On the Analysis and Interpretation of Outcome Measures in Stroke Clinical Trials
Lessons From the SAINT I Study of NXY-059 for Acute Ischemic Stroke
James A. Koziol, PhD; Anne C. Feng, MS

Background and Purpose—A variety of primary end points have been used in acute stroke trials. We focus on the modified Rankin Scale, a reliable and valid ordinal outcome measure that assesses disability on a 7-point scale.

Methods—We provide an abbreviated discussion of analytical methods for ordinal scales, and related effect size measures; we illustrate these methods and their interpretation with outcome data from the SAINT I study of NXY-059 in acute ischemic stroke.

Results—The nonparametric Mann-Whitney statistic provides a straightforward method for analysis of the modified Rankin Scale, and incorporates associated measures of effect size. These measures are directly related to the concepts of Number Needed to Treat and Number Needed to Harm.

Conclusions—Our re-examination of the outcome data from the SAINT I study provides little evidence for the purported efficacy of NXY-059. More broadly, analysis and interpretation of ordinal outcome scales based on ascribed numerical values to the steps of the scale should be done cautiously. Statistical treatment of multiple primary outcome measures in acute stroke clinical trials should be established before analysis. Lastly, conflating statistical and clinical significance should be avoided. (Stroke. 2006;37:2644-2647.)

Key Words: Mann-Whitney statistics ▪ numbers needed to treat or harm

NXY-059 is a free-radical–quenching agent that has been shown to be neuroprotective in animal models of stroke. Lees et al1 have recently reported the results of the Stroke-Acute Ischemic NXY Treatment (SAINT I) study, a randomized, double-blind, placebo-controlled trial of patients with acute ischemic stroke, and found that reduction in cerebral injury and improvement in neurological outcome occurs in patients who are treated with NXY-059 within 6 hours of the onset of stroke. The primary efficacy end point of the SAINT I study was disability at 90 days, as measured according to scores on the modified Rankin Scale (mRS). We here re-examine the summary findings of the SAINT I study relating to the primary efficacy end point, and argue that both statistical and clinical significance of the efficacy of NXY-059 are at best equivocal. Issues raised in our reanalysis are not specific to the SAINT I study, but bear on the general topic of analysis and interpretation of outcomes in clinical trials of stroke. Our aim here is to provide abbreviated commentary and guidelines relating to these issues, using data from the SAINT I study for illustrative purposes.

Primary Efficacy End Point of the SAINT I Study
The primary efficacy end point of the SAINT I trial was the score on the mRS at 90 days posthospitalization (or, the last rating). The mRS is a 7-point ordinal scale; the categories, as originally described,2 are given in Table 1.

We present the mRS primary outcome of the SAINT I study in Table 2, which was derived from the authors’ Figure 2. There are 849 patients in the placebo arm, and 850 patients in the NXY-059 arm, who were included in the efficacy analysis. Note that the authors grouped patients with mRS = 5 or 6; the individual breakdown is not available.

The authors declared that “NXY-059 significantly improved the overall distribution of scores on the modified Rankin scale, as compared with placebo (P = 0.038 by the Cochran-Mantel-Haenszel test)”. We are unable to verify their analysis because their procedure relies on prior knowledge of stratification factors that are also unavailable to us. Nevertheless, the data in Table 2 are prototypical of ordinal outcome measures in stroke clinical trials and will be used for illustrative purposes. Moreover, insofar as “prognostic factors that influence stroke outcome were well matched and showed no interaction with the treatment effect” in the SAINT I study, our reanalyses will provide additional insight into the trial findings.

Nonparametric Analysis of Efficacy
Let us examine the data in Table 2, using a straightforward approach based on the nonparametric Mann-Whitney statistic.
TABLE 1. mRS*

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No symptoms at all</td>
</tr>
<tr>
<td>1</td>
<td>Slight disability: unable to carry out all usual duties and activities</td>
</tr>
<tr>
<td>2</td>
<td>Moderate disability: requires some help, but able to walk without assistance</td>
</tr>
<tr>
<td>3</td>
<td>Moderately severe disability: unable to walk without assistance,</td>
</tr>
<tr>
<td></td>
<td>and unable to attend to own bodily needs without assistance</td>
</tr>
<tr>
<td>4</td>
<td>Severe disability: bedridden, incontinent, and requires constant</td>
</tr>
<tr>
<td></td>
<td>nursing care and attention</td>
</tr>
<tr>
<td>5</td>
<td>Dead</td>
</tr>
</tbody>
</table>

*This modification of the original Rankin scale was codified by van Swieten et al.² We have added Grade 6, dead, to the categories described by van Swieten et al.

We remark that this approach was described over 20 years ago in the medical setting,³ though its application to categorical data predates this. Our null hypothesis is that the distributions of mRS scores are identical in the 2 arms of the trial; we might be particularly interested in a 1-sided alternative hypothesis, that mRS scores on the NXY-059 arm tend to be smaller than those on the placebo arm (though 2-sided alternatives might also be entertained, especially by regulatory agencies). We find that the Mann-Whitney test yields a 1-sided P value of 0.0768, and 2-sided P value of 0.153, in favor of the NXY-059 arm compared with the placebo arm in Table 2.

Incidentally, if we ignore the categorical nature of these data, and merely compare mean mRS scores on the 2 arms with a standard 2-sample t test, we have: mean mRS = 2.841 (SD 1.75) on the placebo arm, mean mRS = 2.713 (SD 1.80) on the NXY-059 arm, t[848] = 1.484, 1-sided P = 0.069, 2-sided P = 0.138. With this large sample size, a Z value of 0.153, in Pr[X > Y] + 1/2 Pr[Y = X] (Pr denoting probability). Again in the present context, larger mRS scores are more deleterious than smaller scores, so a measure of efficacy of NXY-059 relative to placebo would be 1 - Pr[X > Y] + 1/2 Pr[X = Y]. This is simply the probability that an individual randomly assigned to placebo would have a larger (more deleterious) mRS outcome than n individual randomly assigned to NXY-059, plus one-half the probability that there would be no difference in mRS outcome.

With continuous response variables X and Y, the probability of tied data, Pr[X = Y], would be 0, and the effect size measure θ = Pr[X > Y] is immediately interpretable as a measure of the separation of the 2 underlying distributions. Because there is considerable overlap of mRS scores in the arms of the SAINT I study, it is perhaps more reasonable here to consider the 3 measures Pr[X > Y] (better outcomes on placebo), Pr[X > Y] (better outcomes on NXY-059), and Pr[X = Y] (no difference in outcome). Acion et al have termed these measures probabilistic indices, and argued that they provide intuitive and simple effect size measures with categorical outcomes.⁴

Our point estimates of these measures from the SAINT I study are: Pr[X > Y] = 0.392, Pr[X > Y] = 0.431, and Pr[X = Y] = 0.177. However, it is desirable to report CIs for these measures. Newcombe gives a nice discussion of confidence interval construction for measures derived from the Mann-Whitney statistic⁵,⁶; we here adopted a bootstrap approach, which has reasonable properties with a minimum of model assumptions. We first drew 10,000 random samples with replacement of sizes 849 and 850 using the empirical distributions for placebo and NXY-059 respectively, as given in Table 2, and calculated the 3 measures Pr[X > Y], Pr[X = Y], and Pr[X > Y] from each of the random samples. With the bootstrap percentiles,⁷ we then found the joint 95% CIs for these 3 measures to be 0.404 to 0.460, 0.173 to 0.182, and 0.364 to 0.419.

TABLE 2. mRS Scores at 90 Days in the SAINT I Study

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 or 6*</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>93</td>
<td>170</td>
<td>99</td>
<td>108</td>
<td>175</td>
<td>204</td>
<td>849</td>
<td>2.84</td>
<td>1.81</td>
</tr>
<tr>
<td>NXY-059</td>
<td>131</td>
<td>153</td>
<td>97</td>
<td>121</td>
<td>144</td>
<td>204</td>
<td>850</td>
<td>2.71</td>
<td>1.75</td>
</tr>
</tbody>
</table>

*This is a pooled category: all individuals in this category were considered to have mRS score 5, for calculation of the sample means and SDs.

In the context of the SAINT I study, let X₁, X₂, ..., X₈₄₉ denote the mRS scores of the 849 individuals on the placebo arm, and let Y₁, Y₂, ..., Y₈₅₀ denote the mRS scores of the 850 patients on the NXY-059 arm. The Mann-Whitney statistic U is defined as

\[
\sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij}
\]

where \( U_{ij} = 1 \), 1/2, or 0 depending on whether \( Y_i \) is greater than, equal to, or less than \( X_j \), \( m = 849 \), and \( n = 850 \). Then, \( U_{mn} \) is an empirical estimate of the effect size measure \( \theta = Pr(Y > X) + 1/2 Pr(Y = X) \) (Pr denoting probability). Again in the present context, larger mRS scores are more deleterious than smaller scores, so a measure of efficacy of NXY-059 relative to placebo would be 1 - \( \theta = Pr(X > Y) + 1/2 Pr(X = Y) \). This is simply the probability that an individual randomly assigned to placebo would have a larger (more deleterious) mRS outcome than \( n \) individual randomly assigned to NXY-059, plus one-half the probability that there would be no difference in mRS outcome.

The mere presentation of \( P \) values for the difference in efficacy between placebo and NXY-059 in the SAINT I study is less informative than estimation of the magnitude of any putative treatment effect. In this regard, the Mann-Whitney statistic leads immediately to an effect size measure, applicable to both continuous and ordinal data, and without parametric assumptions. We briefly summarize construction of the measure.

Effect Size Measure Based on the Mann-Whitney Statistic

The mere presentation of \( P \) values for the difference in efficacy between placebo and NXY-059 in the SAINT I study is less informative than estimation of the magnitude of any putative treatment effect. In this regard, the Mann-Whitney statistic leads immediately to an effect size measure, applicable to both continuous and ordinal data, and without parametric assumptions. We briefly summarize construction of the measure.

Number Needed to Treat

Lees et al stated that the clinical benefit of NXY-059 "amounts to an average improvement of 0.13 point on the mRS per patient, which suggests that about 8 patients would need to be treated to achieve improvement equal to 1 point on the scale for one patient." Their calculation of number needed to treat (NNT)⁸ is the reciprocal of the observed mean difference in mRS on the 2 arms, \( 1/(2.84 - 2.71) \approx 1/0.13 \approx 7.69 \), where 2.84 and 2.71 are the mean mRS scores on the placebo and NXY-059 arms (Table 2). However, Lees et al fail to report
even an approximate CI for their NNT calculation, even though this is a standard recommendation.9

Unfortunately, in the present setting, the determination of a nominal 95% CI is fraught with difficulties. We calculated a bootstrap 95% CI for the difference in mean mRS (placebo−treatment) as −0.026 to 0.289. As above, the bootstrap CI was based on 10 000 replicate samples. If the lower limit of the 95% CI for mean difference in mRS had been positive, then we would merely have inverted the 2 limits of this CI to derive the corresponding CI for NNT. But, because the CI for the mean difference in mRS includes 0 (that is, no treatment effect), the inverted “CI” will include the value infinity (1/0), and the negative limit −0.026 of the CI is indicative of “harm” with NXY-059, not efficacy.

There is an alternative formulation of the concept of NNT with ordinal data, explicitly formulated with the mRS.10 With the notation introduced in the previous section, Saver would define NNT as 1/Pr(X>Y), and number needed to harm (NNH) as 1/Pr(Y>X). Using Saver’s criteria, we immediately calculate joint 95% CIs as 2.17 to 2.48 for NNT (better outcome on NXY-059 than placebo), and 2.39 to 2.75 for NNH (better outcome on placebo than NXY-059).

It is apparent that invocation of the concept of NNT with ordinal data as with the outcome variable of the SAINT I study (Table 2) is less transparent than with binomial data: in particular, we must also entertain the real possibility of indifference, that is, no difference in outcome between the competing regimens (here NXY-059 and placebo), as well as the converse possibility, reflected in the NNI. We remark that the “continuous variable” formulation used by Lees et al for NNT is highly dependent on the scoring scheme used for the categories: we could have weighted the categorical scores differently, resulting in different NNT values. In contrast, the Saver measures would not be affected, so long as the transformations in scores were monotone (because the underlying Mann-Whitney statistic is invariant to monotone transformations of the data). Regardless, whenever the CI for the “difference parameter” includes 0 (no efficacy) then construction and interpretation of CIs for NNT is troublesome. Indeed, had Lees et al followed conventional practice by constructing a CI for the difference in mRS between the NXY-059 and placebo arms, this would have alerted them and interested readers to the problematic nature of the efficacy of NXY-059 based on the concept of NNT.

**Statistical Versus Clinical Significance**

We have questioned the strength of the statistical significance of perceived differences between the placebo and NXY-059 arms of the SAINT I study. We now address perhaps a more relevant issue, namely, the clinical significance of the findings. How does an observed mean difference of 0.13 U on the mRS scale between the 2 arms translate to the benefit of individual patients? Because the mRS is an integer scale, a 1-point change is the minimal observable difference in this scoring scheme. In this regard, is a 1-point improvement in outcome, say from mRS 1 to mRS 0, of comparable clinical relevance as a 1-point improvement from mRS 6 to mRS 5? We might argue that preventing death is of greater consequence, at least, to the patient, than reduction from 1 to 0, yet both are weighted equally in Lees’ and our analyses. Others might well counter that death (mRS=6) is a “better” outcome than continuation of a debilitating and dignity-stripping condition (mRS=5). It appears, then, that assignment of numerical scores to the steps of the mRS is fundamentally challenging, and that adherence to the numerical scores of the ordinal scaled mRS, or any other categorical scoring scheme, as a proxy for efficacy, can legitimately be challenged. A judgmental approach may well be preferable to a statistical approach for establishing the validity of any numerical scoring scheme ascribed to the steps of the mRS.

If it is agreed that the mRS provides the appropriate numerical scale on which to base efficacy judgments for clinical trials such as SAINT I, then we suggest that a moderate effect size of 0.5,11 equivalently, a “one-half standard deviation” rule12,13 might be invoked as indicative of “clinical significance”. According to this guideline, an observed difference of about 0.8 U (1/2 SD, from Table 2) on the mRS scale would be suggestive of clinical significance. In contrast, the SAINT I study has an observed effect size of 0.17 (0.13/1 pooled SD) in favor of NXY-059, quite modest by these standards.14 Even if one accepts the authors’ declaration of statistical significance for the shift in mRS scores favoring NXY-059, they appear to have conflated statistical with clinical significance.

**Choice of Outcome Measures**

There were in fact 2 primary outcomes in the SAINT I study: the mRS as discussed here, and change in the National Institutes of Health Stroke Scale (NIHSS). A reviewer has pointed out “a glaring incongruity in the results reported by Lees et al: the coprimary outcome measure, the NIH Stroke Scale, showed no difference between NXY-059 and placebo, even though the NIHSS has been convincingly shown to be more sensitive that the mRS.”15,16

Now, it is not at all unusual to have 2 or more primary outcomes in Phase III clinical trials in stroke: indeed, the National Institute of Neurological Disorders and Stroke (NINDS) recombinant tissue plasminogen activator stroke trial incorporated 4 neurological and functional scales (Rankin, NIHSS, Glasgow Outcome Scale, and Barthel Index), in the primary assessment of whether recombinant tissue plasminogen activator was associated with a significantly improved long-term functional and neurological outcome.17 On the other hand, it is highly unusual not to adjust statistically for the 2 primary outcomes of SAINT I, either by declaring a priori that one or the other outcome would need to achieve statistical significance at the α=0.05 level when tested individually, in order to declare statistical significance for the trial at the overall α=0.05 level (this being a simple Bonferroni correction for the multiple testing), or, by combining the 2 outcome measures into 1 global statistic, as assessed at the overall α=0.05 level.18 Unfortunately, Lees et al followed neither standard, instead reported uncorrected P values for their 2 primary outcome measures; they then imputed statistical significance for the trial solely on the basis of their analysis of mRS outcomes. This declaration is puzzling, especially as the underlying clinical protocol would
likely have specified a statistical adjustment to obviate inflation of Type I error with 2 primary outcome variables, particularly when reporting findings to regulatory agencies.

When choosing outcome scales for acute stroke clinical trials, it would be prudent to select scales with proven reliability, validity, and responsiveness (that is, sensitivity to clinical change). We refer interested readers to New and Buchbinder for discussion of the mRS from a clinimetric viewpoint (and also, an extensive and relevant bibliography), and Bruno et al. for discussion of the NIHSS.

Conclusions

We have described a straightforward methodology based on the nonparametric Mann-Whitney statistic for analysis of ordinal outcome measures such as the mRS, including, in particular, measures of effect size. These measures can be immediately related to the concepts of NNT and NNH, and ought to be reported with associated CIs. Nevertheless, these concepts need to be interpreted cautiously in the context of ordinal data, as the numerical scaling of the ordinal category steps is not inviolate.

Multiple primary end points in acute stroke clinical trials are not uncommon. One would hope the various outcome scales are congruent and statistical adjustment for multiple end points is appropriate. Lastly, conflation of statistical and clinical significance should be avoided.

Acknowledgments

We thank the reviewers for their incisive comments.

Sources of Funding

This research was supported in part by the Stein Endowment Fund, Department of Molecular and Experimental Medicine, The Scripps Research Institute.

Disclosures

None.

References


