. These patterns held among the seven actual acute trials reporting full range Rankin outcomes and showing treatment benefit identified in the literature.

Conclusions: The pattern of treatment effect of the intervention determines whether shift analysis or simple dichotomized analysis will be more efficient. Shift analysis is especially advantageous when treatments confer a relatively uniform, mild benefit to patients over a wide range of stroke severities or confer benefits at unexpected but clinically important health state transitions.

Neurology® 2009;72:1310–1315

GLOSSARY

CMH = Cochran–Mantel–Haenszel; **mRS** = modified Rankin scale; **NIHSS** = NIH Stroke Scale; **WMH** = Wilcoxon Mantel Haenszel.

Analyzing alterations in the distribution of patients over the entire range of functional outcome scales (shift analysis, analysis over ranks) is an increasingly employed primary endpoint analytic strategy in acute stroke trials.[1-3] Compared with conventional dichotomized analysis, shift analysis retains all information captured by ordinal outcome scales and frequently improves study power. Among the 55 trial comparisons analyzed by the Optimizing Acute Stroke Trials Collaborative Group, shift analysis was the more efficient endpoint analytic technique, detecting treatment effects in as many as 26% of trials vs as few as 9% with dichotomized analysis.[4]

Figure 1   The four analyzed treatment effect models

Model 1: Neuroprotective effect

	0	1	2	3	4	5/6
Placebo	16.7	16.7	16.7	16.7	16.7	16.7
Rx	22.1	16.7	16.7	16.7	16.7	11.2

Analysis Technique	Sample Size
Shift analysis	1200
Dichotomized 0-1 v 2-6	2508
Dichotomized 0-2 v 3-6	3956

Model 2: Early recanalization effect

	0	1	2	3	4	5/6
Placebo	16.7	16.7	16.7	16.7	16.7	16.7
Rx	24.7	16.7	14.7	14.7	14.7	14.7

Analysis Technique	Sample Size
Shift analysis	1364
Dichotomized 0-1 v 2-6	1200
Dichotomized 0-2 v 3-6	2246

Model 3: Late recanalization effect

	0	1	2	3	4	5/6
Placebo	16.7	16.7	16.7	16.7	16.7	16.7
Rx	16.7	16.7	24.9	13.9	13.9	13.9

Analysis Technique	Sample Size
Shift analysis	3342
Dichotomized 0-1 v 2-6	Infinity
Dichotomized 0-2 v 3-6	1200

Model 4: Benefits cluster at unexpected transitions

	0	1	2	3	4	5/6
Placebo	16.7	16.7	16.7	16.7	16.7	16.7
Rx	27.4	9.5	13.1	27.4	9.5	13.1

Analysis Technique	Sample Size
Shift analysis	1200
Dichotomized 0-1 v 2-6	5660
Dichotomized 0-2 v 3-6	Infinity

For each, the staircase schematic shows the pattern of the treatment effect, the upper table shows the percentile distribution of outcomes among the six Rankin levels in the treatment and control groups, and the bottom table shows the sample size required by each of the endpoint analytic techniques.

However, binary analyses sometimes outperform shift analysis in power. In the pivotal National Institute of Neurological Disorders and Stroke TPA trials, for example, lower *p* values are yielded by a dichotomized test ana-

lyzing only excellent outcome than by ranked tests analyzing benefit over the entire range of outcomes.[5,6] The conditions under which shift analysis or dichotomized analysis will be more efficient have not been well delineated.

For clinical trial designers and interpreters, it is important to better understand when it is advantageous to employ shift analysis and when dichotomized analysis as the prespecified primary statistical analytic technique. Model populations onto which prototypical treatment effects can be projected can clarify the circumstances in which shift or dichotomized analyses will maximize study power.

METHODS **Initial model: Equal distribution of 3-month outcomes in placebo group.** We modeled different patterns of treatment effect upon populations enrolled in parallel group, randomized, controlled, clinical trials. The primary endpoint was a six-level version of the modified Rankin scale (mRS) of global handicap. From the original seven-level mRS,[7] categories 5 (bedridden and severely disabled) and 6 (death) were collapsed into a single worst outcome level, reflecting the finding that many patients consider a severely disabled outcome to be as bad as or worse than death, rather than a more desirable outcome state.[1,8] In each model, half of the patients were assigned to placebo and half to active intervention. In the initial models, among the patients in the placebo group, the outcome distribution was set to an even distribution of patients in each of the six mRS outcome categories.

To generate the mRS outcome distribution in the treated group in each model, the following treatment effect profiles were applied: 1) mild benefits experienced across all ranges of stroke severity (neuroprotective effect), 2) substantial benefits experienced across all ranges of stroke severity (early recanalization effect), 3) substantial benefits experienced across a wide range of stroke severity but with limited ability to achieve fully normal outcome (late recanalization effect), 4) benefits clustered at unexpected health state transitions.

The neuroprotective model was designed to reflect a treatment effect that was mild in degree and experienced equally over the entire range of index stroke severity. The treatment group outcome distribution was generated by starting with the even distribution of the control group and then moving (as a result of treatment) an equal proportion of patients from each outcome category by one step to a better outcome category (5/6 to 4, 4 to 3, 3 to 2, 2 to 1, and 1 to 0).

The early recanalization effect model was designed to reflect a treatment effect that was substantial in degree and experienced equally over the entire range of index stroke severity. The treatment group outcome distribution was generated by moving an equal proportion of patients from each outcome category by three steps, or the maximum possible if less than three, to a better outcome category (5/6 to 2, 4 to 1, 3 to 0, 2 to 0, and 1 to 0).

The late recanalization model was designed to reflect a treatment effect that was substantial in degree, but unable to yield normal or excellent outcomes as patients had already experienced some degree of permanent injury prior to start of therapy. The treatment group outcome distribution was generated by moving an equal number of patients from each outcome category worse than level 2

| Table 1 | Percentile distribution of outcomes among placebo group patients in analyzed models | | | | | |

	Modified Rankin Scale at 3 months					
	0	1	2	3	4	5 + 6
Mean of all trial (n = 16) controls	11.0	18.3	12.8	16.0	18.2	23.9
Mild stroke trial controls (AbBESTT 2)	22.9	21.6	12.8	12.8	12.4	17.5
Severe stroke trial controls (PROACT 2)	12.0	5.0	8.0	22.0	19.0	34.0
Equal distribution	16.7	16.7	16.7	16.7	16.7	16.7

| Table 2 | Sample size requirements for trials with varying treatment effect patterns and placebo group outcome distributions | | |

	Shift analysis	Dichotomized 0–1 vs 2–6	Dichotomized 0–2 vs 3–6
Neuroprotective			
Mean	500	1,916	1,381
Mild	734	2,159	2,092
Severe	447	3,299	629
Equal	1,200	2,508	3,956
Early recanalization			
Mean	75	67	103
Mild	75	542	2,145
Severe	59	43	58
Equal	1,364	1,200	2,246
Late recanalization			
Mean	257	Infinite*	115
Mild	511	Infinite*	192
Severe	152	Infinite*	66
Equal	3,342	Infinite*	1,200
Unexpected benefits			
Mean	373	3,335	Infinite*
Mild	521	3,871	Infinite*
Severe	399	5,891	Infinite*
Equal	1,200	5,660	Infinite*

80% power, alpha = 0.05.
*Infinite since there is no difference between the two groups under these dichotomies.
Mean = placebo group distribution of final modified Rankin scale (mRS) scores the mean of 16 actual trials; Mild = placebo group distribution of final mRS scores from trial with lowest entry stroke severity NIH Stroke Scale (NIHSS) score among 16 actual trials; Severe = placebo group distribution of final mRS scores from trial with highest entry stroke severity NIHSS score among 16 actual trials; Equal = equal proportion of placebo patients with final outcome in each of the mRS categories.

by three steps, or the maximum possible if less than three, to a better outcome category, but no higher than level 2 (5/6 to 2, 4 to 2, 3 to 2).

The unexpected benefit pattern model was designed to reflect a treatment effect that conferred benefits at health state transitions not usually interrogated in dichotomized analyses. Binary analyses of the mRS in acute stroke trials generally employ cutpoints between levels 1 and 2 (0–1 vs 2–6) or between 2 and 3 (0–2 vs 3–6). The treatment group distribution was generated by clustering benefits at the level 1 to 0 and level 4 to 3 transitions, with equal proportions of patients shifting from 1 to 0 and 4 to 3, and half as many patients shifting from 2 to 0 and 5/6 to 3.

For sample size calculations, the significance level (alpha) was set at 0.05 and desired power (beta) at 80%. For dichotomized analyses, requisite sample sizes were derived employing the χ^2 test for comparing two proportions, with separate analyses dichotomizing at excellent vs good or less outcome (mRS 0–1 vs 2–6, dichotomization at excellent outcome) and good or better vs fair or less outcome (0–2 vs 3–6, dichotomization at good outcome).[9] For shift analysis, sample sizes were derived with the Wilcoxon rank sum test, also known as the Mann–Whitney U test (StatXact, Cytel Inc.). The overall treatment effect size was adjusted so that the most efficient statistical analytic technique for a given model required a sample size of 600 patients per treatment arm (1,200 total). The sample sizes of the less efficient methods can be compared to this 600 per treatment standard. Since sample sizes were all large, corrections for very small samples were not needed.

Additional models: Placebo outcome distributions from real trials. The initial hypothetical modeling was performed using the simplifying assumption that the placebo group of patients would experience an even distribution of outcomes among each of the six mRS categories. An additional set of analyses were performed using actual placebo group distributions from acute stroke trials. A systematic Medline search was undertaken to identify all acute stroke randomized trials reporting placebo group outcomes across the six analyzed mRS categories. Three outcome distributions were applied to the mean placebo outcome distribution from all 16 actual trials, and, in sensitivity analysis, to the actual trial with the most extremely mildly affected enrolled patients (lowest entry NIH Stroke Scale [NIHSS] score) and the actual trial with the most extremely severely affected enrolled patients (highest entry NIHSS score). The four patterns of treatment effect (neuroprotective, early recanalization, late recanalization, unexpected response distribution) were applied to these three placebo distributions, in addition to the initial even distribution generating 16 model trials. Sample size estimates for each of these 16 outcomes were calculated using shift analysis (Wilcoxon rank sum), dichotomization at 0–1 vs 2–6 (χ^2 test), and dichotomization at 0–2 vs 3–6 (χ^2 test).

Comparison of treatment effect models with actual trials in each class. To confirm the applicability of the modeled prototypical treatment patterns to treatment patterns actually observed in clinical trials, actual clinical trials were selected for detailed analysis, based on the following criteria: 1) trialists reported positive results on the 3-month mRS endpoint when analyzed by shift analysis, dichotomization at extreme good outcome, or dichotomization at moderate good outcome, 2) outcomes across all mRS categories available for analysis, 3) trial intervention matched one of the model treatment effects based on tested agent's known mechanism of action and trial enrollment time window. For the trials identified by these criteria, the efficiency of the different statistical tests was analyzed by calculating the p values for evidence of a treatment effect from the final mRS outcome distributions using shift analysis, dichotomization at 0–1 vs 2–6 (χ^2), and dichotomization at 0–2 vs 3–6 (χ^2). When the primary paper itself reported p values using a

shift analysis analytic technique (e.g., Wilcoxon rank sum, Cochran-Mantel-Haenszel, bootstrap), these values were employed. When no ranked analysis was described, p values were calculated from the reported outcome distributions using the Wilcoxon rank sum test. To graphically map similarities and differences in the effects of treatment in the prototypical models and exemplar actual trials, treatment effect profile bar graphs were generated. For each model trial and one exemplar trial in each class, a bar graph shows the absolute difference in the percent of patients achieving a good outcome under active treatment vs control, when good outcome is defined as mRS = 0 (column 1), mRS = 0–1 (column 2), mRS = 0–2 (column 3), mRS = 0–3 (column 4), mRS = 0–4 (column 5).

RESULTS Models with simplified, equal distribution of 3-month outcomes in placebo group. Model results when there is an even distribution of outcomes in the placebo group are shown in figure 1. In the neuroprotective model, shift analysis outperformed both binary analyses, with a sample size requirement of 1,200 (600 in each arm), compared with 2,508 for dichotomization at excellent outcome and 3,956 for dichotomization at good outcome. In the early recanalization model, shift analysis and dichotomization at the excellent outcome level showed similar efficiency, outperforming dichotomization at a good outcome level. In the late recanalization model, di-

chotomization at a good outcome level performed best, shift analysis less efficiently, and dichotomization at an excellent outcome level highly inefficiently. In the unexpected benefits pattern model, shift analysis performed best and both dichotomized analyses were highly inefficient. Under the late recanalization and unexpected benefit scenarios, the dichotomization uses the "wrong" cutpoint and therefore does not capture the group difference at all. In this case, the dichotomization chosen causes the two groups to appear identical, leading to a sample size of "infinity."

Models using placebo outcome distributions from real trials. The Medline search identified 16 trials with detailed reporting of 3-month placebo group outcomes across all mRS categories. Among these, entry NIHSS scores were reported in 15 (one trial used the Scandinavian Stroke Scale) and the median baseline score was 12.25. The trial enrolling the least severe deficit patients was AbBESTT 2 (NIHSS 8) and the trial enrolling the most severe was PROACT 2 (NIHSS 17). Table 1 shows the 3-month mRS outcome distributions in the placebo groups for the mean of all 16 trials, the least severe trial, and the most severe trial.

The pattern seen in the initial analysis of trial models with an equal distribution of outcomes among all mRS categories in the placebo group generally held among the trial models with placebo group outcome distributions matching the mean and extremes of actual trials (table 2). Across all four placebo group outcome distributions (equal, mean of actual trials, least severe, most severe), the nominally most efficient statistical test to detect the different treatment effect patterns were as follows: neuroprotective pattern, shift analysis in all four models; early recanalization pattern, dichotomization at extreme good outcome in three models and shift analysis in one model; late recanalization, dichotomization at moderate good outcomes in all four models; unexpected pattern, shift analysis in all four models. Overall, among the 16 models, shift analysis was the most efficient statistical test in 9, second most efficient in 7, and never the least efficient; dichotomization at extreme good outcome was the most efficient in 3, second most efficient in 7, and least efficient in 7; and dichotomization at moderate good outcome was the most efficient in 4, second most efficient in 4, and least efficient in 8.

Comparison of models with actual trials in each class. For the comparison of actual treatment trials within each class of agent effect with the prototypical class effect models, 7 of the 16 trials met selection criteria. For convenience, National Institute of Neurological

Figure 2 Treatment effect profile maps
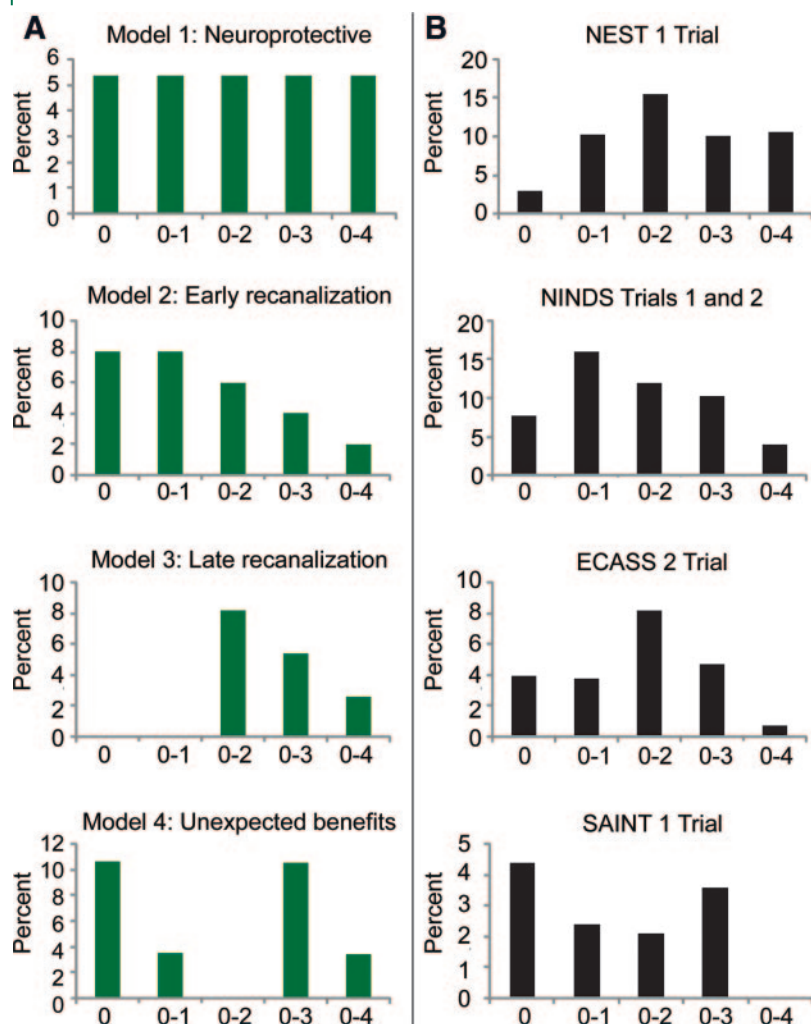
In individual bar graphs, each bar represents the absolute difference in the percent of patients achieving a good outcome under active treatment vs control, when good outcome is defined as modified Rankin scale (mRS) = 0 (column 1), mRS = 0–1 (column 2), mRS = 0–2 (column 3), mRS = 0–3 (column 4), mRS = 0–4 (column 5). In the overall figure, the four bar graphs in A are the treatment effect profile maps for the four models. The four bar graphs in B are the treatment effect profile maps for actual clinical trials that resemble each of the prototypical models.

then dichotomization at good outcome levels. In the two late recanalization trials, PROACT 2 and ECASS 2,[12,13] dichotomization at good outcome levels was the most powerful analytic technique. The treatment profile maps of the model trials and the exemplar actual trials are shown in figure 2.

DISCUSSION Analysis of models incorporating prototypical treatment effects demonstrates that shift analysis is a more efficient statistical technique than dichotomized analysis when treatments yield a small and uniform degree of benefit over all ranges of stroke severity (neuroprotective effect) and benefits that cluster at less frequently interrogated health state transitions (unexpected benefits). In contrast, when treatments confer a substantial benefit over all ranges of stroke severity (early recanalization effect), dichotomizing at excellent outcomes is mildly more powerful than shift analysis, while dichotomizing at good outcomes is not as efficient. When treatments offer substantial benefits but cannot provide a cure (late recanalization effect), dichotomization at good outcome is the most efficient statistical technique, shift analysis less efficient, and dichotomization at excellent outcomes highly inefficient.

These findings accord with prior model studies, actual acute stroke clinical trial observations, and formal statistical theory. In a model population study, an extreme neuroprotective effect was investigated, with every single patient treated benefitting a small amount from therapy.[14] This more extreme version of model 1 showed, as would be expected, an even greater advantage of shift analysis over binary analyses. When the pattern of treatment effect is held constant, changes in the size of treatment effects alter the absolute power and sample size values associated with each test, but not the relative efficiency among the tests.

In actual clinical trials, shift analysis generally performed better than dichotomized analyses in trials of agents showing relatively even neuroprotective effects and in trials in which benefits clustered at less frequently prioritized health state transitions. Dichotomization at excellent outcome levels has been the more powerful technique in early recanalization trials, while dichotomization at good outcome levels has been the most powerful analytic technique in late recanalization trials.

Formal statistical literature studies have investigated the selection of statistical tests to match the pattern of treatment effect.[15-17] In non-normally distributed data, such as highly skewed continuous distributions or ordered categorical outcomes, the most efficient statistical test has been shown to be dependent upon the type of alternative that the study aims

Disorders and Stroke Trials 1 and 2 were analyzed together, yielding six studies for detailed analysis. The performance of the different statistical tests in each of these studies is shown in table 3. In the trial of an agent with a neuroprotective mechanism (NEST),[10] shift analysis clearly outperformed dichotomization at extreme good outcome and nominally outperformed dichotomization at moderate good outcome. In the trial with benefits clustering at less frequently prioritized health state transitions, SAINT 1,[1] shift analysis outperformed both dichotomized analyses. In the two early recanalization treatment studies, the National Institute of Neurological Disorders and Stroke–TPA trial and MELT,[5,11] dichotomization at excellent outcome levels was the more efficient technique, followed by shift analysis and

to detect. The current study is an application of this general insight to the specific case of acute stroke therapy, delineating the relative efficiencies of commonly employed statistical tests when applied to four patterns of treatment effects expected or observed in acute stroke trials.

This study has limitations. The treatment effects captured in the models are idealized; real agents and real trials will generally show less pure patterns of response. Sometimes blends of two models will better capture an agent's likely profile than will an individual model. Sample size calculation for shift analysis in the models was performed using the Wilcoxon rank sum statistical test. In the OAST collaborators analysis, other statistical tests for ordinal data were slightly more efficient, including the robust ranks test, ordinal logistic regression, and bootstrap difference in mean rank.[4] However, the Wilcoxon performed nearly as well and has advantages of less computational complexity than the other tests as well as being well known and having widespread availability. This study investigated shift and binary analyses, but not other analytic approaches to ordinal endpoints, such as proportional odds models, responder analysis, and the global statistic.[3,18] Additional studies are needed to clarify the conditions under which these alternative options may enhance or reduce study power compared with shift and dichotomized analyses.

The pattern of the treatment effect of the intervention under investigation determines whether shift analysis or simple dichotomized analysis will be more efficient. Shift analysis is especially advantageous when treatments confer a relatively uniform, mild benefit to patients over a wide range of stroke severities of cluster at unexpected but clinically important health state transitions. When benefits cluster at a single health state transition that can be predicted beforehand, dichotomized analysis focused upon that state transition, but not others, will outperform shift analysis. These insights can guide clinical trialists in selecting the prespecified primary mode of statistical analysis for acute stroke clinical trials.

REFERENCES

1. Lees KR, Zivin JA, Ashwood T, et al. NXY-059 for acute ischemic stroke. N Engl J Med 2006;354:588–600.
2. Ehrenreich H, Hasselblatt M, Dembowski C, et al. Erythropoietin therapy for acute stroke is both safe and beneficial. Mol Med 2002;8:495–505.
3. Saver JL. Novel end point analytic techniques and interpreting shifts across the entire range of outcome scales in acute stroke trials. Stroke 2007;38:3055–3062.
4. 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;38:1911–1915.
5. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995;333:1581–1587.
6. Savitz SI, Lew R, Bluhmki E, Hacke W, Fisher M. Shift analysis versus dichotomization of the modified Rankin scale outcome scores in the NINDS and ECASS-II trials. Stroke 2007;38:3205–3212.
7. Rankin J. Cerebral vascular accidents in patients over the age of 60, II: prognosis. Scott Med J 1957;2:200–215.
8. Samsa G, Matchar D, Goldstein L, et al. Utilities for major stroke: results from a survey of preferences among persons at increased risk for stroke. Am Heart J 1998;136:703–713.
9. Sulter G, Steen C, De Keyser J. Use of the Barthel index and modified Rankin scale in acute stroke trials. Stroke 1999;30:1538–1541.
10. Lampl Y, Zivin JA, Fisher M, et al. Infrared laser therapy for ischemic stroke: a new treatment strategy: results of the NeuroThera Effectiveness and Safety Trial-1 (NEST-1). Stroke 2007;38:1843–1849.
11. Ogawa A, Mori E, Minematsu K, et al. Randomized trial of intraarterial infusion of urokinase within 6 hours of middle cerebral artery stroke: the middle cerebral artery embolism local fibrinolytic intervention trial (MELT) Japan. Stroke 2007;38:2633–2639.
12. Stingele R, Bluhmki E, Hacke W. Bootstrap statistics of ECASS II data: just another post hoc analysis of a negative stroke trial? Cerebrovasc Dis 2001;11:30–33.
13. Furlan A, Higashida R, Wechsler L, et al. Intra-arterial prourokinase for acute ischemic stroke: The PROACT II study: a randomized controlled trial: Prolyse in Acute Cerebral Thromboembolism. JAMA 1999;282:2003–2011.
14. Berge E, Barer D. Could stroke trials be missing important treatment effects? Cerebrovasc Dis 2002;13:73–75.
15. White IR, Thompson SG. Choice of test for comparing two groups, with particular application to skewed outcomes. Stat Med 2003;22:1205–1215.
16. Walters SJ. Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36. Health Qual Life Outcomes 2004;2:1–17.
17. Federov V, Mannino F, Zhang R. Consequences of dichotomization. Pharm Stat Epub 2008.
18. Stokes ME, Davis CS, Koch GG. Categorical Data Analysis Using the SAS System, 2nd Edition. Cary, NC: SAS Institute Inc., 2000.

Neurology®

Treatment effects for which shift or binary analyses are advantageous in acute stroke trials

Jeffrey L. Saver and Jeffrey Gornbein

This information is current as of December 17, 2008

AMERICAN ACADEMY OF
NEUROLOGY®